Statistical Methods for Estimating Pedigrees and Demographic Parameters Using Genetic Markers

by

Amy Ko

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Associate Professor Haiyan Huang
Associate Professor Elizabeth Purdom
Associate Professor Erica Bree Rosenblum

Fall 2018

Statistical Methods for Estimating Pedigrees and Demographic Parameters Using Genetic Markers

Abstract

Statistical Methods for Estimating Pedigrees and Demographic Parameters Using Genetic Markers

by

Amy Ko

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

The concept of relatedness is fundamental is many areas of genetic studies, such as disease association studies, conservation genetics, and inferences about the demographic history and social structure of a population. Related individuals show signatures of shared ancestry in their genomes, which can then be analyzed to infer the relationship. In this thesis, we present statistical methods for estimating the relationship between individuals at varying time scales and the population parameters that produced such structure. In particular, we develop methods for analyzing genetic markers to estimate the pedigrees of close relatives and the mating parameters, such as the effective population size, that govern the population. Using simulations, we find that our method can infer pedigrees and the effective population size better than existing methods. We also discuss a method to infer regions of Neanderthal ancestry in human genomes, which can then be used to study the distant relationship between Neanderthals and humans. We apply the method on a sample of ancient humans to estimate the date of admixture between Neanderthals and humans.

To my family

For all your love and support.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my advisor Rasmus Nielsen for his guidance, encouragement, and optimism. Without his support, I would not be where I am today. I would also like to thank my labmates for making our little office a fun place to be. I will cherish our memories of surprise birthday parties, complaining (very loudly) about the office temperature, and helping each other procrastinate in later afternoons.

I would like to thank my partner, Jeff, for putting up with me throughout this whole process. His kindness and patience have helped me get through the rough patches and I'm truly grateful to have him in my life. Finally, I would like to thank my family for always being there for me through thick and thin. I look forward to starting the next chapter of my life with all the insights and lessons I've learned from my friends, family, and colleagues.

Chapter 1

Introduction

The concept of relatedness is fundamental in many areas of genetic studies. Relationships among individuals at a short time scale, such as the pedigree structure of close relatives, can be used to study the social organization of a population, such as the degree of polygamy or the offspring distribution among mothers and fathers [8]. In conservation genetics, information about the relatedness among individuals can be used to design an appropriate breeding scheme to prevent inbreeding between close relatives. The concept of relatedness can also extend to the relationship between two distinct populations. One example is the relationship between Neanderthals and modern humans through interbreeding, or admixture, which has been of great interest in genetic studies about human demographic history [53, 77, 58]. In particular, we can analyze the signatures of Neanderthal ancestry in human genomes to address various questions about the demographic events that may have shaped the two populations' relationship to one another. Such questions include the time of admixture between the two populations and how much of the Neanderthal genome was transferred to humans during the admixture event.

In Chapters 2 and 3, we turn our attention to the inference of pedigrees, which contain information about the genealogical relationships among individuals at the finest resolution. Because pedigrees play an important role in a wide array of genetic studies–which will be detailed in Chapter 2–many methods for estimating pedigrees from genetic data have been developed to date. Existing inference methods fall broadly into two categories: those that estimate pairwise relationships only [70, 46, 61, 65, 48, 66] and those that aim to reconstruct the entire pedigree [69, 4, 78, 23, 18, 13, 56, 80, 38, 5, 82, 12, 25, 14, 63, 7, 62, 39, 55]. Although pairwise methods are computationally fast, estimated pairwise relationships do not necessarily translate to the correct pedigree, as piecing together pairwise relationships may not produce a valid pedigree. Furthermore, because the coefficient of variation in genome sharing between two individuals becomes larger as the relationship becomes more distant [28], distinguish competing relationships from each other becomes increasingly difficult. Methods that estimate the entire pedigree has an advantage in this regard. Several studies have shown that the accuracy of pairwise relationship inference can be improved by considering all relationships in the sample simultaneously and resolving uncertain relationships in the context of other individuals [63, 39, 55]. Furthermore, the estimated pedigree is by construction a valid pedigree, which then can be used to study population parameters of interest, such as the level of polygamy in the population.

In Chapter 2, we present a simulated annealing method for estimating pedigrees in large samples of otherwise seemingly unrelated individuals using genome-wide single nucleotide polymorphism (SNP) data. The method supports complex pedigree structures such as polygamous families, multi-generational families, and pedigrees in which many of the member individuals are missing. Computational speed is greatly enhanced by the use of a composite likelihood function which approximates the full likelihood. We validate our method on simulated data and show that it can infer distant relatives more accurately than existing methods. Furthermore, we illustrate the utility of the method on a sample of Greenlandic Inuit.

In Chapter 3, we extend the method discussed in Chapter 2 to develop a Bayesian method to jointly estimate pedigrees and effective population size, N_e, from genetic markers using Markov Chain Monte Carlo. Similar to the simulated annealing method in Chapter 1, the MCMC method supports analysis of a large number of markers and individuals with the use of a composite likelihood. We show on simulated data that our method is able to jointly estimate relationships up to first cousins and N_e with high accuracy. We also apply the method on a real dataset of house sparrows to reconstruct its previously unreported pedigree.

In Chapter 4, we zoom out from looking at close relationships and focus on the relationship between Neanderthals and anatomically modern humans through admixture. More specifically, we develop a method based on a Hidden Markov Model to infer segments of Neanderthal ancestry, or admixture tracts, in human genomes. Analogous to using pedigree structures to infer population parameters, we use the lengths of admixture tracts to estimate various parameters for the admixture event, such as admixture time and the proportion of genetic material contributed by Neanderthals into humans. We then apply our method on samples of ancient humans to estimate the admixture time between Neanderthals and humans.

Chapter 2

Composite Likelihood Method for Inferring Local Pedigrees

2.1 Introduction

Pedigree information is used in many areas of genetic analysis, including discovery of disease-related markers in co-segregation analysis and family-based association studies [50], pedigree-informed haplotype and genotype imputation [43], and in estimating variance components for quantitative traits (e.g. heritability) [72]. At the population level, pedigrees can elucidate the social organization and behavior of a group, such as mating patterns and variance in reproductive success among individuals [8]. Furthermore, pedigree information can be used to infer population parameters such as migration rates between subpopulations at very recent time scales. Most population genetic inference methods are based on coalescence theory, which models the genealogical relationships among samples of genetic data at a time scale of N generations, where N is the effective population size. However, standard coalescence models, such as Kingman's coalescent [35, 33, 34] ignore pedigree structure. Simulation studies have shown that the coalescent is a poor approximation of the genealogical process over short time frames ($< log_2N$ generations, where N is the population size), potentially leading to inaccurate inferences at these time scales [76, 75]. Therefore using the pedigree, which contains more detailed information about the genealogical history of the samples, should provide more power in inferring population parameters for the very recent past. [3]

Considerations of pedigree structure is becoming increasing relevant as the size of population genetic samples increases, as these samples may have an increasing probability of including cryptic relatives. The likelihood of seeing cryptic relatives in population samples depends on the sample size, effective population size, and breeding structure. For example, Moltke [49] found that due to the small population size in Greenland, even a relatively small sample size of 584 Inuits contained many close relatives, and about half of the samples had to be removed to form an unrelated set. Other examples include the HapMap Phase III data in which Pemberton [51] found 166 pairs of cryptic close relatives (i.e. third degree relatives or closer) among the sample population of about 1400; and the San Antonio Family Studies in which Sun [66] found 4 cryptic relative pairs among 154 putatively unrelated samples. Performing association studies on samples harboring cryptic relatedness may result in spurious associations [74]. In such cases, pedigree information can be used to remove related samples or explicitly model relatedness to increase the power of association studies [2].

Pedigree information is undoubtedly valuable. In many cases, however, pedigrees are not directly observable and must be inferred from genetic data, which is the topic of this paper. However, we note that using estimated pedigrees as a replacement for known pedigrees may not be an optimal procedure in many cases, if the statistical uncertainty in the estimation of the pedigree is ignored. For example, the consequences of using estimated pedigrees in linkage analyses are largely unknown and we warn against the use of such methods without further studies of their properties.

Although numerous pedigree inference methods have been developed to date, most are limited to inferring very close relationships or require a prior knowledge of the sample structure. Many existing methods support only single- or two-generation samples. The single-generation methods are sibship inference algorithms which partition the sampled individuals into sibship clusters [4, 61, 69, 78]. The parentage inference methods for two generations find the best parent-offspring combinations from a set of offspring and candidate parents [23, 80, 82]. Several methods that can support more than two generations have been developed [5, 7, 13, 12, 25, 38, 56]. But they are either limited in the number of markers that can be analyzed [7, 56]; do not support polygamous pedigrees [25, 38]; assume a complete sample (i.e. every member in the pedigree is sampled) [13, 12, 14]; or assume all sampled individuals belong to a single generation [25, 38]. The state-of-the-art method, PRIMUS [63], is the most flexible of the existing methods; it accommodates missing data and is able to infer multi-generational, polygamous pedigrees. Although PRIMUS is a notable improvement from other methods, its accuracy decreases significantly as the number of missing individuals increases. This is problematic as we expect samples to contain only a small fraction of pedigree members unless the sample represents a large portion of the total population or is specifically designed to include close family members. Extending the work of PRIMUS, PADRE [62] connects PRIMUS-reconstructed family networks to estimate distant relatives. However, PADRE estimates only the degree of relationship between the founders connecting the family networks, which is not equivalent to estimating the pedigree.

The difficulty in pedigree inference comes from three sources. First, the number of possible pedigrees is enormous even for a small sample size [64, 68], making naive enumeration of pedigrees in search for the best one infeasible. Second, computing the likelihood of a pedigree is very expensive. Algorithms for computing the likelihood of a pedigree are either exponential in the number of loci [16], or in the number of individuals [40], which makes the likelihood computation of large pedigrees at many loci prohibitively slow. Finally, inference of pedigree relationships from genetic relationships, measured by the proportion of the genome shared by identical-by-descent (IBD), has high uncertainty. As the pedigree relationship between two individuals becomes more distant, the coefficient of variation and the magnitude of skew in genome sharing become larger [28]. For example, the distribution of genome sharing between second cousins overlaps significantly with that of third cousins, making these two pedigree relationships difficult to distinguish based on pairwise genome sharing alone.

In this chapter, we present CLAPPER (Composite Likelihood Approach to Pedigree Reconstruction), a method that estimates the unknown pedigree from the genotype data of a sample of individuals. Note that our parameter of interest is the pedigree, which is not equivalent to the set of all pairwise relationships. In fact, pairwise relationships do not necessarily define a unique pedigree. Our new inference method addresses the drawbacks of the existing methods. More specifically, our method can utilize many markers genome-wide, support multi-generational pedigrees (up to 5 generations) and polygamous reproduction, and allows many missing individuals in the sample. We assume that all individuals are outbred and that the pedigrees do not create cycles, except in the case of full-sibs. To increase computation efficiency, we use a composite likelihood to approximate the full likelihood based on pairwise likelihoods, and use simulated annealing as a heuristic optimization algorithm for maximizing the composite likelihood. We validate our method on simulated data and show that it outperforms existing methods for inferring distant relatives. Furthermore, we demonstrate our method's application to real data on a sample of Greenlandic Inuit.

2.2 Materials and Methods

Composite Likelihood

CLAPPER is based on the idea of forming a composite likelihood function based on marginal likelihood functions calculated for pairs of individuals. While even pairwise likelihoods are slow to calculate for full genomic data, they can be tabulated and stored in computer memory. It is thereby possible to estimate pedigrees, based on a composite likelihood function, by only calculating the likelihood function between pairs of individuals once. This makes our method potentially applicable to large data sets containing thousands of individuals. As we will later discuss, using some heuristics, the method may even be applicable to large GWAS data sets.

We define a pedigree as undirected graphs where a node represents an individual and an edge represents a parent-offspring relationship. Each individual has a sex and is associated with 0, 1 or 2 edges connecting the individual to its parents, which must be of different sexes if the individuals has two identified parents (See Section 2.2 for more detail). An individual in the pedigree may or may not be represented in the sample, but if individual i is represented in the sample it is associated with genotype vector, X_i.

For each pedigree, the set of k sampled individuals is denoted by H, and the composite likelihood for such a pedigree is defined as

$$CL(H) = \begin{cases} P(X_i), & \text{if } k = 1 \\ \frac{\prod_{(i,j)\in H} P(X_i, X_j | R_{i,j})}{\prod_{i \in H} P(X_i)^{k-2}}, & \text{otherwise} \end{cases} \tag{2.1}$$

where $R_{i,j}$ is the relationship between i and j induced by the pedigree. For a pedigree consisting of one individual, the likelihood is simply the probability of the individual's observed genotypes. For $k > 1$ the composite likelihood is obtained as the product of marginal pairwise likelihoods. However, to obtain a more natural scaling of the composite likelihood we note that the probability of the data for each individual has been calculated $k - 1$ times and we therefore divide the composite likelihood function with the marginal likelihood of each individual $k - 2$ times. This has several desirable properties such as convergence of the composite likelihood to the true likelihood as the relatedness among individuals goes to zero. Another way to think of this composite likelihood function is in terms of products of conditional likelihoods. We can factor the full likelihood as

$$P(X_1, \cdots, X_k|H) = P(X_1)P(X_2|X_1, H) \cdots P(X_k|X_1, \cdots, X_{k-1}, H).$$

Since computing the conditional likelihoods $P(X_i|X_1, \cdots, X_{i-1}, H)$ is difficult, we approximate them with

$$P(X_i) \prod_{j=1}^{i-1} \frac{P(X_i|X_j, H)}{P(X_i)}.$$

That is, we multiply the marginal probability of our current observation $P(X_i)$ by the likelihood ratio $\frac{P(X_i|X_j,H)}{P(X_i)}$ for each previous observation X_j. If the previous observation informs our current observation, then $\frac{P(X_i|X_j,H)}{P(X_i)} \neq 1$, so the likelihood of the current observation increases or decreases accordingly. Using this approximation, we arrive at (2.1). Note that $P(X_i|H) = P(X_i)$ since $P(X_i)$ is simply the likelihood of observing the genotypes X_i, which is independent of the pedigree, H.

The pairwise likelihood $P(X_i, X_j|R_{i,j})$ can be computed efficiently using the Hidden Markov Model (HMM) approximation by [3], which is used in this study. However, we note that any other definition of the pairwise likelihood function could have been used. For a set of possible outbred relationships in a 5-generation pedigree (see Types of Pairwise Relationships), the pairwise likelihood for each pair (i, j) is precomputed and stored in memory.

The total pre-computation time for $\binom{n}{2}$ pairs of individuals, s types of relationships, and L loci, therefore, is $O(n^2sL)$. Since the composite likelihood of a pedigree is a simple function of the pairwise and marginal likelihoods, it can be computed fast by accessing the precomputed values stored in memory. The full composite likelihood for a set of local pedigrees is then computed by taking the product of the composite likelihood for each local pedigree.

It is worthwhile to note alternative ways to construct a composite likelihood. Another, perhaps more intuitive, formulation that also ensures that the composite likelihood converges to the true likelihood as the relatedness among individuals goes to zero, is

$$\prod_{i \neq j} P(X_i, X_j)^{\frac{1}{n-1}}, \tag{2.2}$$

which scales the product of pairwise likelihoods by $\frac{1}{n-1}$ to account for the multiple counting of each sample. However, as we will discuss in the Results section, this formulation leads to a worse approximation of the full likelihood function.

Types of Pairwise Relationships

Tables 2.1 summarizes the pairwise relationships supported by our method. Pairwise relationships can be represented by the paths between the two individuals in the pedigree. Without loss of generality, let the age of the first individual be less than or equal to the age of the second individual. Then a pairwise relationship is defined by the number of unique paths connecting the two individuals (k); the number of meiosess between the first individual and the most recent common ancestor (MRCA) of the two individuals ($m1$); and the number of meioses between the second individual and the MRCA ($m2$). For example, pairwise relationship with $k = 2$, $m1 = 2$, and $m2 = 2$ corresponds to full first cousins. Note that in direct ancestor-descendant relationships, the second individual acts as the MRCA. For instance, pairwise relationship with $k = 1$, $m1 = 1$, and $m2 = 0$ corresponds to the parent-offspring relationship. Finally, an unrelated relationship corresponds to $k = 0$, $m1 = 0$, and $m2 = 0$.

Also given in Table 2.1 are the total number of meioses between the two individuals (α) and the Jacquard coefficients (w_1, w_2) [30]. Here, $w1$ and $w2$ denote the probability of two individuals having one or two pairs of alleles, respectively, identical-by-descent (IBD) at a random locus.

Representing Pedigrees as a Graph

In a pedigree graph, a node represents an individual and an edge represents a parent-offspring relationship. Each node has a set of features: sex (male, female, or unknown), sample status, and depth. As shown in Figure 2.1, the unshaded nodes represent ghost individuals for whom we do not have genotype data (unsampled); the shaded nodes represent individuals for whom we have genotype data (sampled). To represent the pedigree compactly, an unsampled individual is represented only if it connects at least two sampled individuals. For example, individual C in Figure 2.1 has one unsampled parent that connects it to A and B. But the other parent of C is not represented since it does not connect two or more sampled individuals. In addition, each individual belongs in a particular depth. For example, individuals A and B belong in depth 1 and individual C belongs in depth 0.

When we explore pedigrees, we first check whether the configuration in question is a valid pedigree graph. For our method, we restrict ourselves to outbred, non-cyclic pedigrees (except cycles formed by full siblings). We say that a pedigree is valid if all of the following conditions are satisfied:

Table 2.1: Cousin-like relationships. Each row represents a unique pairwise relationship.

k	m_1	m_2	α	w_1	w_2
1	1	1	2	1/2	0
1	2	1	3	1/4	0
1	2	2	4	1/8	0
1	3	1	4	1/8	0
1	3	2	5	1/16	0
1	3	3	6	1/32	0
1	4	1	5	1/16	0
1	4	2	6	1/32	0
1	4	3	7	1/64	0
1	4	4	8	1/128	0
2	1	1	4	1/2	1/4
2	2	1	5	1/2	0
2	2	2	6	1/4	0
2	3	1	6	1/4	0
2	3	2	7	1/8	0
2	3	3	8	1/16	0
2	4	1	7	1/8	0
2	4	2	8	1/16	0
2	4	3	9	1/32	0
2	4	4	10	1/64	0
1	1	0	1	1	0
1	2	0	2	1/2	0
1	3	0	3	1/4	0
1	4	0	4	1/8	0
0	0	0	4	0	0

1. Each node has zero, one, or two parent nodes. If the node has two parents, the two parents have opposite sexes.

2. Parents of a node belong in the same generation.

3. The pedigree does not create loops except in the case of full siblings. For example, double first cousins are not allowed. Inbreeding is also not allowed.

4. The pedigree does not span more than a maximum number of generations. For this report, the maximum number of generations is five.

5. An ancestor node is older than its descendants, if age information is available.

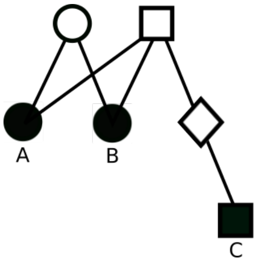

Figure 2.1: Pedigree graph. The shapes indicate the sex of the node (circle=female, square=male, diamond=unknown) and the color indicates whether the individual was sampled (shaded=sampled, unshaded=unsampled).

Simulated Annealing

Because the number of possible pedigrees grows very rapidly with sample size, an exhaustive search for the most likely pedigree is infeasible for even a moderate number of individuals. Therefore, we use simulated annealing [36] to maximize the composite likelihood function. In this algorithm, a perturbation of the pedigree is generated by locally modifying the edges and nodes of the current pedigree. We explore the pedigrees with high likelihoods by always accepting proposals with higher likelihoods and occasionally accepting those with lower likelihoods to avoid getting stuck in local maxima. We implemented 22 different perturbations (moves) detailed in Appendix A. These moves can be broadly categorized into three classes. The first class of moves involves choosing two individuals and modifying their pairwise relationship. These moves include transitions between: parent-offspring and full siblings; parent-offspring and half siblings; uncle-nephew and nephew-uncle; grandparent-grandchild and half siblings; and full siblings and self. Related to these are moves that add or subtract an edge between two nodes. For example, adding an edge causes parent-offspring relationships to become grandparent-grandchild relationship, whereas subtracting an edge has the opposite effect. The motivation for this class of moves is that these pairs of relationships have similar IBD coefficients, hence similar likelihoods. So these perturbations allow transitions between pedigrees with similar likelihoods.

The second class of moves allows bigger perturbations in the current pedigree. These moves include splitting a pedigree into two, joining two pedigrees into one, or the combination of splitting and joining. Splitting a pedigree can be done in two ways: we can either detach a chosen individual's sub-pedigree (i.e. its descendant and itself) from its ancestors, or split off a randomly selected subset of its children to form a new pedigree. Joining two pedigrees involves creating a common ancestor between two individuals that belong to different local pedigrees.

The last class of moves is designed to transition between similar pedigrees when sex or age information is missing. For example, one move allows an individual and its descendant to swap places if age information is not present to resolve the directionality of the relationship. Another move changes the sex of an individual if sex information is not available, which in turn switches the sex of its potential spouses.

All of these transitions modify a small part of the current pedigree to generate a new configuration. Since the composite likelihood is a function of the pairwise and marginal likelihoods, the likelihood of the new configuration can be computed fast by adjusting the old likelihood by the changes made to the modified part of the pedigree.

The outline of the simulated algorithm is described below:

Initialization: Let each individual be a singleton pedigree (i.e. everyone is unrelated). Compute and store the composite likelihood of the current configuration.

Recursion:

1. Choose one of the 22 moves at random and generate a new configuration accordingly.
2. If the new configuration is an invalid pedigree, reject and go back to step 1. If it is a valid pedigree, compute the composite likelihood $CL(H_{new})$ for the new configuration. Accept with probability $min[(CL(H_{new})/CL(H_{old}))^t, 1]$, where t is the annealing temperature.
3. Repeat steps 1-2 C times.
4. Decrease the temperature to t/f, where $f > 1$ and go to step 1.

Termination: Terminate after I iterations or when the change in composite likelihood is less than e.

The tuning parameters C, f, I, and e were optimized to achieve a balance between convergence and computational efficiency using a number of trial runs on different simulated data sets. Table 2.2 shows an example of the composite likelihood score at different stopping times determined by the maximum number of iterations. We run multiple instances of the algorithm with different random seeds. The algorithm then reports the pedigree with the highest likelihood encountered among all runs.

Table 2.2: Composite Likelihood Convergence. Composite likelihood score at various stopping times given by I for a particular instance of simulation B.

Maximum Number of Iterations (I)	Composite Likelihood
5×10^5	-113131
1×10^6	-113120
1.5×10^6	-113117
2×10^6	-113117
2.5×10^6	-113117

Background Relatedness

Since the composite likelihood function is based on pairwise likelihood values, any inference based on it is limited by the quality of the pairwise likelihoods. One important factor that confounds the likelihood computation is linkage disequilibrium (LD), which often causes relationships to be overestimated [67]. As shown in Figure 2.2, unrelated pairs of individuals often have higher likelihoods for being distantly related, which leads to false detection of relatives. The method of [3] attempts to correct for LD by conditioning on nearby markers. However, in our experience residual effects of LD will still tend to bias inferences when markers are in high LD. One way to further reduce the effects of LD is pruning, or thinning, of markers. However, there is no consensus on how best to choose a set of markers that contains minimal LD and yet harbors enough information to detect distant relatives. To get a better sense of the effects of LD pruning on relationship inference, we simulated various pairwise relationships (i.e. second cousins, third cousins, unrelated) at linked loci. We pruned the markers based on LD in 100 unrelated founders and measured the pairwise prediction accuracy for the test pair. We repeated this procedure under different levels of LD pruning to choose an appropriate level of pruning threshold (See Results).

In addition to LD pruning, we further controlled for false detection of relatives by adding a regularization term to the composite likelihood. The regularizer was designed to weight against individuals from forming family clusters, motivated by the fact that in large data sets there are so many potential pedigree relationships for each individual, that most individuals will be inferred to have some pedigree relationship to at least one individual in the sample, even when they are unrelated. This is essentially a multiple testing problem in which an increasing number of individuals in the sample implies a reduced probability of inferring an individual to be unrelated to all individuals in the sample. There are natural ways of addressing this problem in a Bayesian framework that we might also be able to appeal to in the current framework. In particular, we will assign a probability distribution on the number of local pedigrees inferred. More specifically, we used the regularized composite likelihood

$$CL^*(X) = CL(X)Pr(Q = q)^{\beta}, \tag{2.3}$$

where q is the number of local pedigrees and $\beta > 0$. We chose a Poisson distribution with mean n, the sample size, as the distribution of Q. This regularization is conservative in the sense that it favors every individual to remain a singleton unless there is strong evidence otherwise. Our choice to use the Poisson distribution was made, in part, for computational convenience but, as we will discuss in the Results section, resulted in good statistical properties of the method.

Simulated Dataset

We tested the performance of our method on simulated pedigrees. We generated human autosomal haplotypes using msprime [32] with effective population size of 10,000, average

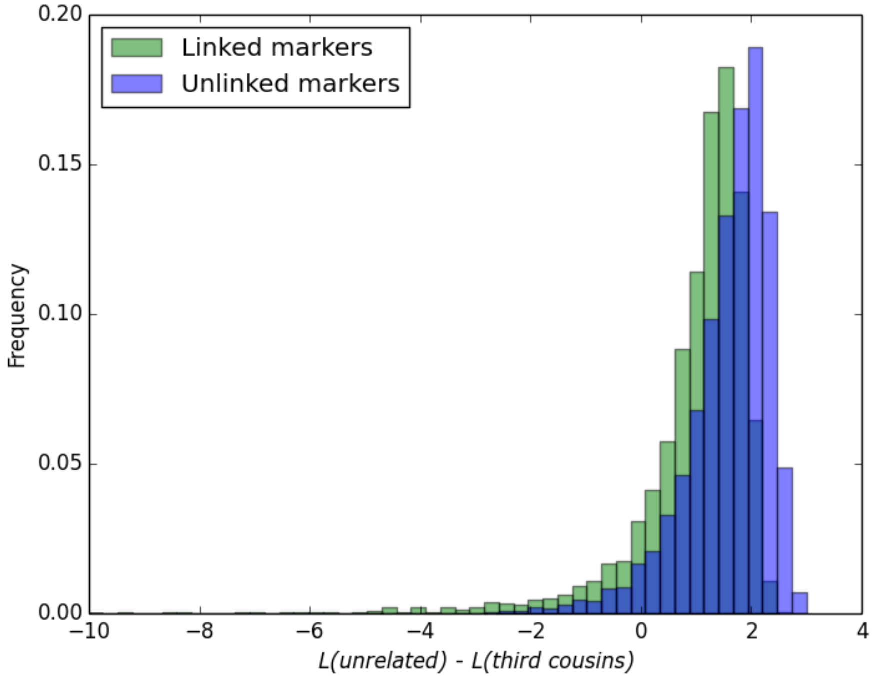

Figure 2.2: Effects of LD on Relatedness Estimation. The figure shows the histogram of the log likelihood difference, $L(\text{unrelated}) - L(\text{third cousins})$, when the true relationship is unrelated. Unrelated pairs often have higher likelihoods for being third cousins when LD is present in the data, as shown by the histogram corresponding to linked markers. The data were simulated with msprime and the likelihoods were computed using RELATE.

recombination rate of 1.3e-8, and mutation rate of 1.25e-8. Using these founder haplotypes, we simulated four pedigree structures shown in Fig 2.3.

Simulation A consisted of 10 singletons and a 45-person family that spanned 5 generations. Of the 45 family members, 10 were sampled and 35 were missing. The kinship coefficients of the sampled relative pairs ranged from 1/4 (e.g. full siblings) to 1/256 (e.g. third cousins). Simulation B was designed to study the performance of our method on smaller family clusters. It consisted of 4 family clusters and 4 singletons. Each family cluster contained 15 to 18 members, of which only 4 of them were sampled. The sampled individuals spanned multiple generations and formed pairwise relationships with kinship coefficients ranging from 1/4 to 1/256. Simulation C was designed the test the method on pedigree structures in which every sampled individual, excluding singletons, has at least one close relative in the data. It consisted of 9 singletons and a 16-person pedigree that spanned 5 generations. The 16-person pedigree contained 7 missing individuals and 9 sampled individuals, where each sampled individual formed a parent-offspring relationship with at least one other sample. Finally, simulation D was designed to test the method on a pedigree that

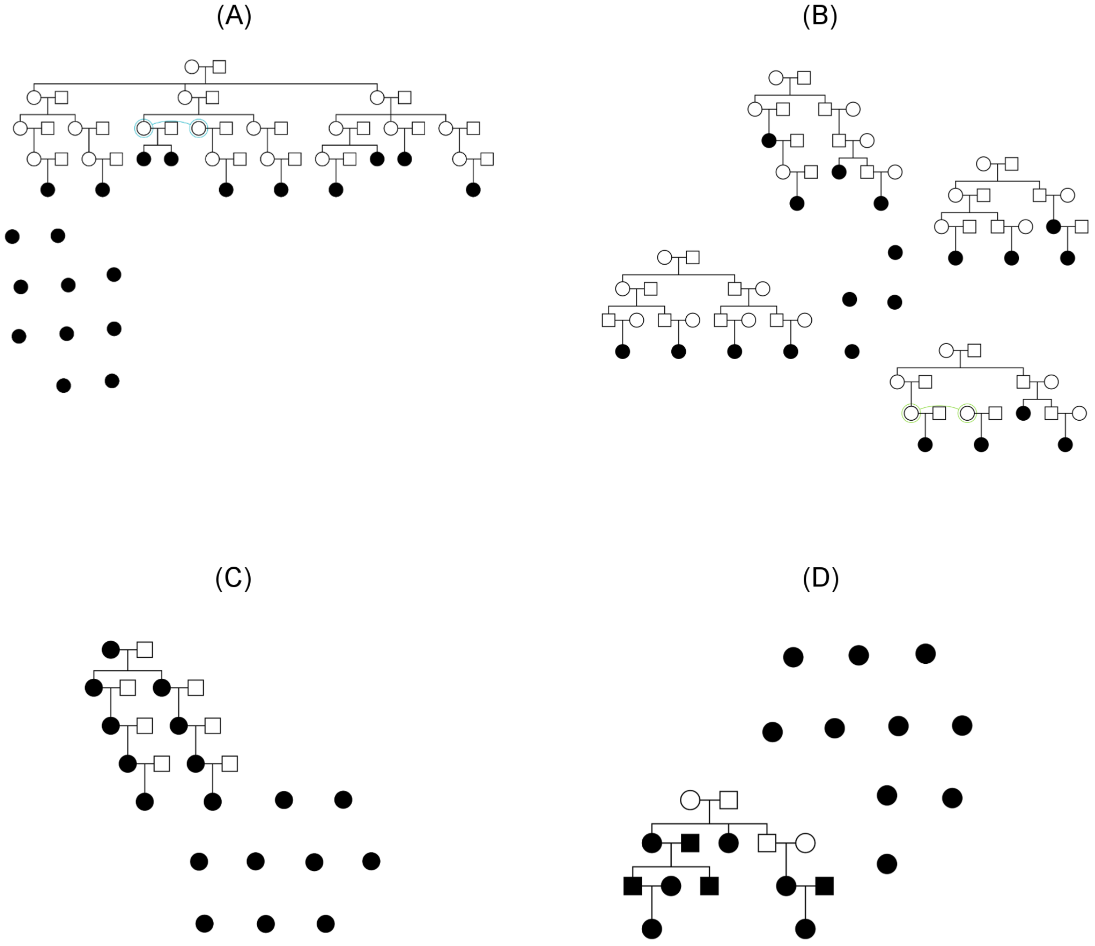

Figure 2.3: **Simulated pedigrees.** Shaded nodes indicate sampled individuals for which we have genotype data and unshaded nodes indicate unsampled individuals. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

is relatively easy yet more difficult to infer than simulation C. Whereas every sample was connected by parent-offspring relationships in simulation C, some samples in simulation D were connected only by an avuncular relationship.

Each simulation scenario was replicated 100 times. For each sampled individual, we simulated genotyping error by switching each allele to the alternate allele with probability 0.01. To reduce the level of LD among markers, we used PLINK [9] to prune the original set of markers at $r^2 = .05$, resulting in about 10,000 markers. The sex of each sample was assumed known, whereas the age was assumed unknown.

Inbred Pedigrees

The current version of CLAPPER does not support inbred pedigrees and rejects any such pedigrees during its search algorithm. To test how CLAPPER performs on inbred pedigrees, we simulated two inbred individuals who form a first cousin relationship. In the first scenario, the two individuals are inbred via their grandparents who form a first cousin relationship (Figure 2.4A). In the second scenario, the grandparents form a full sibling relationship (Figure 2.4B).

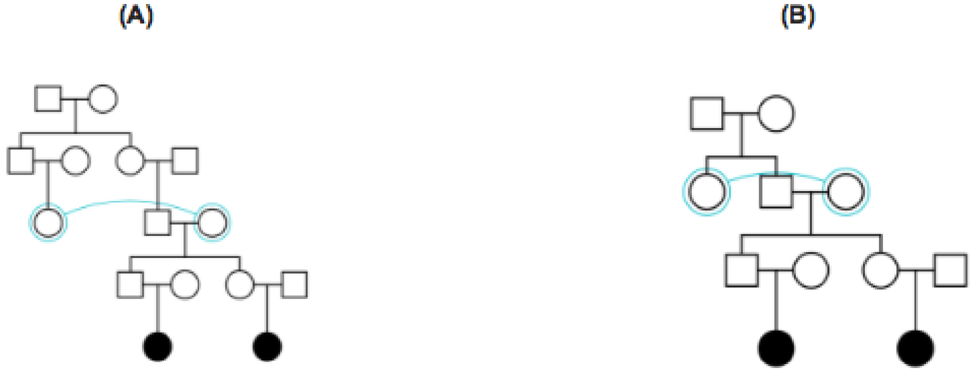

Figure 2.4: Simulated inbred pedigrees. The shaded nodes indicate sampled individuals and unshaded nodes indicate unsampled individuals. The nodes connected by a blue line indicate that the two nodes are the same. (A) Two sampled first cousins whose grandparents are first cousins. (B) Two sampled first cousins whose grandparents are full siblings.

For each scenario, we ran our method on 100 independent datasets. We allowed the two sampled nodes to be in any generation.

Empirical Dataset

We applied our method to reconstruct the previously unreported pedigrees of 100 individuals in Tasiilaq villages in Greenland which had been genotyped [49] using the Illumina CardioMetaboChip, consisting of 196,224 SNPs. Since the European admixture into the Greenlandic population can confound relationship inference, we selected individuals from Tasiilaq villages, which showed one of the lowest levels of European admixture in the sample. In particular, the 100 individuals we selected were estimated to have European admixture proportion of 5 percent or less. To reduce the effects of LD, with pruned the markers using PLINK at $r^2 = 0.05$. Due to the unusually high level of LD in the Greenlandic population, we were left with 2173 SNPs after LD-pruning.

Competing Methods for Comparison

We compared the performance of our method on simulated data to PRIMUS (v1.9.0), arguably the state-of-the-art pedigree reconstruction method. Although many pedigree inference methods exist, we chose to use PRIMUS as a benchmark since it is the most flexible of the existing methods in the types of pedigrees it can infer. More specifically, PRIMUS supports the inference of multi-generational, polygamous pedigrees and allows for missing individuals. PRIMUS reconstructs pedigrees that are consistent with pairwise IBD estimates and reports high-scoring configurations.

To estimate the pairwise IBD coefficients for the simulated data, we used two different methods: PLINK and RELATE [3]. To use PLINK, we first estimated the population allele frequencies from 100 founder individuals. We then used PLINK to estimate the IBD coefficients for the individuals in our simulated pedigrees, where the population allele frequency estimates were provided as input. This mimics the inference procedure recommended in the PRIMUS documentation. A similar procedure was used to run RELATE to estimate the pairwise IBD proportions (Appendix B). The IBD estimates were then used by PRIMUS to reconstruct likely pedigrees. We denote the combined method of PLINK and PRIMUS as PP, and Relate and PLINK as RP. Since PRIMUS was designed to reconstruct pedigrees where samples are connected by third-degree relationships or closer, we applied PP and RP only to simulations C and D.

We used PADRE [62] for simulations A and B, where PRIMUS was inappropriate to use due to the presence of samples connected only by distant relationships. PADRE takes as input relationship likelihoods by ERSA[29] and output by PRIMUS, and reports the degree of relationship for each pair of samples. To generate the results by PRIMUS, we used PP and RP as described before. ERSA uses estimates of IBD segments to compute the pairwise relationship likelihoods. Since RELATE was used to compute the pairwise likelihood of IBD proportions for CLAPPER, we used RELATE also to estimate the pairwise IBD segments to generate the input for ERSA. We denote the combined method of PP and PADRE as PPP, and RP and PADRE as RPP. The command lines used for running the softwares are provided in Appendix B.

Recall that CLAPPER maximizes a statistic that incorporates both the likelihood score and the number of family networks (2.3). In PP and RP, however, all reported pedigrees have the same number of family networks, which makes maximizing both the likelihood score and the number of family networks equivalent to maximizing the likelihood score alone. The same is true for PPP and RPP, which report a single best estimate of family networks.

We also compared our method to the pairwise inference method. In this method, we used RELATE to compute the pairwise likelihood under each possible relationship [S1 TABLE] for all pairs of individuals. Then we assigned each pair the relationship with the highest pairwise likelihood. We controlled the false positive rate by multiplying the likelihood of being unrelated by a scalar $c > 0$, in order to provide comparable results between methods. The pairwise inference method produces only the best relationship for each pair, which may not result in a valid pedigree when all pairwise relationships are pieced together. Still, it serves as a useful benchmark to evaluate the accuracy of pairwise predictions by our method.

Measuring the Error Rate

We measured the performance of our method in two ways: the frequency of estimating the true pedigree; and the distance between the estimated pedigree and the true pedigree in terms of pairwise relationships. We note that since CLAPPER does not consider inbred pedigrees whereas PP and RP do, we pre-processed the output of PP and RP before measuring the error rate to make a fair comparison. More specifically, we removed all inbred pedigrees from the output of PP and RP and measured the error rate using just the remaining pedigrees.

Frequency of Estimating the True Pedigree Configuration

We say that the estimated pedigree is correct if there is a one-to-one mapping between the nodes of the estimated pedigree and the nodes of the true pedigree such that each edge in the estimated pedigree has a corresponding edge in the true pedigree. Note that for PP and RP, which potentially report multiple highest-scoring pedigrees, we say that the estimated pedigree is correct if the true pedigree is in the set of highest-scoring pedigrees.

Pairwise Error Rate

To measure the error rate of the pairwise method, which estimates pairwise relationships directly, we compared the true relationships to the estimated relationships. Therefore, we define the error rate for each pair as

$$e = \begin{cases} 0, & \text{if } \hat{w}_1 = w_1 \text{ and } \hat{w}_2 = w_2 \\ 1, & \text{otherwise} \end{cases}$$

where w_i is the probability that two individuals share i pairs of alleles IBD at a random locus under the true relationship; and $\hat{w}_i$ is the corresponding probability for the estimated relationship. In other words, the estimated relationship is correct if its three Jacquard coefficients [30] are exactly the same as those of the true relationship.

Furthermore, to measure the distance between the estimated relationship and the true relationship for each pair, we computed the kinship coefficient distance

$$d = \frac{|\hat{\phi} - \phi|}{\phi},$$

where $\hat{\phi} = \frac{1}{4}\hat{w}_1 + \frac{1}{2}\hat{w}_2$ and $\phi = \frac{1}{4}w_1 + \frac{1}{2}w_2$.

We also used e and d to measure the pairwise error rate of CLAPPER, where the inferred pairwise relationships are those induced by the estimated pedigree, and the true pairwise relationships are those induced by the true pedigree. For PP and RP, which report all pedigrees with high likelihood scores, we computed the error rate by taking the average across all highest-scoring pedigrees. For PPP and RPP, which report a single best degree of relationship for each pair, we measured the error rate by e and d as defined above.

2.3 Results

Behavior of the Composite Likelihood

To examine the behavior of the composite likelihood, we simulated a nuclear family with two parents and their four children at 3,000 independent loci. We then computed the likelihood of the data under various pedigree configurations, ranging from the pedigree in which no one is related to the true pedigree. For each pedigree configuration, we computed the likelihood value with three different formulas: the full likelihood using MERLIN[1], composite likelihood A, given by (2.2), and composite likelihood B, given by (2.1).

The comparison of the three likelihood formulas are shown in Figure 2.5. The x-axis is the distance of the test pedigree to the true pedigree, measured by the proportion of pairwise relationships that are correct in the test pedigree. As expected, the full likelihood increases as the test configuration becomes closer to the true pedigree. Both composite likelihood formulas preserve the ordering of the pedigrees induced by the full likelihood. That is, the order of pedigrees from the least likely to the most likely based on the full likelihood corresponds to the ordering based on the composite likelihood formulas. Although both composite likelihood formulas preserve this ordering, the likelihood surface given by (2.2) is much flatter than the full likelihood, whereas the likelihood surface of (2.1) is roughly on the same order of magnitude as the full likelihood.

Effects of Linkage Disequilibrium on Pairwise Relationship Inference

As mentioned in the Methods section, we examined different thresholds for LD pruning. The appropriate level of pruning depends both on the genome length and the types of relationships we want to infer accurately. As shown in Fig 2.6, there is a trade-off between keeping enough markers to estimate distant relationships and removing markers to reduce false detection of relatives. For unrelated pairs, the most stringent LD pruning we tested ($r^2 = .025$) showed the best relationship prediction accuracy. For third cousin relationships, however, pruning the markers too severely caused too much information loss, leading to a decrease in prediction accuracy. A similar pattern is observed for the second cousin relationships. For our simulated and empirical data, we prune the markers at $r^2 = .05$, which according to our simulations, retained enough information to estimate second and third cousins while keeping the false positive rate (i.e. estimating unrelated pairs as related) relatively low. We note that finding optimal strategies for dealing with background LD when inferring relatedness is an important topic that merits further research.

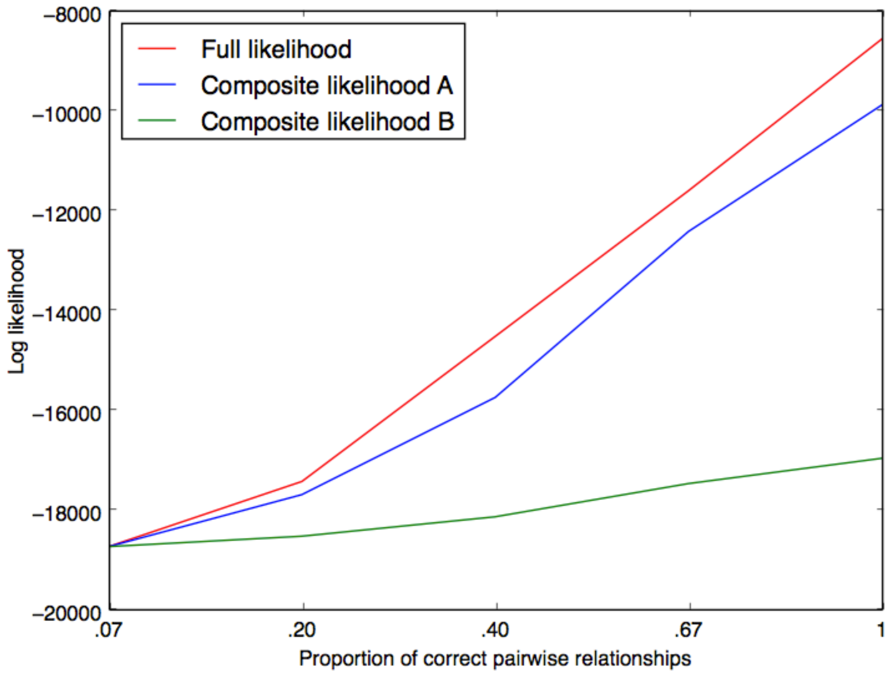

Figure 2.5: Comparison of Various Likelihood Formulas on Simulated Data. The x-axis measures how close the test pedigree is to the true pedigree; the test pedigree becomes closer to the truth from left to right. In this simulation, the composite likelihood given by (2.1) approximates the full likelihood more closely than (2.2).

Estimating Simulated Pedigrees

Table 2.3 summarizes the frequency of estimating the true pedigree and the average number of best pedigrees reported by each method. For simulation C, where all samples were connected by parent-offspring relationships, CLAPPER was able to find the true pedigree in all 100 experiments. This showed that when the sampled individuals are connected by very close relationships, CLAPPER can unambiguously find the correct pedigree. Similarly, RP inferred the true pedigree as the single best estimate in 96 out of 100 experiments. The remaining 4 experiments did not output any pedigrees because all likely pedigrees exceeded the maximum number of generations we imposed (5 generations). On the other hand, PP showed a lower accuracy rate than both CLAPPER and RP. Several experiments finished with errors due to too large a number of likely pedigrees to process, while some only produced inbred pedigrees. However, the true pedigree was estimated in the majority of the experiments that finished successfully.

For simulation D, all methods had a lower accuracy rate for estimating the true pedigree compared to simulation C. Some of the samples in this scenario were connected only through

Table 2.3: Accuracy for estimating the true pedigree.

Simulation	# Reported[a] (CLAPPER)	# Correct[b] (CLAPPER)	# Reported[a] (PP)	# Correct[b] (PP)	# Reported[a] (RP)	# Correct[b] (RP)
C	1	100/100	6	65/76 *	1	96/96 **
D	1	56/100	49	18/79 ***	3	20/100

[a] Number of highest scoring pedigrees reported.

[b] Numerator is the number of times the true pedigree was among the highest scoring pedigrees; denominator is the the number of successful experiments that produced at least one outbred pedigree.

*Excludes 6 runs that finished with errors and 18 runs that did not produce any outbred pedigrees.

**Excludes 4 runs that did not produce any pedigrees.

***Excludes 20 runs that finished with errors and 1 run that did not produce any outbred pedigrees.

an avuncular relationship, which made the inference more difficult than the pedigree given in simulation C. Nonetheless, CLAPPER showed a higher accuracy rate than both PP and RP even though we counted the estimated pedigree as correct if the true pedigree was found in any of the best reported pedigrees for RP and PP. Simulations A and B were omitted from our analysis since they contained samples that were not connected by third degree relationships or closer, which made PP and RP inappropriate to use to estimate the full pedigree.

Fig 2.7-2.8 show the average pairwise error rate across all replicate experiments, categorized by different levels of true relatedness, ϕ. For simulation A, PPP did not finish successfully in 19 out of 100 experiments due to errors encountered in PRIMUS (e.g. too many likely pedigrees to process). Similarly, PPP did not finish successfully in 24 experiments for simulation B. Furthermore, PP and RP encountered errors or did not produce any outbred pedigrees in some experiments (Table 2.3). These experiments were removed from our analyses and are not reflected in Fig 2.7-2.8.

For simulations A and B, all methods had a very low false positive rate (i.e. error rate for $\phi = 0$), and relatively low error rates for estimating close relationships (Fig 2.7A-B). For more distant relatives such as those beyond first cousins ($\phi \leq 1/32$), however, CLAPPER was able to estimate the relationships more accurately than both PPP and RPP. For simulation C, all methods had zero error rates for all relationship categories except PP, which showed a nonzero false positive rate (Fig 2.7C). For simulation D, CLAPPER outperformed RP across all relationship categories, but had a lower accuracy rate than PP in many relationship categories. However, PP showed a significantly higher false positive rate than CLAPPER (Fig 2.7D).

Furthermore, Fig 2.8 shows that even when the estimated relationship by CLAPPER is wrong, it is generally close to the true relationship. For example, the median error rate for $\phi = 1/128$ was 0.5, which is equivalent to estimating second cousins once removed as third cousins. Overall, the median error rate of CLAPPER was equal to or lower than that of the competing methods across all relationship categories.

CLAPPER also performed considerably better than the pairwise inference method. The likelihoods in the pairwise prediction were weighted so that its false positive rate roughly matched that of our method. Fig 2.9 show that at similar false positive rates, our method estimated pairwise relationships with a greater accuracy than the pairwise method across almost all relationship categories. Fig 2.10 further demonstrates that our method has a significant advantage over the pairwise prediction method in detecting relatives. If the purpose of relationship inference is to find relatives–to discover the number of family clusters present in the data, for example–Fig 2.10 demonstrates that our method is able to detect relatives far more accurately than the pairwise method. These figures show that even though our method and the pairwise inference method both use the same pairwise likelihood values to estimate relationships, leveraging information from all pairs of relationships improves the inference significantly compared to considering each pair in isolation.

Each experiment was run 3 times with different random number seeds, where each run consisted of 2 million iterations. The runtime of our method depends on many factors, including the number of individuals, the hidden pedigree structure, the number of missing individuals, and the annealing schedule in the simulated annealing algorithm. That said, each run on our simulated data, excluding the pre-computation time for calculating the pairwise likelihoods, took about 9 seconds on 2.5 GHz Intel Core i5 processor.

Effect of Inbreeding

Figure 2.11 shows four different types of pedigrees inferred by our method for the pedigree shown in Figure 2.4A. In all of these pedigrees, the two sampled individuals had the kinship coefficient of 1/16, equivalent to the kinship coefficient of first cousins.

Figure 2.12 shows the inferred pedigree for the pedigree shown in Figure 2.4B. Due to inbreeding, our method inferred that the two individuals had a uncle-nephew relationship, which is one degree of relationship closer than the truth.

These simulations showed that although CLAPPER does not consider inbred pedigrees, it estimates relationships close to the truth. When the level of inbreeding was relatively low such as that shown by Figure 2.4A, the degree of relationship was correctly inferred by CLAPPER. However, when the level of inbreeding was very high such as that shown by Figure 2.4B, CLAPPER estimated that individuals are more genealogically closer than they actually are.

Estimating the Greenlandic Inuit Pedigrees

To demonstrate our method's ability to infer pedigrees in practical applications, we estimated the previously unreported pedigrees of 100 individuals from Tasiilaq villages in Greenland. Because the Greenlandic Inuit population has high levels of LD, only 1868 SNPs remained after pruning the markers at $r^2 = .05$. Our simulation study showed that at this number of SNPs, regularization with $Poi(n)$ caused the error rate for estimating distant relatives ($\phi < 1/32$) to be very high; but using no regularization at all led to a high false positive rate (Figure 2.13). So we chose to use $Poi(n/2)$ as our regularization, which still produced a lower false positive rate, yet performed better in inferring distant relatives on simulated data.

We ran our algorithm 5 times with different random number seeds, resulting in 5 pedigrees estimates. The top three estimates with the highest composite likelihood scores were within 1.2 likelihood units of each other. The other two estimates were both about 20 likelihood units away from the top three. The inconsistency of the multiple runs was likely caused by the existence of multiple local peaks on the likelihood surface, which makes finding the global optimum difficult in our heuristic optimization. Each run, which consisted of 80 million iterations, finished in about 24 minutes on 2.5 GHz Intel Core i5 processor. Figure 2.14 shows the estimated pedigree drawn by PhenoTips [21]. The reconstructed pedigree consisted of 38 singletons and 8 non-singleton family clusters. Many of these clusters consisted of close relationships such as parent-offspring, full siblings, half-siblings, and avuncular relationships.

Based on our simulations, we expect more than 90 percent of the estimated relationships in these categories to be correct.

2.4 Discussion

In this report, we have shown that the use of composite likelihood allows us to analyze pedigrees containing many individuals at many loci, where computing the full likelihood would be prohibitively slow. Our method can estimate pedigrees when the number of possible pedigrees is too large to enumerate, which is true even for tens of individuals in a multigenerational pedigree. Our method is also one of the very few methods that can support complex pedigree structures such as polygamy, multigenerational pedigrees (up to 5 generations), and missing individuals. In addition, we can incorporate information about sex, age, and the number of generations spanned by the sample to better estimate the pedigree.

We have shown that our method has a significant advantage over the pairwise inference method. It can better estimate relationships beyond first cousins (Fig 2.9) and is able to detect relatives much more accurately (Fig 2.10). The composite likelihood considers all pairwise likelihoods jointly, which in turn can help resolve uncertain relationships in the context of other pairwise relationships. Therefore, even for pairwise relationship inference, where estimating the entire pedigrees may not necessarily be of interest, our method can be used to estimate the relationships more accurately.

Our method also showed an improvement over PRIMUS (PP and RP) and PADRE (PPP and RPP). PRIMUS's reconstruction algorithm relies on accurate pairwise relationship assignments based on IBD estimates. If the sample consists mostly of distant relatives, however, relationship assignment becomes uncertain due to high variance in IBD sharing, which often leads to incorrect pedigree reconstruction. Although our method also relies on pairwise information, we showed that working directly with pairwise likelihood values rather than IBD-based relationships assignments improved the power significantly. Furthermore, PRIMUS's enumeration of possible pedigrees becomes computationally cumbersome as the number of likely pedigrees increases rapidly for a set of distantly related samples. If the data contains many close relationships, however, PRIMUS can reconstruct all likely pedigrees very fast, whereas our method produces a single best pedigree, which may be close but not exactly correct. Thus the performance of each method depends on the sample structure and a suitable method must be chosen accordingly. Similar to PRIMUS, the performance of PADRE depends crucially on accurate estimates of IBD proportions and segments, and poor estimates of either parameter can lead to biases in the relationship inference. We note that IBD estimation is a difficult problem and better estimates of IBD would improve the performance of both PRIMUS and PADRE.

We applied our method on the Greenlandic Inuit dataset to demonstrate its ability infer previously unknown pedigrees from genetic data. Although the estimates of distant relationships are uncertain, we can still get a general sense of pedigree structures hidden in the data and take appropriate actions for downstream analyses. For example, the inferred pedigree can be used to filter out close relatives or model relatedness among samples in association studies. Furthermore, we can validate or improve the estimated pedigree with other evidence such as age.

Pedigree inference based on our composite likelihood is heavily influenced by how well we can compute the pairwise likelihoods. An important factor that affects the pairwise likelihood computation is LD, which often leads to overestimation of relatedness. Although the HMM by [3] conditions on nearby markers, it does not remove the effects of LD completely and necessitates LD-pruning. Unfortunately, there is no consensus on how best to prune markers while still retaining enough information to infer distant relatives. Although we carried out a simple simulation study to get a rough sense of appropriate level of pruning, it is by no means a complete solution. More work is needed on the effects of LD on relatedness inferences and how to remedy the problem, whether it be by more extensive simulations studies, or by modeling LD in the likelihood computation. Furthermore, care must be taken to use appropriate allele frequencies in likelihood computation to account for other potentially confounding factors such as population substructure [6, 83] and admixture [57, 71]. As better methods for estimating pairwise likelihoods become available, our method for estimating pedigrees should also improve.

There are limitations to our method that require further work. Our method assumes that all individuals are outbred, which may not be true of many systems including some human populations [41, 19]. It currently does not support pedigrees with cycles caused by inbreeding or complex cyclic relationships such as double first cousins. When inbreeding is present, CLAPPER infers pedigrees that are close to the underlying truth under the assumption that there is no inbreeding (see Effect of Inbreeding). Pedigree non-identifiability also poses a challenge to pedigree estimation. Donnelley [15] remarked that two pairs of cousin-type pedigrees that have equal numbers of meioses are not identifiable (e.g. half cousins vs. great half avuncular) no matter how much genetic data are available. Furthermore, Kirkpatrick [37] gave examples of non-identifiable 3-person pedigrees where no likelihood-based methods, including the full likelihood, can find the correct pedigree for certain. Another limitation of our method is that it does not provide an uncertainty measure on the estimated pedigree. This could be solved in two ways: by block-bootstrapping the data and repeating the inference, which would be slow; or using a Bayesian approach by assigning a prior to pedigrees and attempting to sample from the posterior distribution. Furthermore, while computationally efficient compared to full likelihood methods, our method is still based on calculation of pairwise relationships and does, therefore, not scale up to GWAS data sets with hundreds of thousands of individuals. However, it may be possible to use a divide-and-conquer approach in which individuals are first divided into clusters using methods such as [45], then estimating the pedigree of each cluster separately, and finally estimating more distant relationships among clusters.

Overall, our method provides a computationally efficient way to estimate pedigrees of seemingly unrelated individuals. It improves our ability to validate and discover pedigrees in realistic genetic datasets where we expect a high level of missing data. The ability to estimate pedigrees more accurately opens up possibilities to develop and improve numerous pedigree-based or pedigree-aware studies, from correcting cryptic relatedness in GWAS to estimating demographic parameters of the very recent past. However, as noted in Introduction, the naive use of estimated pedigrees in downstream analyses may not be justified when there is significant statistical uncertainty in the estimation of the pedigree. Such analyses would need to take the statistical uncertainty in pedigree estimation into account, a topic of potential future research.

Our software is available for download at https://github.com/amyko/clapper.

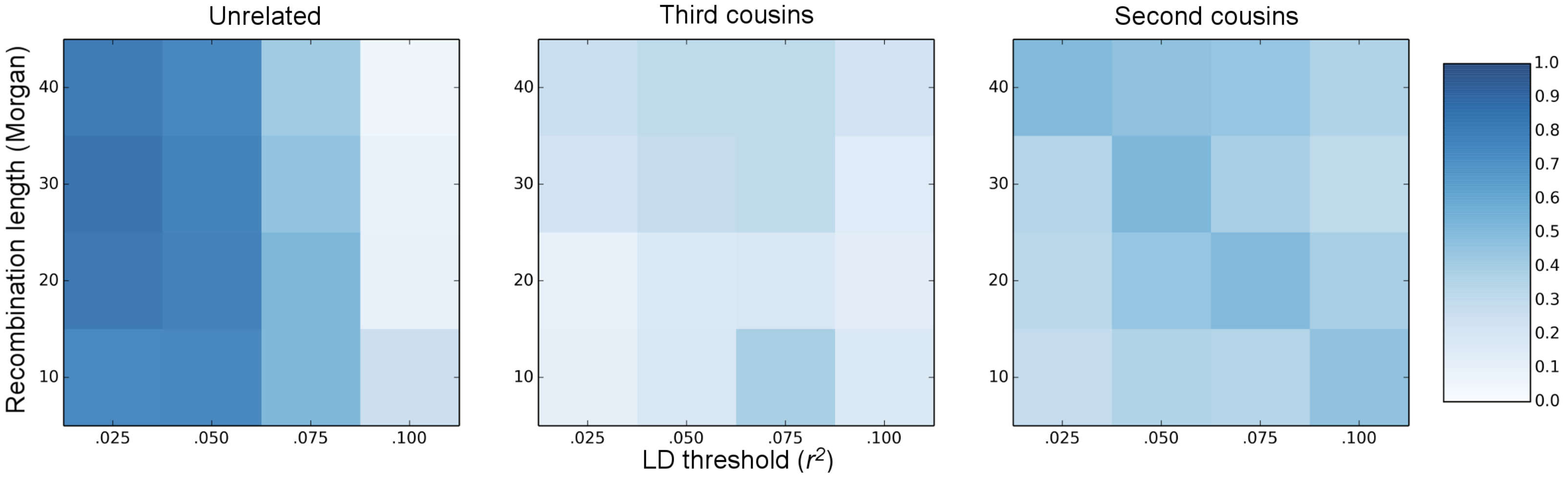

Figure 2.6: Effects of LD-pruning on pairwise prediction accuracy. The three panels show different true pairwise relationships: unrelated, third cousins, and second cousins. Each square in a panel corresponds to the relationship prediction accuracy for a particular genome length and LD-prune threshold. The color indicates the accuracy rate between 0 and 1.

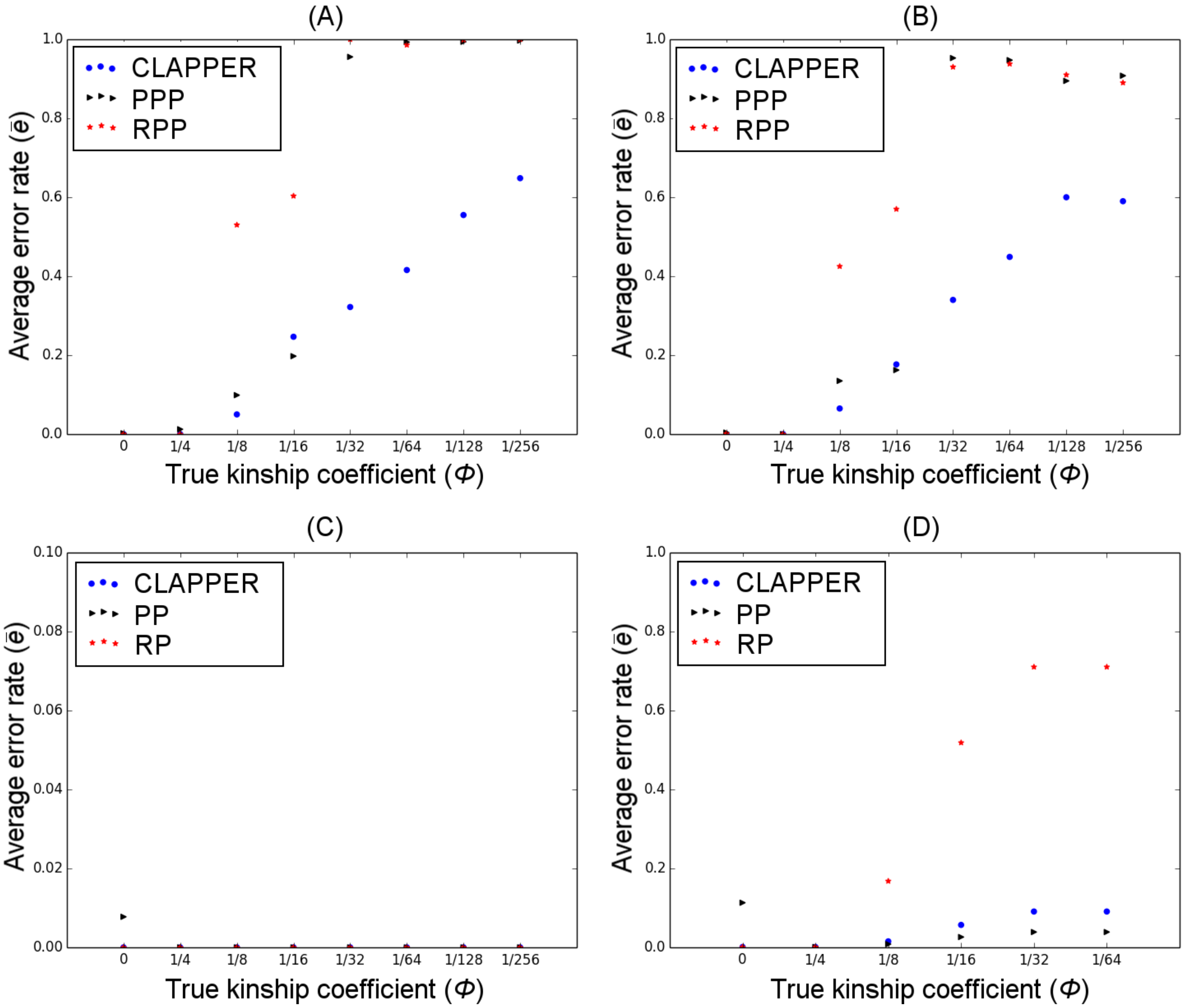

Figure 2.7: Comparison of prediction error rates. Each panel compares the average error rate between CLAPPER and competing methods for a particular simulation scenario: (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D. The x-axis shows different relationship categories measured by the kinship coefficient; the y-axis is the average error rate $\bar{e}$ (See Measuring the Error Rate). Analysis excludes all experiments that did not finish successfully or did not produce any outbred pedigrees.

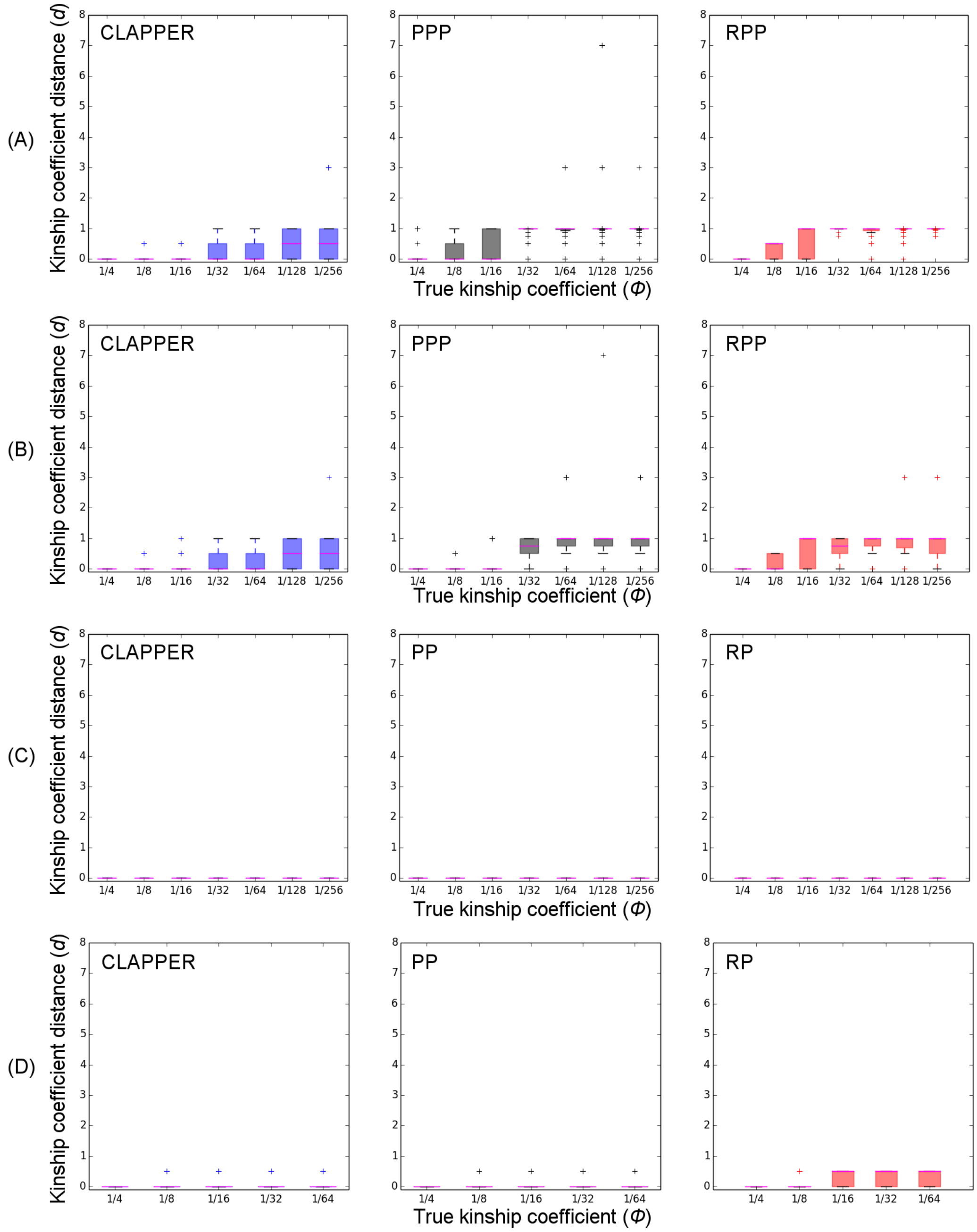

Figure 2.8: Absolute between the expected kinship coefficient under true and inferred relationships, normalized by the true kinship coefficient. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D. The x-axis is the relationship category measured by the kinship coefficient; the y-axis is the distance d between the true relationship and the relationship estimated by our method (See Measuring the Error Rate in Materials and Methods section). The magenta line indicates the median value for each box plot. Analysis excludes all experiments that did not finish successfully or did not produce any outbred pedigrees.

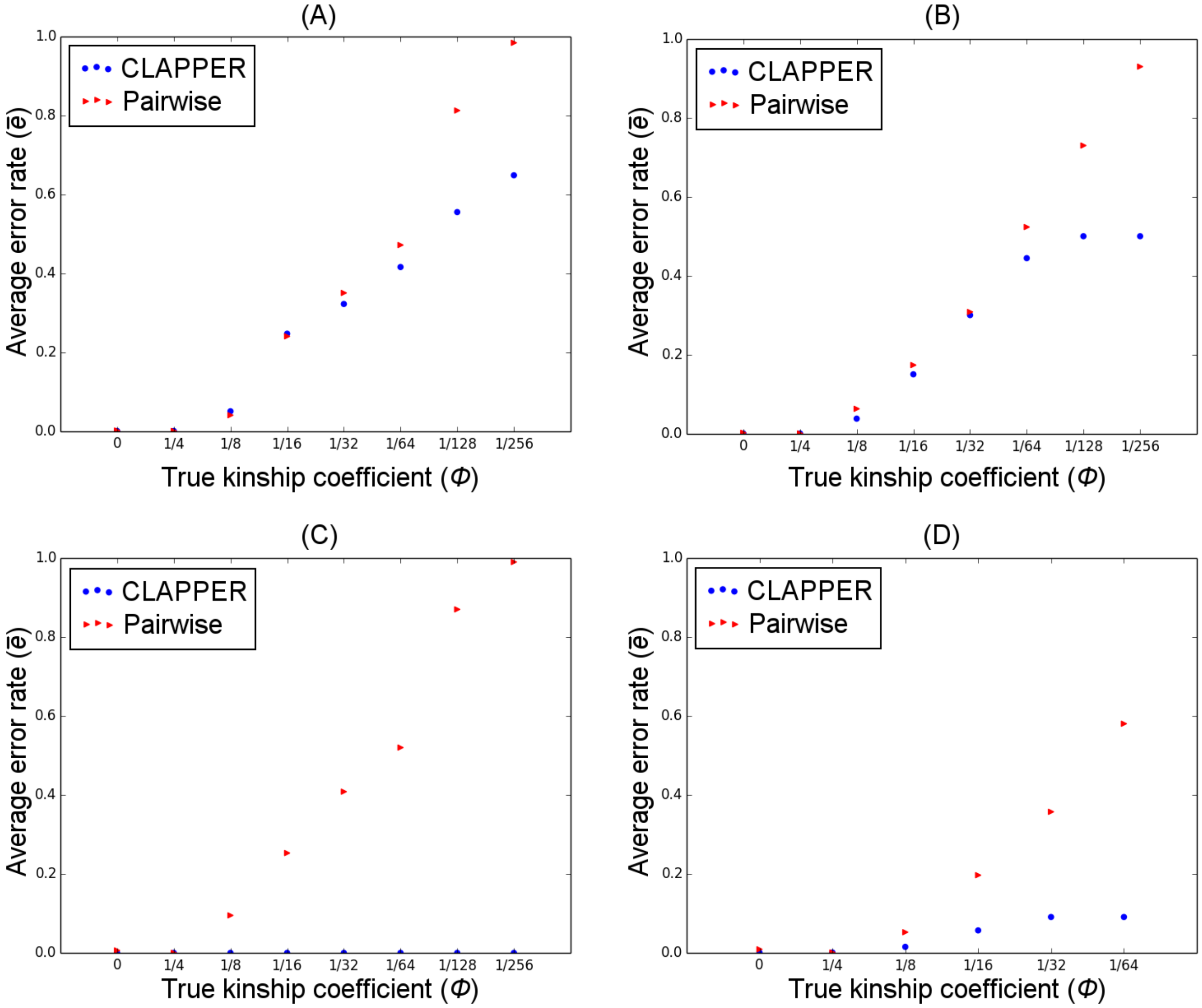

Figure 2.9: Comparison of prediction error rates between CLAPPER and pairwise inference. Each panel compares the average error rate between the pairwise method and CLAPPER for a particular simulation scenario: (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

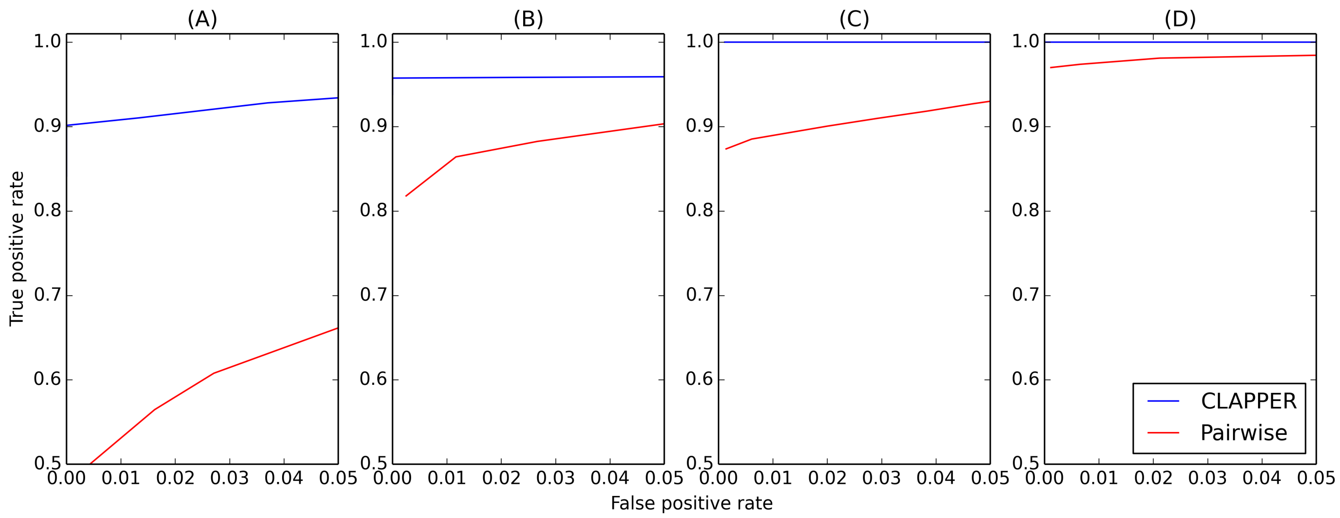

Figure 2.10: ROC curve for detecting relatives in a sample: pairwise vs. CLAPPER. (A) simulation A; (B) simulation B; (C) simulation C; (D) simulation D.

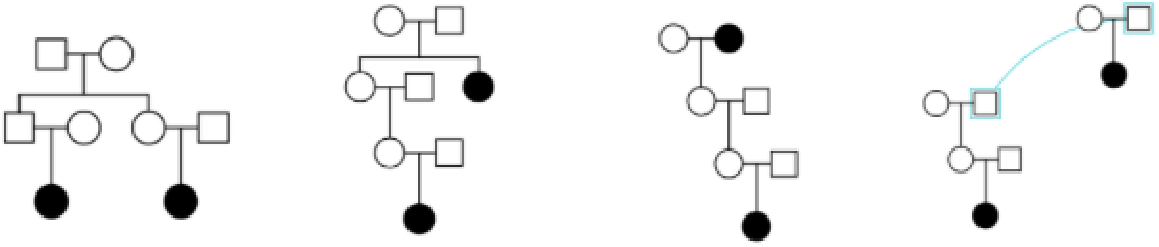

Figure 2.11: Different types of pedigrees inferred by CLAPPER for the pedigree shown by Figure 2.4A.

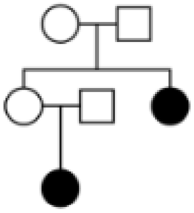

Figure 2.12: Different types of pedigrees inferred by CLAPPER for the pedigree shown by Figure 2.4B.

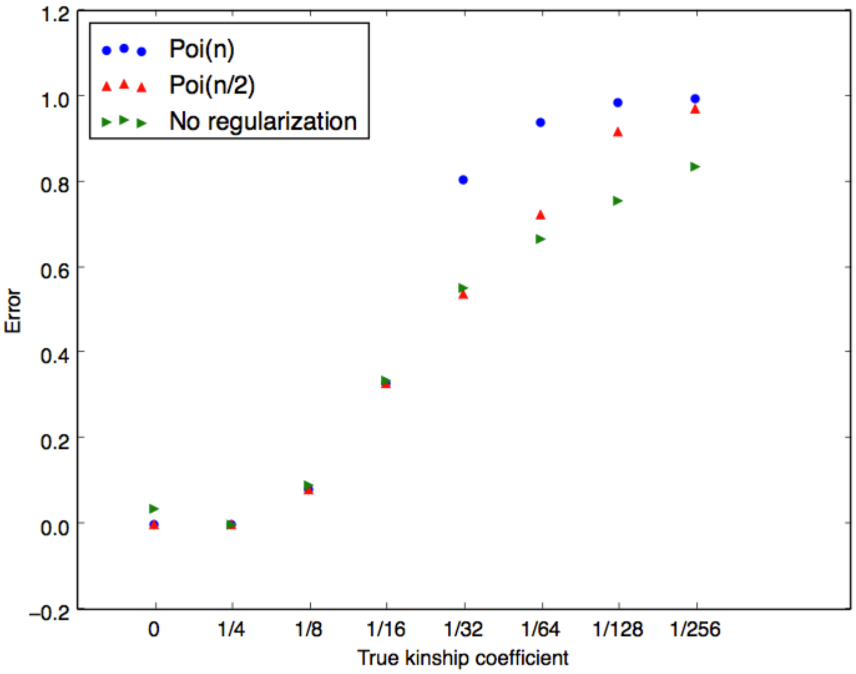

Figure 2.13: Effects of Regularization Term. Accuracy of simulated annealing method on simulated data at 2000 markers under different levels of regularization.

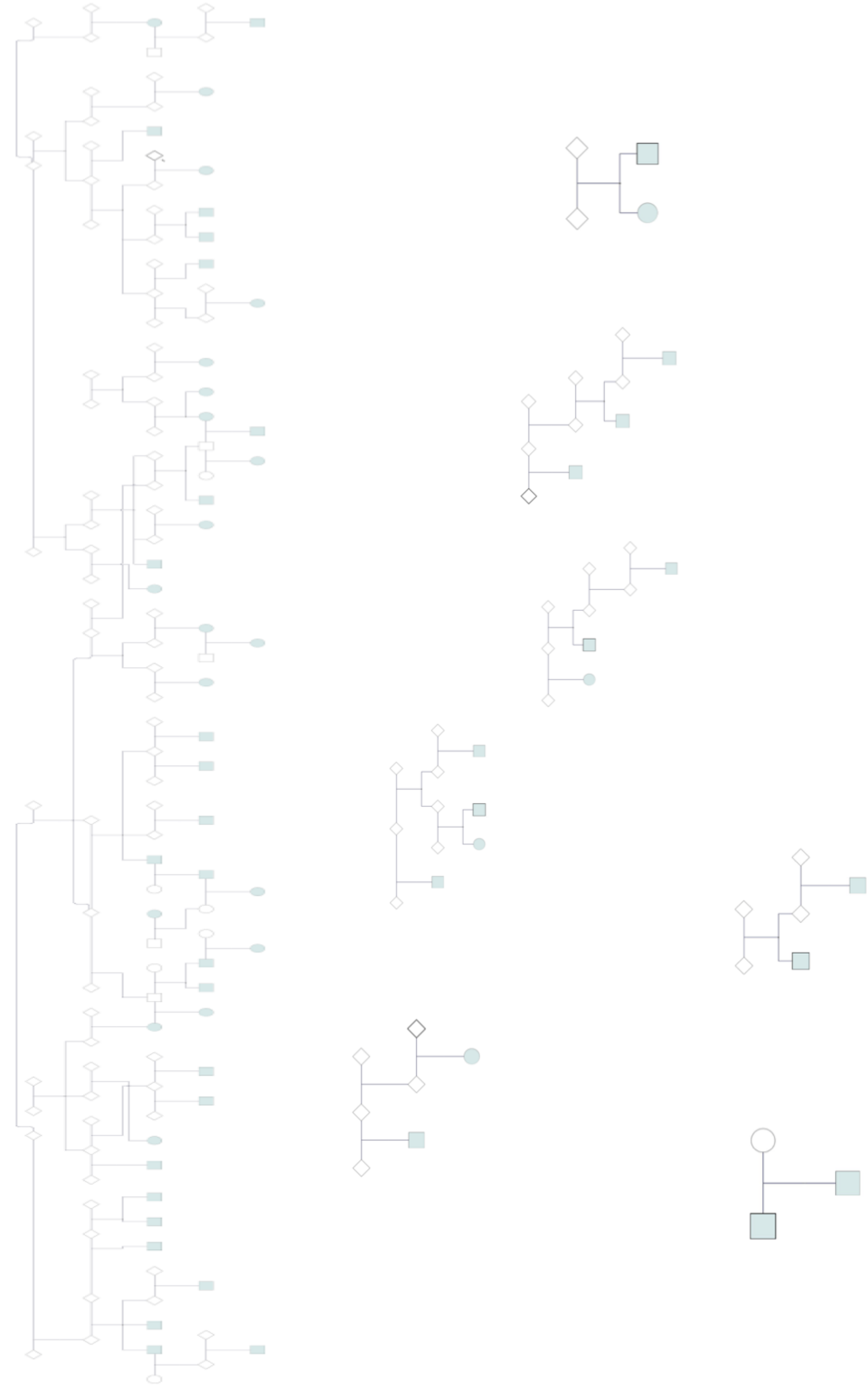

Figure 2.14: Estimated pedigree of 100 Tasiilaq Individuals in the Greenlandic Inuit Dataset. Shaded nodes indicate sampled individuals; unshaded for unsampled; squares for male; circles female; diamonds for unknown sex.

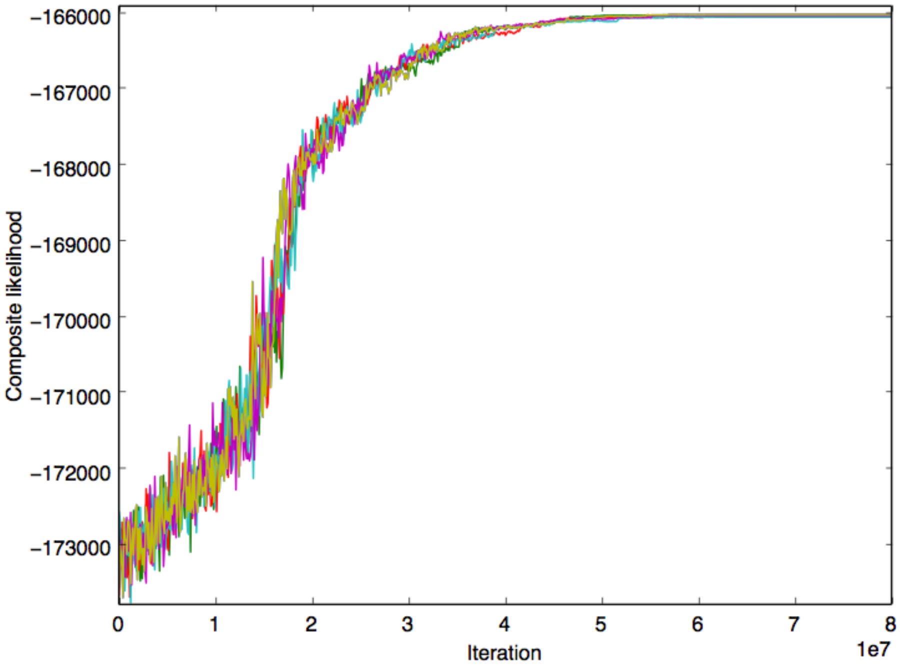

Figure 2.15: Likelihood Convergence for the Greenlandic Inuit Pedigree Estimation.

Chapter 3

Joint Estimation of Pedigrees and Effective Population Size Using Markov Chain Monte Carlo

3.1 Introduction

As discussed in Chapter 2, the genealogical history embedded in pedigrees can be used to estimate demographic parameters for the recent past, such as short-term effective population size (N_e) [81]. Most population genetic models are based on Kingman's coalescent [33, 34, 35], which is a poor approximation of the genealogical process for time frames shorter than $log_2 N$, where N is the population size [76, 75]. Pedigrees, which provide a finer resolution on the genealogical history of the samples than the coalescent, may therefore be more appropriate to use for estimating demographic parameters of the very recent past.

In this study, we present a Bayesian method that jointly estimates pedigrees and N_e. We use the composite likelihood developed in [39] to make the likelihood computation efficient for a large number of markers and individuals. We also use Markov Chain Monte Carlo (MCMC) [24] to sample pedigrees from high probability regions, circumventing the need to enumerate all possible pedigrees. Our method is different in several important ways from previous methods such as [82, 63, 39] that also use composite likelihoods and sampling algorithms to explore the pedigree space. The previous methods take a maximum likelihood approach and produce a list of pedigrees with highest likelihoods, and does not provide a principled way to compute the uncertainty of the estimated pedigrees. In contrast, our method casts the problem in a Bayesian framework and estimates the posterior probability distribution of the parameters, which in turn quantifies the uncertainty in the parameter estimation.

Furthermore, by assigning a prior to the pedigrees, which is a function of population parameters that govern mating behavior of the population, we can estimate these parameters jointly with the pedigree. In particular, we focus on estimating short-term N_e, a key parameter in areas such as conservation genetics as it quantifies the level of genetic drift and inbreeding in the current population. Various approaches have been developed for estimating short-term N_e, including methods based on relatedness, heterozygosity excess, linkage disequilibrium, or changes in allele frequency over time [79]. Our pedigree-based approach for estimating N_e is most closely related to the estimation method based on the frequency of siblings in a sample by [81], which was shown to be more accurate and robust than other approaches.

In our method, we assume that all individuals belong to a single generation and infer pedigrees going up to two generations back in time (i.e. up to first cousins). Furthermore, we assume that the population is outbred with non-overlapping generations and the pedigrees do not contain cycles other than those caused by full sibling relationships. We validate our method on simulated data and show that it can estimate relationships and N_e accurately. Furthermore, we apply our method on a real dataset containing a sample of house sparrows to reconstruct its previously unreported pedigree.

3.2 Materials and Methods

Bayesian Inference of Pedigrees and Mating Parameters

Our method aims to estimate the joint posterior distribution of pedigrees and mating parameters. Let n be the sample size, H the pedigree of the sample, θ the set of mating parameters for the population, and $X = (X_1, ..., X_n)$ the set of genotype vectors for the n individuals. Then the joint posterior probability of H and θ can be written as

$$Pr(H, \theta | X) \propto Pr(X|H) Pr(H|\theta) Pr(\theta), \tag{3.1}$$

where $Pr(X|H)$ is the likelihood of the pedigree, $Pr(H|\theta)$ is the prior for the pedigree under a mating model parameterized by θ, and $Pr(\theta)$ is the hyperprior on the mating parameters. We describe below how to compute each of these component terms in more detail.

Composite Likelihood

As discussed in Chapter 2, computing the likelihood of a pedigree, $Pr(X|H)$, is computationally prohibitive for even a moderately large set of markers or individuals. We therefore approximate the likelihood with the composite likelihood introduced in [39] to make the computation more efficient. The composite likelihood is based on the marginal pairwise likelihoods. See Chapter 2 for a more detailed description of the composite likelihood.

We pre-compute and store in memory the pairwise likelihoods $Pr(X_i, X_j | R_{i,j})$ for each pair (i, j) for a specified set of pairwise relationships, where $R_{i,j}$ is the relationship between individuals i and j induced by pedigree H. For pedigrees going up to two generations back in time, this set includes full siblings, half siblings, full first cousins, half first cousins, and unrelated. The pairwise likelihoods can be computed efficiently using the method described in [84] for unlinked markers or by [3] for linked markers. The pairwise likelihoods can then be accessed from memory to compute the composite likelihood efficiently.

Prior

For the prior on the pedigrees, $Pr(H|\theta)$, we used a modified version of the mating model introduced in [17]. The model is defined by three parameters: α, β, and N, which we describe in more detail below.

The probability of a pedigree under the Gasbarra mating model is most naturally described by the procedure by which each child stochastically chooses its mother and father. We assume a homogeneous population of constant size N with non-overlapping generations and equal proportions of males and females (i.e. $N/2$ males and $N/2$ females). Let n be the number of children in the current generation. One by one, each child chooses a parental pair (f, m) where $f \in \{1, 2, ..., N/2\}$ and $m \in \{1, 2, ..., N/2\}$.

Let $C_f(k)$ be the number of children that mother f has after the first k children have chosen their parents. Then the probability that the $(k+1)^{th}$ child chooses mother f is given by

$$\frac{\alpha + C_f(k)}{\alpha(N/2) + k}, \tag{3.2}$$

where α is a parameter that controls the offspring distribution among mothers in the population. A small value of α corresponds to the mating model where a few mothers have many offspring, whereas a large value of α corresponds to the model where children are distributed more evenly among all mothers.

After selecting mother f, the child chooses a father next. Let $C_{fm}(k)$ be the number of children that parental pair (f, m) has after the first k children have chosen their parents. Then the probability of the $(k+1)^{th}$ child choosing father m is given by

$$\frac{\beta + C_{fm}(k)}{\beta(N/2) + C_f(k)}, \tag{3.3}$$

where β is a parameter that governs the degree of polygamy of fathers. If β is small, then the child is more likely to choose father m if the father already shares offspring with the child's mother, f (i.e. parental pairs tend to stay monogamous). On the other hand, $\beta = \infty$ corresponds to the case where the child chooses a father at random (i.e. random mating model).

After all n children in the current generation have chosen their parents, we continue recursively backwards in time by treating the chosen mothers and fathers in the current stage as the offspring for the next stage. Using this sequential sampling scheme, we can compute $Pr(H|\theta)$, where $\theta = (\alpha, \beta, N)$.

Furthermore, we can relate the mating parameters α, β, and N to the effective population size, N_e, using the formula derived in [17].

For the hyperprior, $P(\theta)$, we assume a uniform distribution for each of the parameters in θ. For instance, we assume $\alpha \sim U(\alpha_{min}, \alpha_{max})$ for some fixed α_{min} and α_{max}. We treat β and N in a similar way.

Finally, we combine the composite likelihood, prior, and hyperprior to approximate the joint posterior distribution of H and θ with

$$CL(X|H)Pr(H|\theta)Pr(\theta) \tag{3.4}$$

Markov Chain Monte Carlo

To explore the vast parameter space in a computationally feasible way, we use Markov-Chain Monte Carlo (MCMC) to sample from the posterior distribution of H and θ, approximated by Equation 3.4.

We represent the pedigree for a sample of individuals as an directed graph, where a node corresponds to an individual with a particular sex (i.e. male or female) and an edge represents a parent-offspring relationship. Individual i in the graph is not necessarily represented in the sample; but if it is sampled, the node is associated with a genotype vector X_i. A more detailed description of how a pedigree is represented as a graph and what constitutes as a valid pedigree is provided in [39].

The MCMC explores pedigrees and mating parameters simultaneously. To explore the pedigree space, we make local modifications to the edges and the nodes in the graph using 10 reversible updates. The 10 updates can broadly be categorized into two groups. The first category of updates involves inserting or deleting edges to join or split pedigrees. The second category is modifying the pairwise relationship between two randomly chosen individuals, such as changing half-siblings to full-cousins, and vice versa. To explore the mating parameters, we use three different updates–one for each mating parameter–where we propose a new state by sampling the new parameter value from a normal distribution centered at the current value. A more detailed treatment of the updates is given in File S1.

Here, we outline the MCMC algorithm. Let $Q = (H, \theta)$ denote the set of parameters we want to estimate (i.e. pedigree and mating parameters).

1. Initialize pedigree H to be the one in which every individual is unrelated to each other. Initialize α by sampling from $U(\alpha_{min}, \alpha_{max})$, for some fixed α_{min} and α_{max}. Initialize β and N in a similar way. Compute and store Equation 3.4 for the current configuration.

2. Choose one of the 10 updates at random and generate a new configuration.

3. If the new configuration is invalid, reject and go back to step 1. If it is valid, accept the new configuration with probability

$$min\left(1, \frac{CL(H_{new})Pr(H|\theta_{new})Pr(Q_{old}|Q_{new})}{CL(H_{old})Pr(H|\theta_{old})Pr(Q_{new}|Q_{old})}\right),$$

4. Repeat steps 1-3 T times.

The total number of samples, T, was chosen to achieve a balance between convergence of the Markov chain and computational time. Since we only want to keep samples after the Markov chain has converged to the stationary distribution, we discarded the first B samples as burn-in. To check the convergence of Markov chains, we ran multiple independent MCMC chains and checked that all chains fluctuated in a similar, stable range of log likelihood values. We note that this is only a proxy for checking convergence and there are other, albeit more involved, ways to check convergence, such as checking the potential scale reduction factor for some specified quantity [20]. Furthermore, we keep only every tth sample to avoid storing correlated samples.

For both simulated and empirical datasets, which will be described next, we ran the MCMC for $T = 6 \times 10^6$ iterations with the burn-in period of $B = 4 \times 10^6$ iterations. The hyperprior for the mating parameters was set as follows: $\alpha \sim U(.1, 100)$, $\beta \sim U(1 \times 10^{-5}, .1)$, and $N \sim U(5, 5000)$. We also thinned the MCMC samples by keeping only every 50th sample.

Simulated Data

We tested the performance of our method on simulated data. We simulated pedigrees up to two generations back in time using the mating model described in Prior with $\alpha = 15$, $\beta = 15$, and $N = 1000$, which translates to $N_e = 650$ using the formula given in [17].

We then simulated 10,000 independent single nucleotide polymorphic sites (SNPs) for each of the N founders in the pedigree, where the population allele frequency for each marker was sampled from the site frequency spectrum under neutrality. We assumed that the markers were spread evenly among 20 independent chromosomes of length 100Mb, and assumed sequencing error rate of .01. To test the effect of marker type on our parameter inference, we also simulated 20 microsatellites with 10 alleles of equal frequency per marker. Furthermore, we assumed that each marker was on an independent chromosome, and had sequencing error rate of .01 and allele dropout rate of .05.

We then simulated the genotypes for the children in the pedigree by recombining parental haplotypes at rate 1.3e-8 per base pair per generation. We generated 50 independent datasets for both SNP and microsatellite simulations. For convenience, we refer to the simulations with SNPs as Simulation A and those with microsatellites as Simulation B in later sections.

Empirical Data

We applied our method to reconstruct the previously unreported pedigree of house sparrows collected from an archipelago off the Helgeland coast of northern Norway [44]. The individuals were genotyped using a custom Affymetrix 200K SNP array, with markers distributed across 29 of the chromosomes in the genome. Also provided were the location and year in which each each individual was collected.

We used individuals from a single island (island 27) to avoid any potential substructure in the sample. Furthermore, we restricted our analysis to the individuals born in 2009 to ensure that all samples belong in a single generation. We pruned the markers for linkage disequilibrium (LD) using PLINK [9] at $r^2 = .05$ to get a set of independent or loosely linked markers. The filtering steps resulted in 79 individuals and 4519 SNPs.

Evaluation of Method

We compared the performance of our method to that of COLONY [31], one of the most widely used pedigree reconstruction methods. We chose COLONY for several reasons. First, it supports full likelihood computation, which provides a gold standard to which we can compare our composite likelihood method. Second, it supports both SNPs and microsatellites data, allowing us to compare the performance of different marker types. Third, COLONY can estimate short-term N_e based on the estimated frequency of siblings in the sample, which was shown to be more accurate than other methods of estimating short-term N_e [81].

Because the sample size in our simulations was much smaller than the population size, many pedigrees for the sample had similar likelihoods, making it difficult for both our method and COLONY to find the correct pedigree in its entirety. So we used pairwise prediction accuracy as a proxy for the accuracy of pedigree inference. In our method, we assigned pairwise relationship R to pair (i, j) if it had the highest posterior probability among all competing relationships. We approximated the posterior probability of R by counting the proportion of times pair (i, j) had relationship R in the MCMC samples. Similarly, we assigned relationship R to pair (i, j) in COLONY if it had the highest probability among all candidate relationships. Because the number of possible pedigrees is large, COLONY archives only the top w pedigrees with highest likelihoods. Suppose S is the set of indices for the pedigrees where (i, j) has relationship R. Then the probability of R is estimated by

$$\frac{\sum_{k \in S} L_k}{\sum_{i=1}^{w} L_i},$$

where L_i is the likelihood of the ith pedigree.

Furthermore, since COLONY restricts its inference to pedigrees going back only one generation back in time (i.e. siblings), we also limited our inference to the same scope when comparing the performance of our method to COLONY. The parameters used to run COLONY are detailed in File S2.

3.3 Results

Simulated Datasets

To illustrate some of the issues involved in estimating multi-generation pedigrees, we first turn our attention to an example from Simulation A. Figure 3.1 shows the two most likely local pedigrees involving three sampled individuals (shaded) and their estimated posterior probabilities. In the first pedigree, individual 3 forms a full first cousin relationship with the other two individuals (1 and 2), as opposed to a half first cousin relationship as in the second pedigree. Here, the true pedigree is shown by the first pedigree (Figure 3.1A), which had the highest posterior probability.

The uncertainty in the pedigree estimation, shown by the similar posterior probabilities of the two pedigrees (.55 and .45), was consistent with the fact that the pairwise likelihood values were similar under different relationships. More specifically, individuals 1 and 3 had a higher likelihood of being full cousins than half cousins by about one log likelihood unit. On the other hand, individuals 2 and 3 had a higher likelihood of being half cousins than full cousins by roughly the same amount. Based on pairwise likelihoods alone, individuals 1 and 3 would be classified as half cousins, and individuals 2 and 3 as full cousin. Piecing together such pairwise assignments, however, would not produce a valid pedigree. Such uncertainties in cousin inference were not uncommon: about 20 percent of true cousin pairs in Simulation A had nonzero posterior probabilities for both full and half cousins.

Table 3.1a shows the pairwise prediction accuracy by MCMC for 50 independent datasets in Simulation A, where the pairwise likelihoods were computed using the method by [3]. Full siblings, half siblings, and half cousins were classified correctly in almost all instances, whereas about seven percent of full cousin pairs were classified as half cousins. The rate of false detection of relatives was very low at about .01 percent, where the unrelated pairs were estimated as half cousins.

Figure 3.3A shows the posterior distribution of N_e estimated from the MCMC samples aggregated over 50 datasets in Simulation A. The mode of the posterior distribution was close to the true value, indicated by the red vertical line. Similarly, Figure 3.3B shows that the distribution of maximum a posteriori (MAP) N_e for the 50 datasets was concentrated around the true value. The three mating parameters that make up the components terms of N_e (i.e. α, β, and N) showed high correlations among them. Figure 3.2 shows that high values of N tended to co-occur with low values of α for this simulation, which suggests that these parameters should not be estimated independently of each other and marginal point estimates of any of these parameters are likely to be misleading.

Tables 3.1b and 3.1c compare the performance between our method and COLONY. Since COLONY estimates up to first degree relatives only, we also restricted the inference of our method to the same scope. Furthermore, we computed the likelihoods using the method discussed in [84] which assumes unlinked markers, an assumption that COLONY makes in its likelihood computation. Here, both our method and COLONY classified full siblings, half siblings, and unrelated without error. Both the methods also estimated all half cousin pairs to be unrelated. Furthermore, a similar proportion of full cousin pairs were misclassified as half siblings by both methods: 22 percent by COLONY and 24 percent by our method. As shown in Figure 3.4, N_e was underestimated by both methods, which is consistent with the higher proportion of half siblings in the estimated pedigrees, caused by the misclassification of some of full cousin pairs as half siblings.

Table 3.2a shows the pairwise prediction accuracy by MCMC for Simulation B (i.e. mi-

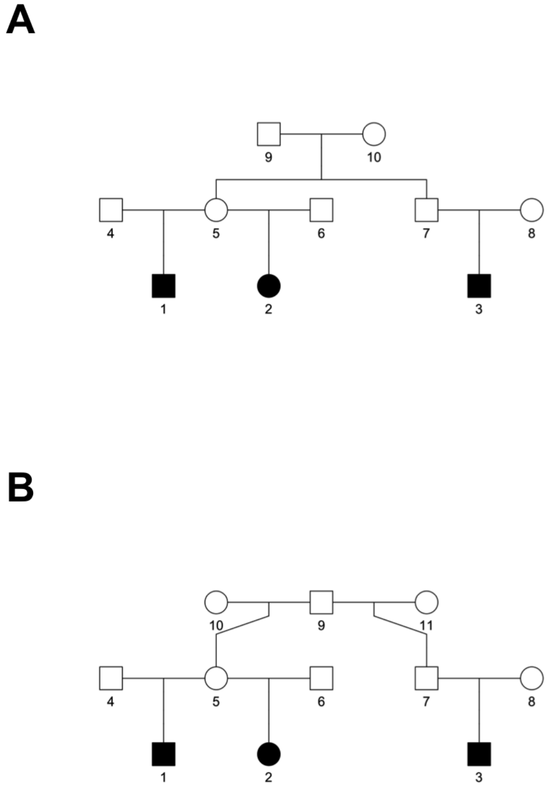

Figure 3.1: An example output pedigrees for three sampled individuals (shaded) from a dataset in Simulation A. Sex of the unsampled individuals (unshaded) are unknown but are drawn in for illustration only. (A) Pedigree with the highest estimated posterior probability ($p = .55$). (B) Pedigree with the second highest estimated posterior probability ($p = .45$). The true pedigree is shown in panel A.

(a) Two-generation Inference by MCMC

		Predicted				
		FS	HS	UR	FC	HC
True	FS	106	0	0	0	0
	HS	0	136	0	1	0
	UR	0	0	59996	0	4
	FC	1	0	0	445	32
	HC	0	0	0	4	526

(b) One-generation Inference by MCMC

		Predicted[a]		
		FS	HS	UR
True	FS	106	0	0
	HS	0	137	0
	UR	0	0	60000
	FC	0	117	360
	HC	0	0	530

[a] The likelihoods were computed without using the linkage information between markers to make the likelihood computation comparable to COLONY's.

(c) One-generation Inference by COLONY

		Predicted [a]		
		FS	HS	UR
True	FS	106	0	0
	HS	0	137	0
	UR	0	0	60000
	FC	0	106	371
	HC	0	0	530

[a] Inference was based on the full likelihood method under the assumption of independent markers.

Table 3.1: Pairwise Prediction Accuracy for Simulation A (SNPs)

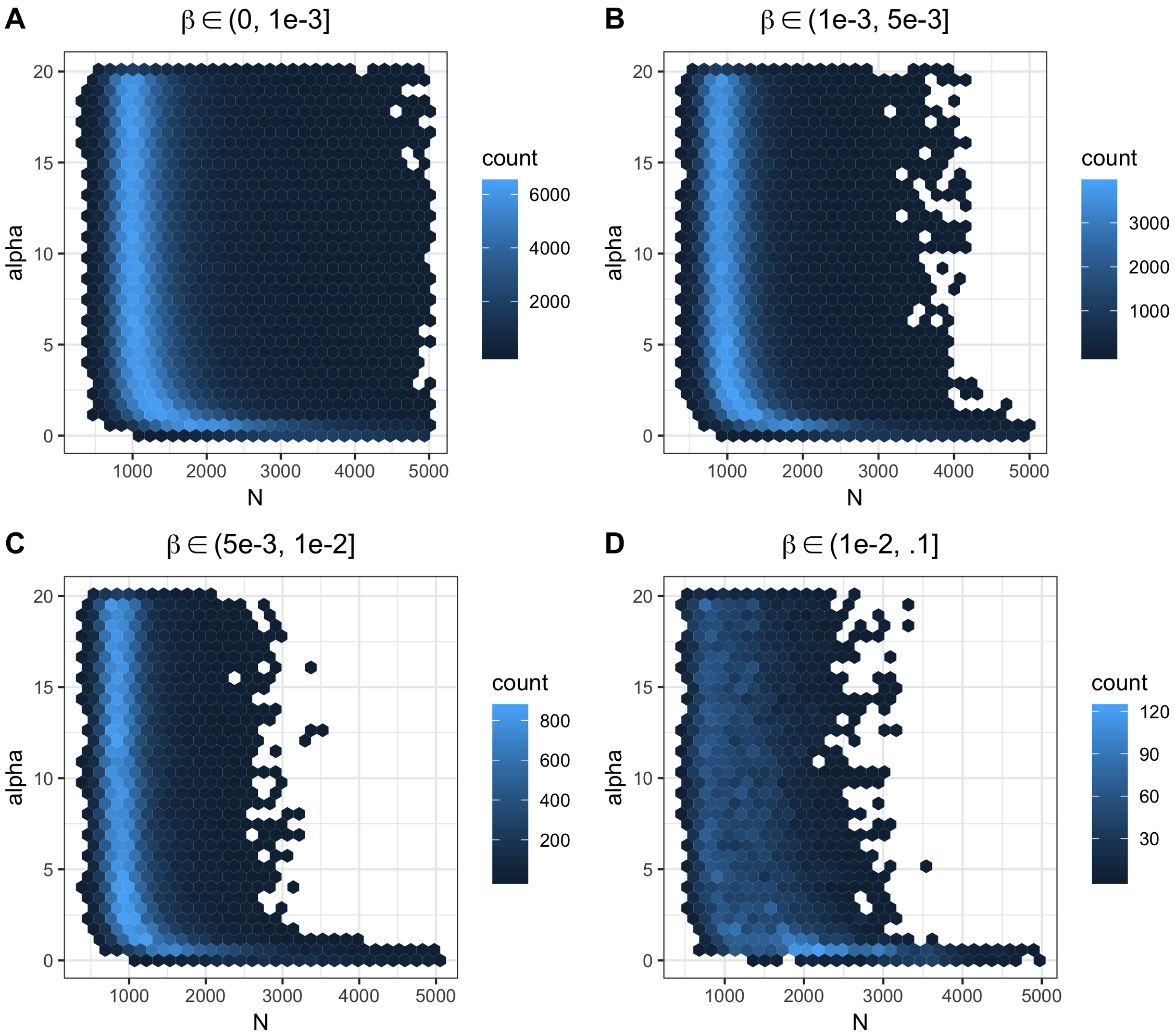

Figure 3.2: Each panel represents a heatmap of α and N for some fixed values of β, indicated on the top of each panel. The plots were generated from the MCMC samples aggregated over all 50 datasets in Simulation A.

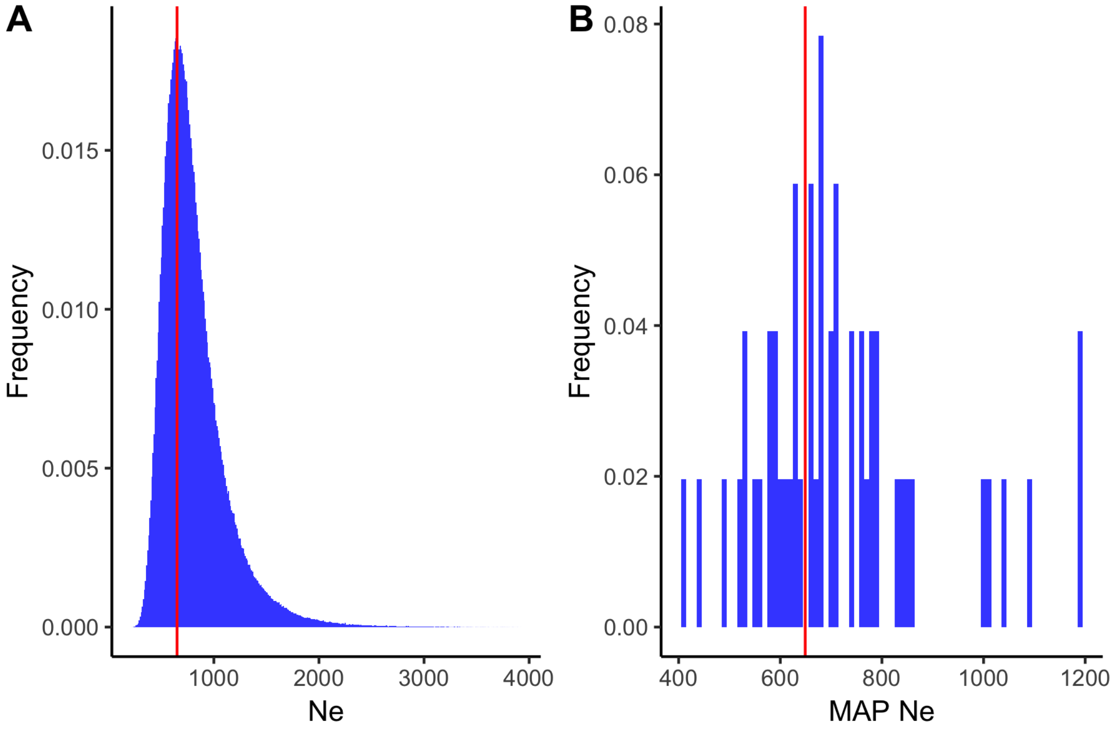

Figure 3.3: (A) Estimated posterior distribution of Ne from MCMC samples aggregated over 50 datasets in Simulation A. (B) Distribution of maximum a posterior (MAP) N_e for 50 datasets in Simulation A. The red vertical line in each panel corresponds to the true value of the parameter.

crosatellites), where the likelihoods were computed using the method by [78]. The accuracy rates were significantly lower than those in Simulation A (i.e. 10,000 SNPs). About 77 percent of full siblings and 27 percent of half siblings were classified correctly, and virtually all cousin pairs were estimated to be unrelated. This is likely due to the prior, which puts higher probabilities on sparsely connected pedigrees, overwhelming the likelihoods that do not show strong evidence for individuals being related. The distribution of MAP N_e also had a much higher variance compared to that of Simulation A (Figure 3.5A).

Tables 3.2b and 3.2c compare the performance of our method with that of COLONY for Simulation B. Again, we restricted the inference by our method to sibships to make a fair comparison with COLONY. Here, COLONY performed better than our method in correctly inferring full siblings and half siblings, but it also had a much higher false positive rate of 2.8 percent compared to .04 percent in our method. In fact, about 87 percent of the pairs estimated as half siblings by COLONY were actually unrelated. We note, however, that this problem may be addressed by adding an appropriate prior that is more conservative in half sibling assignments. Furthermore, due to the large number of unrelated pairs and cousins that were misclassified as half siblings, N_e was significantly underestimated by COLONY

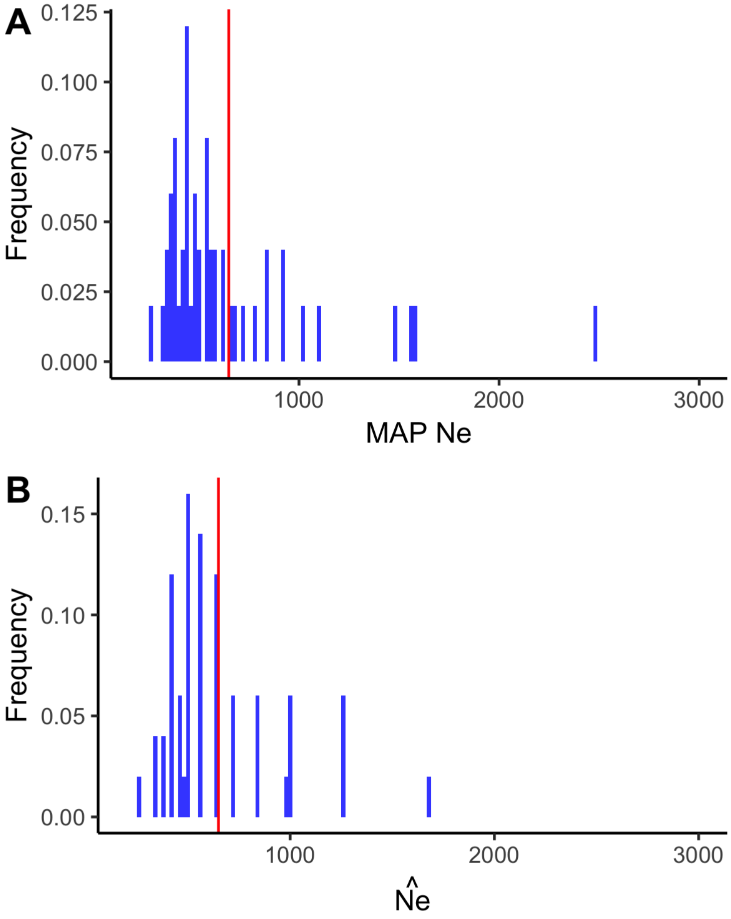

Figure 3.4: (A) Distribution of MAP N_e by MCMC, where the pedigree inference was restricted to one generation and the likelihood computation did not use the linkage information between markers. (B) Distribution of N_e by COLONY based on full likelihood computation and assuming nonrandom mating.

(Figure 3.5C).

For all the experiments, we checked the convergence of MCMC by studying the trace of the log likelihood values of multiple independent chains. As an illustration, we show an example of the log likelihood trace for the last one million iterations for a single experiment in Simulation A (Figure 3.6).

The running time for our method depends on many factors, such as the sample size, the underlying pedigree structure, and the maximum number of generation allowed in the pedigree inference. As an example, an MCMC run with 6 million iterations for a two-generation pedigree inference took about 36 seconds on a laptop with 2.3 GHz Intel Core i5 processor for a single dataset in Simulation A.

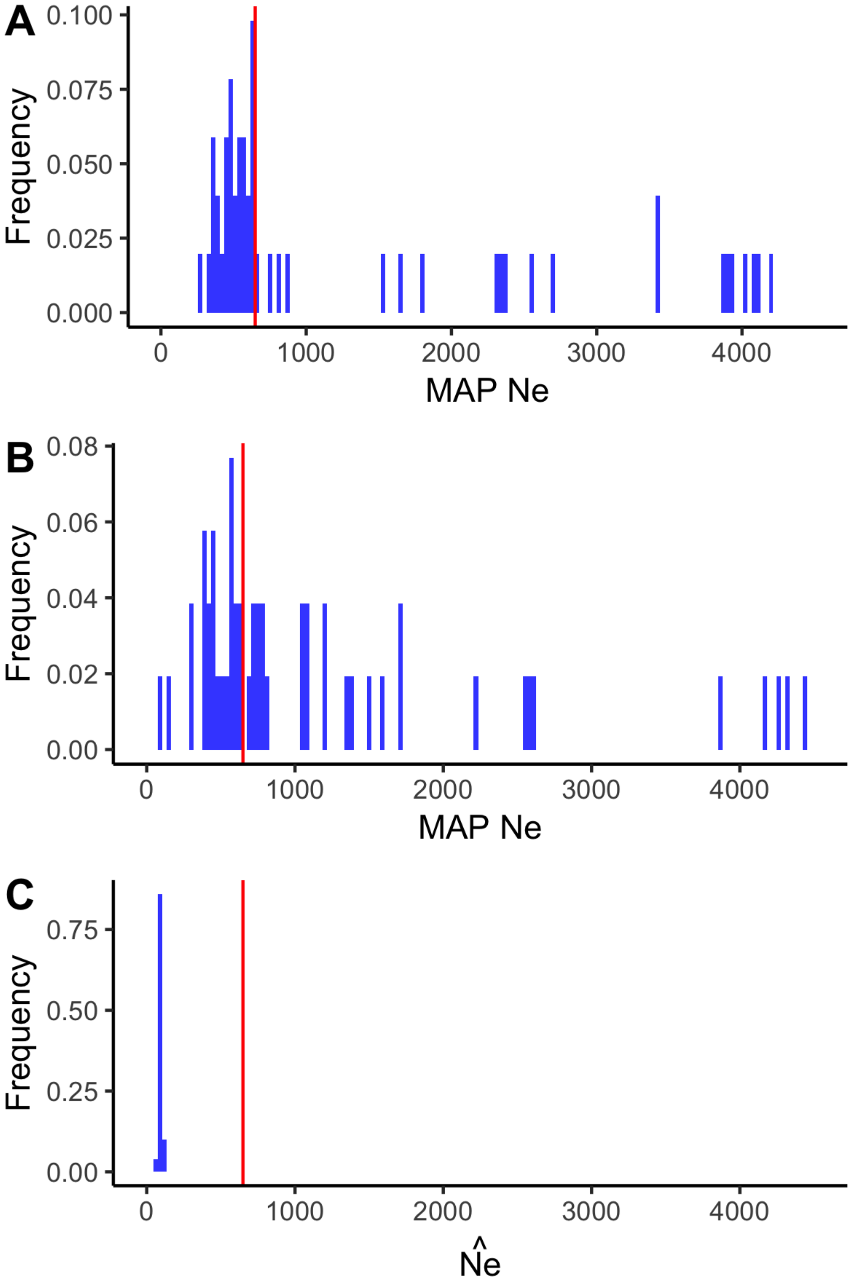

Figure 3.5: Distribution of the N_e estimate in Simulation B (i.e. microsatellites). (A) Distribution of MAP N_e estimated from MCMC samples under two-generation inference. (B) Distribution of MAP N_e estimated from MCMC samples under one-generation inference. (C) Distribution of N_e estimate by COLONY under nonrandom mating.

(a) Two-generation Inference by MCMC

		Predicted				
		FS	HS	UR	FC	HC
True	FS	96	22	7	0	0
	HS	2	31	81	0	0
	UR	0	23	60054	0	0
	FC	1	8	445	0	0
	HC	0	0	480	0	0

(b) One-generation Inference by MCMC

		Predicted		
		FS	HS	UR
True	FS	91	22	12
	HS	2	25	87
	UR	1	23	60053
	FC	0	3	451
	HC	0	0	480

(c) One-generation Inference by COLONY

		Predicted		
		FS	HS	UR
True	FS	102	22	1
	HS	2	92	22
	UR	3	1675	58399
	FC	1	105	348
	HC	0	39	441

Table 3.2: Pairwise Prediction Accuracy for Simulation B (Microsatellites)

Effect of Presence of Relatives Beyond First Cousins

For real datasets, it is often unreasonable to assume that the sample does not contain relatives more distant than first cousins. Here we show the effect of having second cousins in the sample on the inference of pedigrees and N_e. Table 3.3 shows the prediction accuracy for a simulation scenario where second cousins were present in the sample. The simulation parameters were identical to those of Simulation A, except for the number of generations under which the pedigrees were simulated. Instead of going back up to two generations back in time as in Simulation A–which generated relatives up to first cousins–here we simulated pedigrees up to three generations back in time, which generated second cousins as well.

As we can see in Table 3.3, the accuracy rates were similar to those of Simulation A for relationships up to first cousins. However, about 73 percent of full second cousins (2FC) were classified as half first cousins (HC), the most distant relationship type our method is designed to estimate. Similarly, about 22 percent of half second cousins (2HC) were classified as HC.

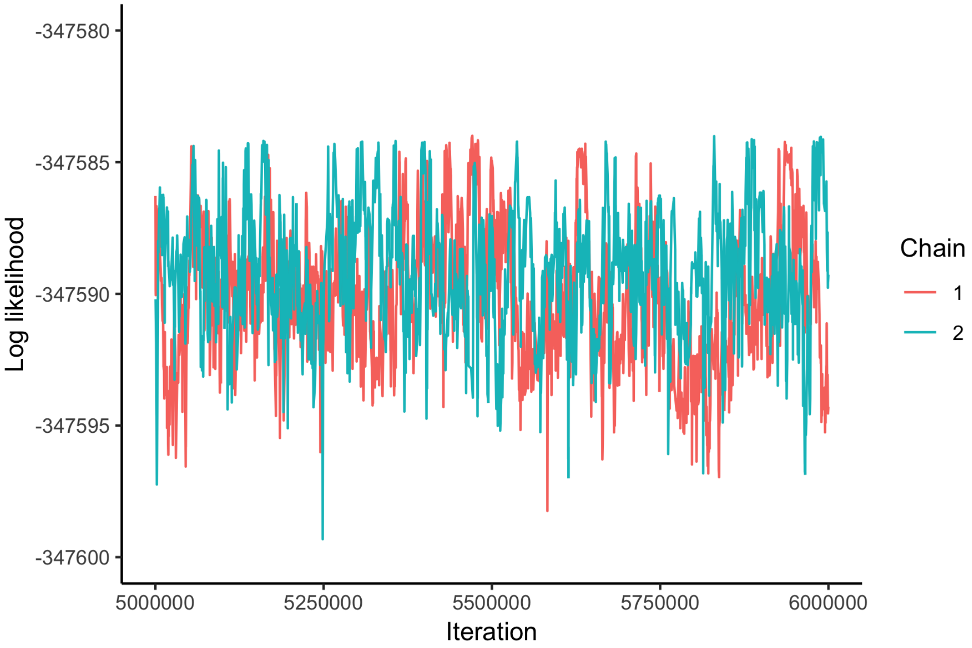

Figure 3.6: Likelihood trace of two independent MCMC chains for the last million iterations for a dataset in Simulation A. Both chains fluctuate in a similar, stable range of log likelihood values

As expected, N_e was biased downward due to the high frequency of HC in the estimated pedigrees, caused by the misclassification of second cousins as HC (Figure 3.7).

Table 3.3: Pairwise Prediction Accuracy for Datasets Containing Second Cousins (Inference by MCMC)

		Predicted				
		FS	HS	UR	FC	HC
True	FS	118	1	0	0	0
	HS	0	108	2	0	1
	UR	0	0	56189	0	3
	FC	5	5	0	386	95
	HC	0	0	9	4	499
	2FC	0	0	523	2	1388
	2HC	0	0	1482	0	430

To correct the downward bias in N_e estimation, we took advantage of the fact that our method can still infer siblings with high accuracy (Table 3.3). More specifically, we simulated pedigrees under various N_e to find a value that generated a number of siblings close to the one estimated by our method. Let $S_{IBD} = N_{FS} + .5N_{HS}$ be the summary statistic that measures

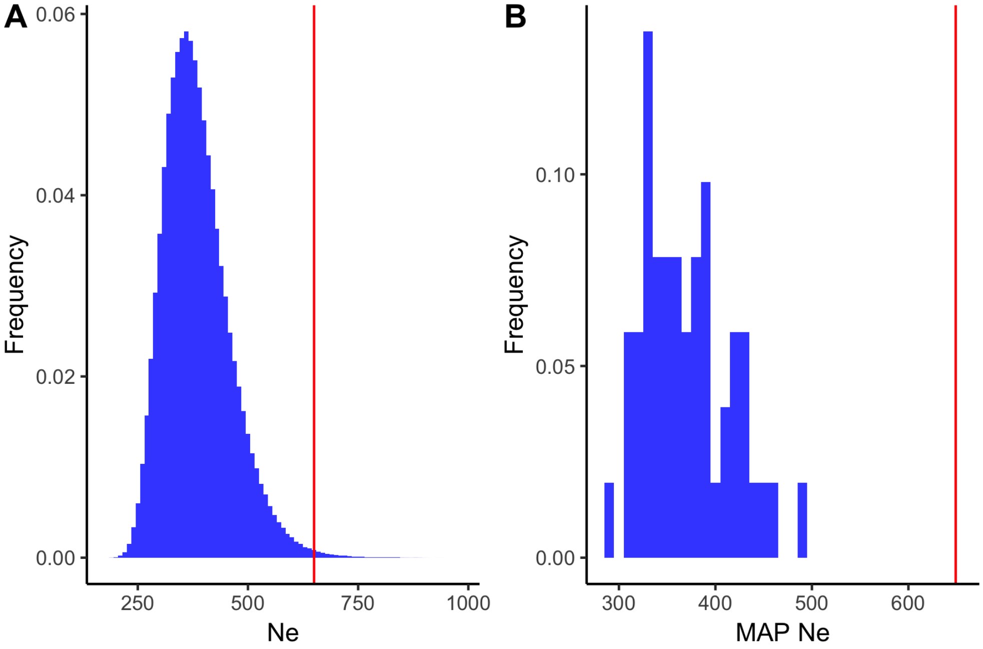

Figure 3.7: (A) Estimated posterior distribution of Ne from MCMC samples aggregated over 50 datasets. (B) Distribution of MAP N_e for the 50 datasets.

the level of identical-by-descent (IBD) contributed by siblings in the sample, where N_{FS} and N_{HS} are the number of full siblings and half siblings, respectively; and denote $\hat{S}_{IBD}$ to be the statistic obtained from the MCMC inference on the sample. Let α_{MAP} and β_{MAP} be the MAP estimates of α and β, respectively, computed using the marginal posterior distributions obtained from the MCMC samples. We then simulated pedigrees going back up to one generation in time under α_{MAP}, β_{MAP}, and various values of N–which translates to different values of N_e–and computed S_{IBD} from the simulated pedigrees. We then chose the value of N_e that produced S_{IBD} that most closely matched $\hat{S}_{IBD}$.

Figure 3.8 shows the distribution of the N_e estimates after correcting for bias as described above. Although the standard error was higher than that of uncorrected estimates, the median of the distribution (657) was much closer to the true value (650) than before.

Effect of Ignoring Linkage Information in Likelihood Computation

Table 3.4 shows the pairwise prediction accuracy for Simulation A, where the likelihoods were computed without taking into account the linkage information between markers [84]. For first-degree relatives, the accuracy rates were similar to those when linkage information was used in the likelihood computation. For second-degree relatives, however, the accuracy rates decreased significantly. For example, about a quarter of full cousin pairs were classified as half cousins and about 50 percent of half cousins were classified as unrelated. Furthermore,

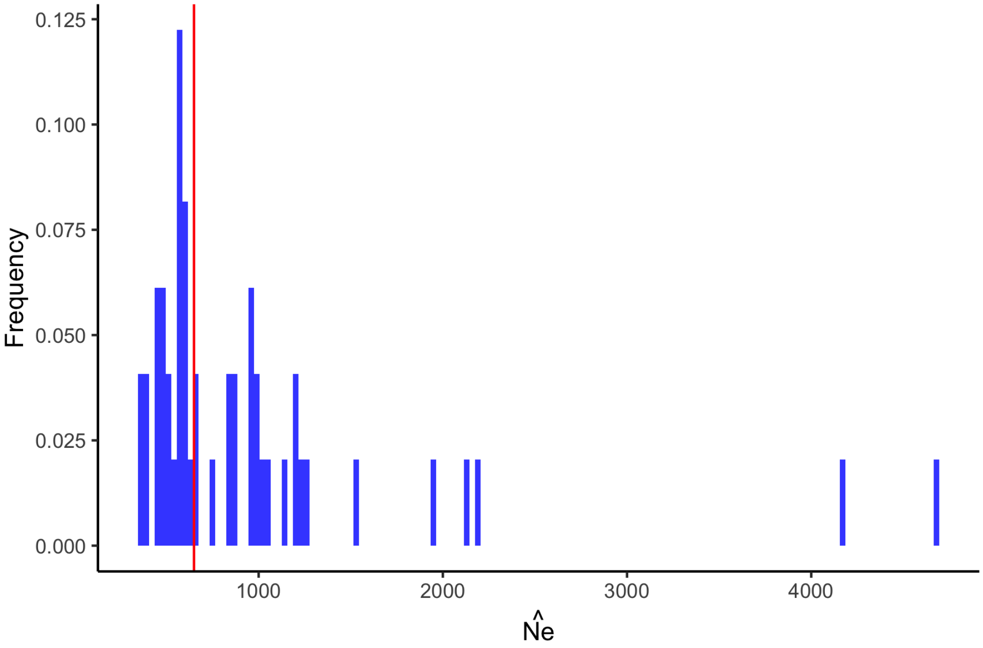

Figure 3.8: Distribution of $\hat{N}_e$ in 50 independent datasets after bias correction. The red vertical line indicates the true value of N_e.

N_e was overestimated (Figure 3.9), which is consistent with the fewer number of cousin pairs that were estimated in the pedigrees. The results show that likelihood computation methods that take into account the linkage between markers should be used, if possible, instead of those that assume independent markers.

Table 3.4: Pairwise Prediction Accuracy

		Predicted				
		FS	HS	UR	FC	HC
True	FS	106	0	0	0	0
	HS	0	136	0	1	0
	UR	0	1	600000	0	0
	FC	1	0	0	374	103
	HC	0	0	246	3	281

Sparrow Dataset

We analyzed a subset of the house sparrow dataset sequenced by [44]. After the filtering steps described in Empirical Data, the sample consisted of 75 individuals and 4,519 SNPs

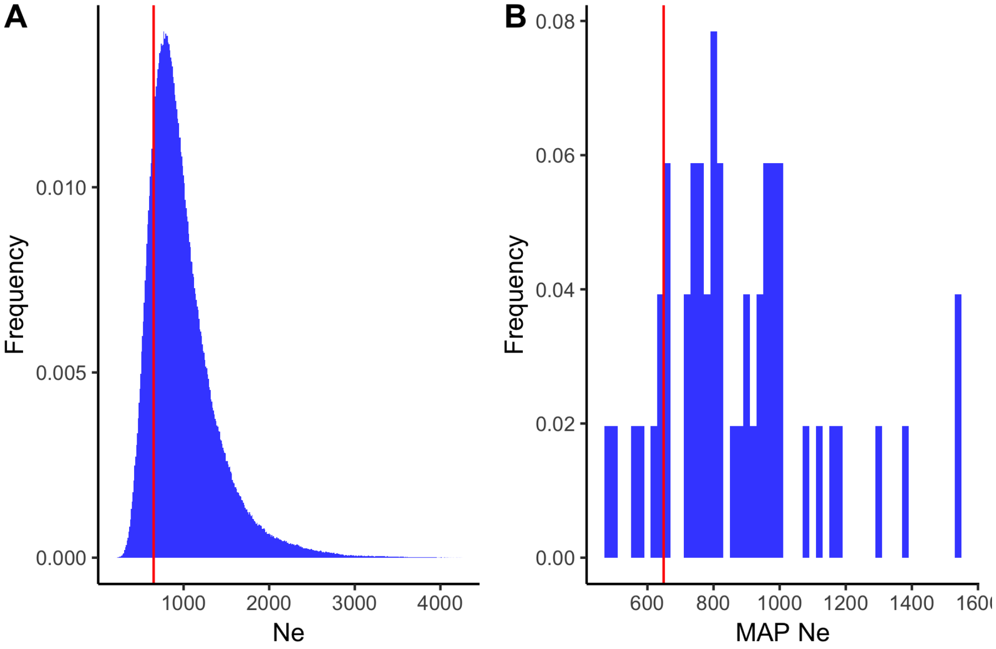

Figure 3.9: (A) Estimated posterior distribution of Ne from MCMC samples aggregated over 50 datasets in Simulation A, where likelihoods were computed without linkage information between markers. (B) Distribution of MAP N_e for the 50 datasets.

distributed across 29 autosomes. Here we show an example of the inferred pedigrees by our method and compare them to those estimated by COLONY.

Figure 3.10 shows the likely local pedigrees involving five individuals (shaded) in the sparrow dataset. The estimated posterior probabilities of the pedigrees shown in panel A and B were .77 and .23, respectively. The difference between the two pedigrees was the pairwise relationship between individuals 1339 and 1450, which was estimated to be full cousins in panel A and half cousins in panel B. Figure 3.10C shows the pedigree with the highest likelihood estimated by COLONY. This pedigree had posterior probability of zero in our method. We see that the half sibling relationship between individuals 1390 and 1450 were recovered by COLONY but all cousin relationships that our method detected were estimated to be unrelated. Based on the simulation studies in Simulated Datasets, however, we expect the full first cousin relationships inferred by our method to be either true first cousins or, with considerably smaller probability, more distant relatives (e.g. second cousins).

Table 3.5 compares the pairwise relationship classifications between our method and COLONY. Pairs that were classified as full siblings, half siblings, or unrelated by our method largely agreed with the classifications by COLONY. On the other hand, about 29 percent of pairs that were estimated to be full cousins by our method were estimated to be half siblings by COLONY, which is consistent with what was observed in simulation studies in Simulated Datasets. Furthermore, most of the relationships that were inferred as half cousins by our

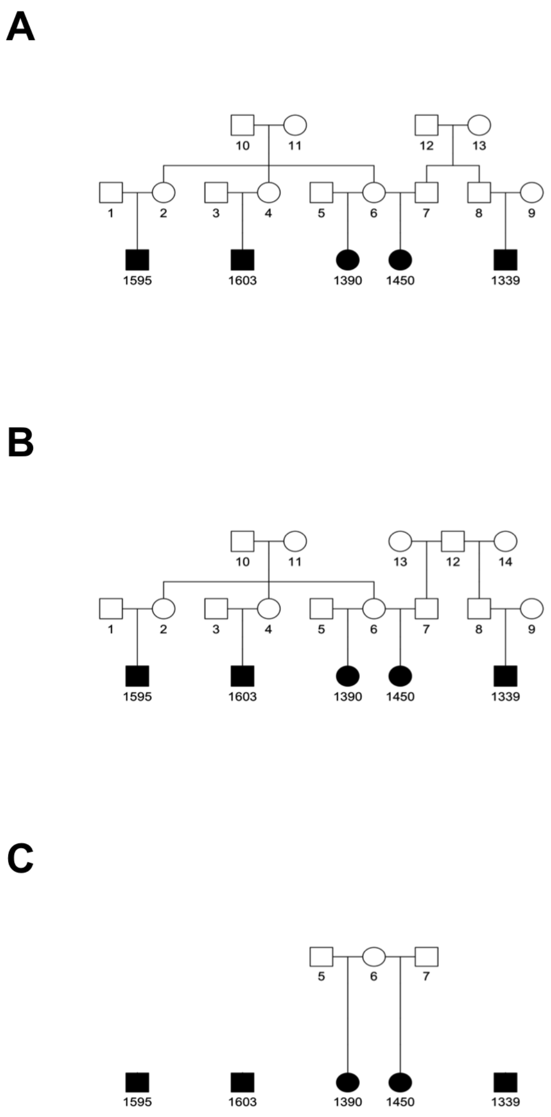

Figure 3.10: Estimated pedigrees of five sampled individuals in the sparrow dataset. (A) Pedigree with estimated posterior probability of .77. (B) Pedigree with estimated posterior probability of .23. (C) Most likely pedigree estimated by COLONY, but whose posterior probability was zero in our method.

method were classified as unrelated by COLONY.

Table 3.5: Comparison of Pairwise Relationship Classification by MCMC and COLONY.

		COLONY [a]		
		FS	HS	UR
MCMC [b]	FS	33	0	0
	HS	0	23	0
	UR	0	1	2909
	FC	0	15	37
	HC	1	4	57

[a] Inference was based on the full likelihood method.

[b] The likelihoods were computed by [3] for linked markers and the inference allowed pedigrees going up to 2 generations back in time (i.e. up to first cousins).

3.4 Discussion

We have shown that, given enough marker information, our method is able to jointly estimate N_e and relationships up to first cousins accurately and efficiently. Unlike existing pedigree inference methods, our method not only allows estimation of pedigrees and N_e, but also provides an uncertainty measure on the estimates via posterior probabilities. Furthermore, our method provides a framework for incorporating different types of population models in the prior for the pedigree, which can potentially allow us to estimate other population parameters, such as migration rates between subpopulations.

Our method also improves upon one of the most widely used pedigree reconstruction programs, COLONY, by estimating relationships beyond sibships. This not only expands the types of pedigrees we can infer but also increases the accuracy of sibship inference. In particular, first cousins were often misclassified as half siblings if the estimation method did not allow inference of cousins. For example, about 44 percent of half siblings estimated by COLONY using 10,000 SNPs were actually first cousins (Table 3.1c). Furthermore, we showed that N_e can be underestimated if the sample contains cousins but the pedigree inference is restricted to sibships only (Figure 3.4). By explicitly including first cousins in the inference, our method was able to infer half siblings with higher precision (Table 3.1a), as well as estimate N_e more accurately (Figure 3.3). However, we note that the problem persists when the sample contains relatives more distant than first cousins. When datasets contained second cousins, for example, they were often estimated as half first cousins–the most distant relationship our method is designed to estimate–and consequently caused a downward bias in N_e estimates. Therefore, we must use caution in interpreting inferred half cousins, as the true relationship could be more distant, and use the simulation method discussed in Effect of Presence of Relatives Beyond First Cousins to correct for potential bias in N_e estimates.

We note that the performance of our method relies heavily on the accuracy of pairwise likelihoods. The accuracy of pairwise likelihoods depends on many factors, such as marker density, level of linkage disequilibrium, sequencing error rates, and population allele frequency estimates. Ignoring the linkage between markers, in particular, significantly decreased the power to detect first cousins (File S3). Due to linkage, close relatives such as first cousins are expected to share, with high probability, long IBD segments that are on the order of megabases in length, although the probability of IBD per marker is relatively low [10]. The presence of such long IBD segments should make detecting relatives quite easy even though identifying the exact relationship can be more difficult. Treating the markers as independent, however, does not take advantage of the presence of long IBD segments and thus decreases our ability to detect relatives (Tables 3.1b, 3.1c). Therefore, likelihood computation methods, such as [3], that take into account the linkage information between markers should be used instead for detecting relatives, and naturally, for pedigree inference as well.

Marker type and density also have a significant impact on the quality of pairwise likelihoods. We have seen that using 20 highly informative microsatellites performed worse than using 10,000 SNPs. The accuracy rates of COLONY (Table 3.2c) suggest that the use of microsatellites to estimate sibships might be misguided in practice since first cousins can often be misclassified as half siblings in methods that do not explicitly model first cousins. Furthermore, microsatellites may not provide enough information to easily distinguish between full and half siblings (Table **??**). Also, 20 microsatellites with 10 alleles of equal frequency in our simulations is more generous than what is available in many real datasets, and the performance on less informative datasets is likely to be worse than what was shown in this study. We note that finding the best ways to address the various challenges in pairwise likelihood computation is an active area of research and requires further investigation.

There are limitations to our method that require further work. Our method does not support pedigrees that contain cycles, except those caused by full sibling relationships. More specifically, we do not consider pedigrees that are inbred or have complex, cyclic relationships such as double first cousins. A simulation study by [39] suggests that in the presence of inbred individuals, the method will tend to estimate individuals to be more genealogically closer than they actually are (e.g. first cousins estimated half siblings). Furthermore, our method assumes that all samples belong in a single generation, which may not typically be true for many real datasets. This may be addressed by adding updates in the MCMC that allow sampled individuals to move between generations. Furthermore, our method does not yet scale up to sample sizes typical of GWAS as the number of pairwise comparisons still increases rapidly with sample size. One possible approach to address this issue is partitioning the sample into smaller sets using methods such as [45] and estimating the pedigrees for each smaller subset of individuals.

Overall, our method provides a way to jointly estimate pedigrees and N_e, and measure the uncertainty of the estimates in a computationally efficient way. Importantly, our method also provides a basic framework for estimating demographic parameters of the current population from pedigrees–analogous to population genetic methods based on coalescent trees–thus opening up new possibilities for learning about the demographic history of the recent past.

Our software is available for download at https://github.com/amyko/mcmcPed.

Chapter 4

Estimation of Neanderthal Admixture Tracts and Time

4.1 Introduction

Neanderthals, a group of archaic hominins that lived until about 40,000 years ago, have been of great interest in studies of human demographic history. Archaeological record suggests that anatomically modern humans and Neanderthals coexisted in parts of Europe and Asia for as long as 5,000 years before the disappearance of Neanderthals [27]. Previous studies have found signatures of Neanderthal admixture in the genomes of Europeans and Asians, but not in sub-Saharan Africans [53, 77, 58], consistent with the hypothesis that early humans interbred with Neanderthals in Eurasia after African and non-African populations had split (Figure 4.1). Consequently, non-African haplotypes that are shared with Neanderthals but absent from Africans are likely to be of Neanderthal ancestry–an observation that forms the foundation of the admixture inference method we will discuss in Section 4.2.

Natural questions that arise about the admixture event between modern humans and Neanderthals include: 1) when did inbreeding between the two populations occur? (admixture time) 2) how much of the modern human genome is of Neanderthal ancestry? (admixture proportion), and 3) what parts of the modern human genomes are of Neanderthal ancestry? (admixture tracts). In this chapter, we discuss a Hidden Markov Model (HMM) method to address these questions and analyze ancient human samples, Kostenki 14 (K14) [59] and individuals from Sunghir [60], to estimate the admixture time between early modern humans and Neanderthals.

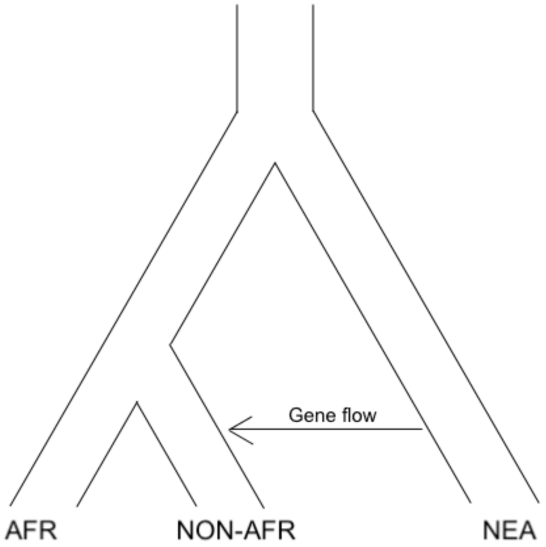

Figure 4.1: Demographic model for admixture between modern humans and Neanderthals. In this model, the gene flow between Neanderthals and non-Africans occurs after non-African and African populations split.

4.2 Hidden Markov Model for Estimating Admixture Parameters

Model

Here we describe a Hidden Markov Model (HMM) method to infer Neanderthal admixture time, proportion, and tracts. The HMM has two states representing the ancestry in a haploid genome: Neanderthal or human. The objective is to estimate the parameters of this two-state HMM and use the subsequent posterior decoding to infer admixture tracts. In contrast to some previous methods [52, 58], the emission probabilities are not estimated using simulated training data, but are instead estimated directly from observed data.

Consider a test haplotype $h = (h_1, ..., h_k) \in \{0,1\}^k$ of length k from an admixed population (e.g. Europeans), where 1 denotes a derived allele and 0 ancestral. We denote $z = (z_1, ..., z_k) \in \{0,1\}^k$ as the hidden ancestry vector for the test haplotype. We assign $z_i = 0$ if site i is of Neanderthal ancestry, and $z_i = 1$ if it is of modern human ancestry. We also define an observation vector $y = (y_1, ..., y_k) \in \{0,1\}^k$ as follows:

$$y_i = \begin{cases} 1, & \text{if } h_i = 1, f_i^{afr} < \epsilon, f_i^{nea} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where $f_i^{afr}$is the derived allele frequency estimated from a panel of individuals from an unadmixed population such as the Yoruba (YRI), and f_i^{nea} is the derived allele frequency estimated from a panel of Neanderthal genomes. In other words, the site is informative of Neanderthal ancestry if it is 1) derived in the test haplotype, 2) the derived allele is present in the Neanderthal panel, but 3) absent in the panel of unadmixed individuals. The parameter $\epsilon \geq 0$ is added to account for possible sequencing errors in the unadmixed population. We consider as missing data sites that are fixed in the test population or sites for which the observed state cannot be determined.

Using these hidden and observed states, we define a two-state, time-homogeneous HMM. The ancestry vector z is the sequence of hidden states along the test haplotype and y is its associated sequence of observed states (Figure 4.2). Following a previously developed model by [22], the transition rates between the hidden states are given by

$$\begin{aligned} Pr(z_j = 1 | z_{j-1} = 0) &= rm(t-1), \\ Pr(z_j = 0 | z_{j-1} = 1) &= r(1-m)(t-1), \end{aligned} \tag{4.2}$$

where m is the admixture proportion, t is the admixture time, and r is the recombination rate. In other words, admixture tracts become shorter over time due to recombination after the initial admixture event (Figure 4.3), so the transition rates between the two states is proportional to the time since admixture. It is worthwhile to note that the assumption of exponentially distributed tract lengths given by Equation 4.2 is not quite true, but is a fair assumption for the level of divergence and admixture between archaic and modern humans [42].

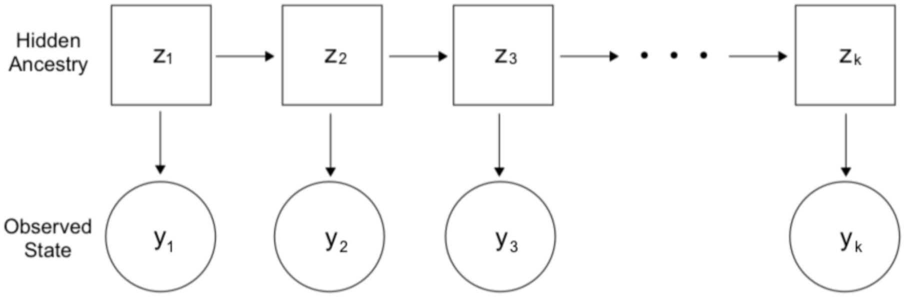

Figure 4.2: The sequence of hidden states is the ancestry along a test haplotype (0 for human ancestry and 1 for Neanderthal ancestry). The observed states are defined in Equation 4.1.

To estimate the parameters of the HMM, we first fix the transition probabilities using initial guesses for m and t; recombination rate r is assumed to be known, a realistic assumption for human data sets. Then we compute the product of the likelihoods of all individuals

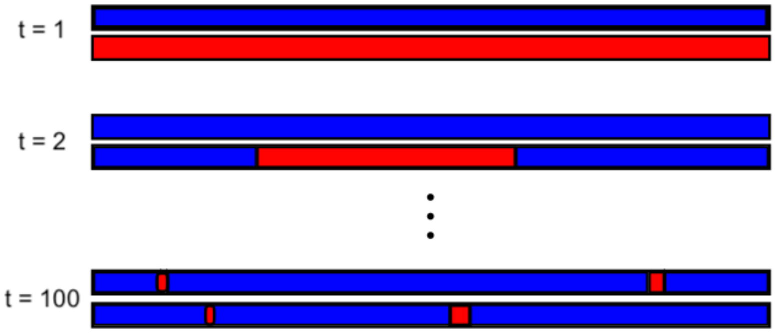

Figure 4.3: Admixture tracts (red) become smaller over time due to recombination.

in the study sample, $\prod_{i=1}^{n} Pr(y^i|r,m,t)$, using the forward algorithm [54] and maximize it with BFGS to estimate the emission probabilities. Here, y^i denotes the observation vector for the ith individual. We then fix the emission probabilities at the estimated values and maximize the likelihood of each individual to estimate the transition probabilities. Transition probabilities then give the maximum likelihood estimates of the admixture proportion and admixture time. More specifically, we estimate admixture proportion, m, with

$$\hat{m} = \frac{\widehat{Pr}(z=1|z=0)}{\widehat{Pr}(z=1|z=0)+\widehat{Pr}(z=0|z=1)} \tag{4.3}$$

(i.e. proportion of times the HMM jumps into Neanderthal state), and admixture time with Equation 4.2.

We then use the Viterbi algorithm [73] to infer the most likely sequence of local ancestry along the test haplotype using the estimated HMM parameters. To control for false detection of admixture tracts, we filter out tracts that are closer to any of the haplotypes in the panel of unadmixed individuals (e.g. YRI) in the corresponding regions. Here, we use the Hamming distance to measure the distance between haplotypes.

Simulated Data

We simulated SNP data using a modified version of msHOT [26] following the demography shown in Figure 4.1. We simulated 100 non-African, 100 African, and 2 Neanderthal haploid genomes, where each haploid genome consisted of 100 independent one mega-base segments. Following the demographic parameters discussed in [58], we set the admixture time and proportion to 1,900 generations ago and .02, respectively. We set the split time between humans and Neanderthals to 13,000 generations ago, and between Africans and non-Africans to 2,500 years ago. The mutation rate in humans was set to 1.25e-8 per site per generation. We then pruned the markers to filter out sites that were in high linkage disequilibrium.

Performance on Simulated Data

Figure 4.4 shows the distribution of maximum likelihood estimates of admixture proportion from 100 non-African genomes. The time estimates had a mean value of .021 and standard deviation of .001. Figure 4.5 shows the distribution of admixture time estimates, with mean of 1,980 and standard deviation of 281. Figure 4.6 shows the precision-recall curve for admixture tract inference, where the color indicates the threshold probability at which a site was called as having Neanderthal ancestry. Overall, the precision rates were high at relatively high recall rates. However, we note that further study is required to test the robustness of the method to various factors such as the level of linkage disequilibrium in the data.

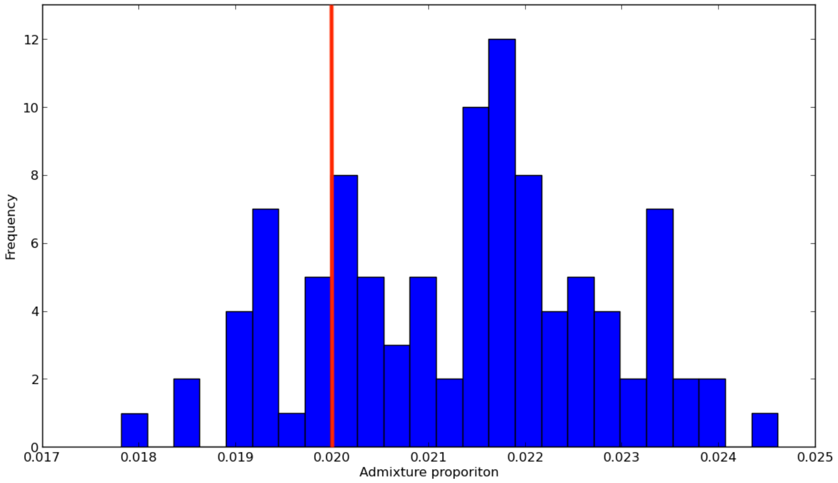

Figure 4.4: Distribution of admixture proportion estimates.

4.3 Application to Real Data

Kostenki 14

The Kostenki 14 (K14) skeleton, one of the oldest anatomically modern human fossils in Europe, was excavated in 1954 in Kostenki-Borshcevo in Russia and was dated to be 33,250 ± 500 radiocarbon years old–about 36,000 to 39,000 calendar years [59]. After filtering, the data set consisted of 148.9 million unique reads, corresponding to an average depth of coverage of 2.42X. We used the HMM discussed in Section 4.2 to estimate the maximum likelihood (ML) admixture time between Neanderthals and modern humans using the K14 sample.

The algorithm had two steps. In the first step, emission probabilities were estimated from the samples in the 1000 Genomes Project [11]. We considered only two possible type

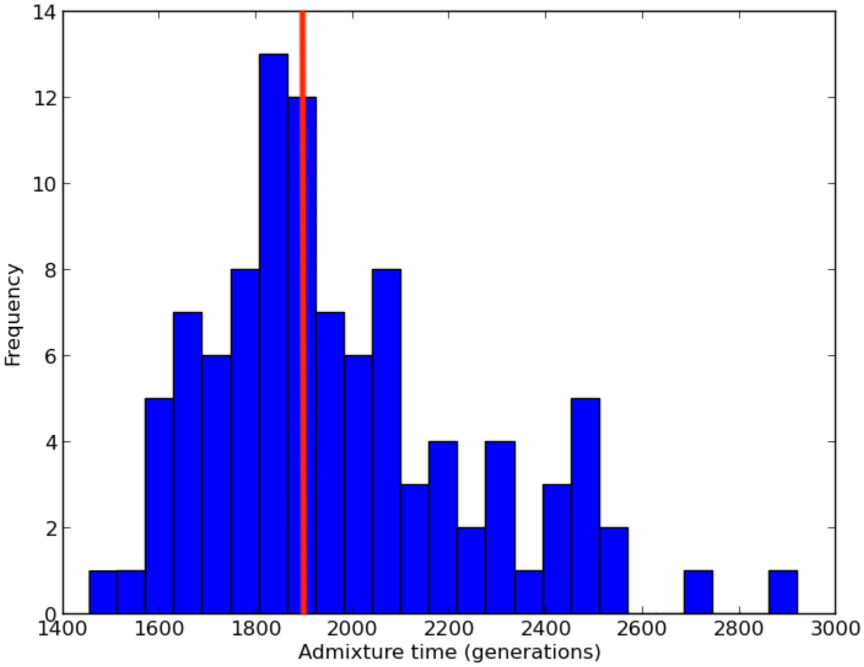

Figure 4.5: Distribution of admixture time estimates.

of emissions: (A) sites variable in humans (phased 1000 Genomes v.2) in which the focal haplotype had a derived allele as determined by the 4-way EPO alignments, in which Neanderthals contained at least one derived alleles, and in which all reference African populations were invariant ancestral; and (B) variable sites (including Neanderthals) that do not fulfill the conditions in (A). Invariable sites and sites that did not have information to determine whether condition (A) was true were considered missing data. The HMM was then applied to these data, after removing sites in linkage disequilibrium, by multiplication of the likelihood function among individuals, using transition probabilities given by 4.2. The likelihood function was calculated using standard algorithms and optimized using the BFGS algorithm. Note that individuals may not be independent, they may share the same tracts. However, the estimator should still perform well as a composite likelihood estimator for obtaining point estimates. Using this method we obtained emission probabilities of 0.0155 and 6.67e-9 for observations of type A when in the Neanderthal and human state, respectively.

The second step was to estimate transition probabilities in K14 using maximum likelihood on the same set of sites as those used for estimating emission probabilities. The transition probability estimates directly provide ML estimates of the admixture time by the invariance principle of maximum likelihood.

To enable application to the low-coverage, unphased ancient DNA such as K14, we considered a modified state space in which we only use at most one read for the K14 individual

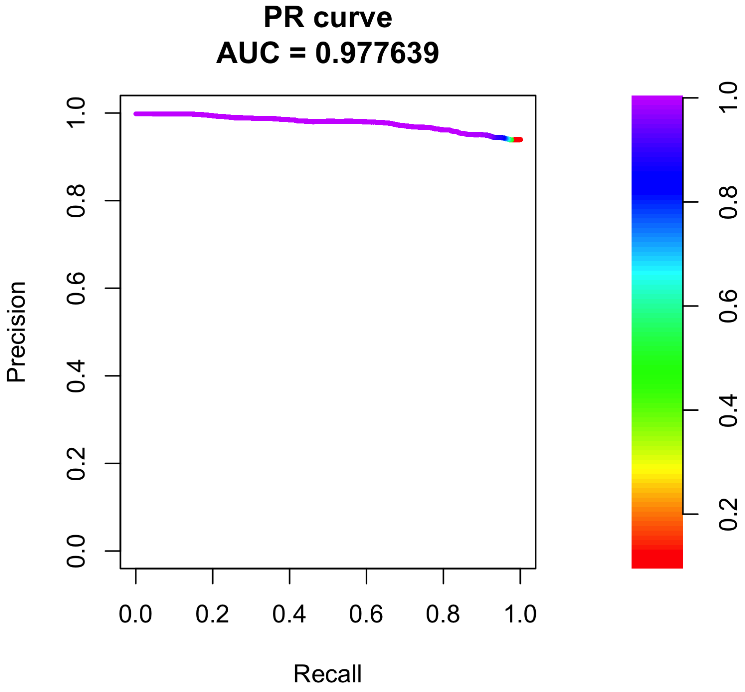

Figure 4.6: Precision-recall curve for detecting Neanderthal tracts. The color represents the threshold probability at which Neanderthal ancestry was called.

for each position. When more than one read was available for a position, we sampled one uniformly at random. As a first approximation, we assumed that K14 was not homozygous for a Neanderthal segment any place in the genome. The state space is then reinterpreted as having two states: a state in which K14 is heterozygous for a Neanderthal haplotype ($z_i^* = 1$) and a state in which both alleles in K14 are of human origin ($z_i^* = 0$). The emission probability for state $z_i^* = 0$ is then identical to the previously described emission probability for $z_i = 0$. However, for state $z_i^* = 1$, the emission probabilities are obtained as a 50:50 mixture between those of $z_i = 0$ and $z_i = 1$ in the previous analysis. Sites that were invariable in the 1000G data were considered missing data. This largely eliminated the effect of errors, and also eliminated the effect of unique mutations on the K14 lineage, ensuring that the previously estimated emission probabilities are applicable to the K14 individual.

Using these methods and assuming generation time of 29 years and a recombination rate of 1.26e-8 per base pair per generation, we obtained an estimate of the admixture time of approximately 16,600 year before K14 was deposited. Therefore we estimate that the time of admixture between Neanderthals and early humans to be approximately 54,000 years ago.

Sunghir

Here we analyzed the ancient human samples from Sunghir, sequenced at an average depth of coverage between 1X and 11X by [60]. The samples consisted of four individuals whose skeletons were excavated in Sunghir, an Upper Paleolithic burial site in Russia. The age of the individuals was dated to be between 34,600 and 33,600 years.

We used a modified version of the HMM method described in Section 4.2 to date the Neanderthal admixture event in Sunghir samples. The strategy we used here was slightly different than in the original version. Instead of estimating the age directly by training the HMM on the data, we instead used fixed parameters for the HMM based on simulations and analytic predictions for a fixed admixture time and proportion. We then inferred tracts using this model, and used parametric simulations to obtain an estimate of the admixture time from the inferred tracts. This procedure corrects for any biases associated with the tract inference because we use the exact same procedure for inferring tracts on the simulated and the real data. We chose this approach because it allowed us to directly incorporate SNP filtering and biases due to inadequate modelling of background LD in the inference procedure, thereby producing a more robust method for inferring admixture times.

We used an HMM with two states 1) sites heterozygous for the Neanderthal allele and 2) sites homozygous for the human alleles. Similar to the model in Section 4.3, this construction allowed us to analyze unphased data. We analyzed sites from the 1000 Genomes phased data that were variable in the CEU and YRI samples. We then divided the variable sites into two categories: 1) sites in which the focal haplotype had a derived allele, Neanderthals have at least one derived allele, and YRI have zero derived alleles; 2) variable sites that do not meet the conditions in (1). Invariable sites in CEU and YRI, and sites for which we could not determine the conditions in (1) were considered missing data. The emission probabilities, estimated from simulated data, were 0.0162 and 0.000410 for observing a site of type (1) given the Neanderthal state and human state, respectively.

The transition rates were computed from $Pr(z = 0|z = 1) = m(t - 1)$ per Morgan and $Pr(z = 1|z = 0) = (1-m)(t-1)$ per Morgan, where $m = .03$ and $t = 728$ generations. (Note: 728 generations = 1900 generations - sample age of 34,000 years in generations, where we use 29 years per generation.) The prior of being in the Neanderthal state was set at $m = .03$. For both the real and simulated data, CEU and YRI sample sizes of 85 and 88, respectively, were used to identify the variable sites within the union of the two populations. We then LD-pruned these sites at $r^2 = 0.7$ via a sliding window approach (window size = 200kb). This procedure was applied to all four Sunghir samples.

We simulated data using SLiM [47] under demographic histories of Neanderthal admixture into Sunghirs with varying admixture times. For each admixture time, 400 admixed chromosomes, with 100Mb/chromosome were simulated. A recombination rate of 1.3e-8, a mutation rate of 1.5e-8, and an admixture proportion of 0.03 were used in the simulations. The inference procedure is robust to the assumed admixture proportion as long as it is relatively small. We fit an exponential distribution with parameter λ, truncated at 10 kb, to both the real and the simulated data. We then identified the simulated value of the admixture time that generated the same value of λ as observed in the real data. To obtain approximate confidence intervals (CIs) we block- bootstrap inferred admixture tracts.

The results are presented in Figure 4.7. The mean value of λ for the real data was 1.05e-05 with an approximate 95 percent confidence interval (CI) of [1.026 e-05, 1.067 e-05], resulting in an admixture time estimate of 770 generations before the age of the sample with an approximate 95 percent CI of [755, 786].

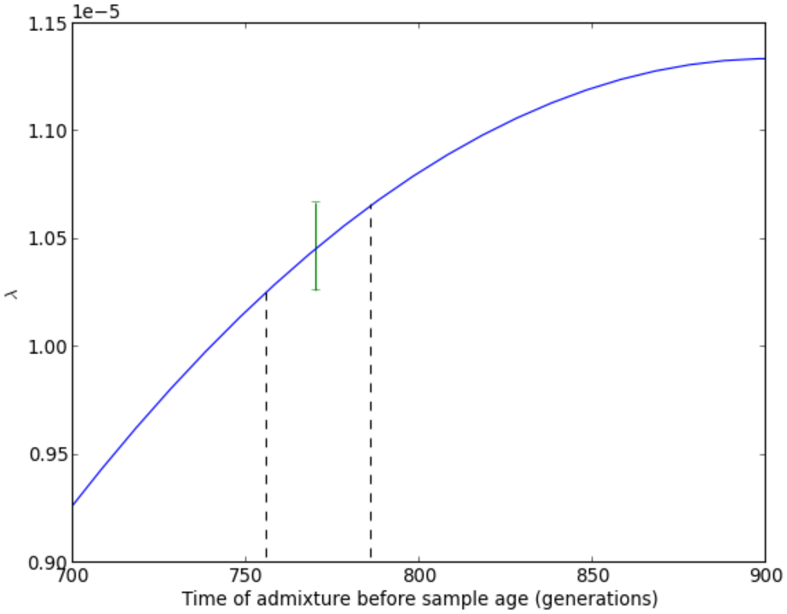

Figure 4.7: Neanderthal admixture time estimate for Sunghir using tract length distribution. The green vertical line indicates the 95 percent confidence interval for λ in the real data. The dotted vertical lines indicates the corresponding interval for admixture time in generation.

Appendix A

Transitions Between Pedigree Graphs

This section describes the possible transitions between pedigree graphs. For each of the moves described below, we reject the move if the proposed pedigree is not a valid pedigree. We denote nodes that represent unsampled individuals as "ghost nodes" and nodes that represent sampled individuals as "sampled nodes."

1. Link: join two pedigrees

   a) Choose a random pair of nodes i and j, where each pair has an equal probability of being chosen.

   b) Choose target depth d, where d is drawn from the geometric distribution. The target depth is the depth at which i and j will share a common ancestor. If the target depth exceeds the maximum depth, reject.

   c) Choose the target sex for the would-be common ancestor, male or female, with equal probability. Follow a random path from i to reach an ancestor of the target sex at depth d. That is, starting from i, choose a male or female parent with equal probability, and move up through the pedigree until depth $d-1$ is reached; at depth d, choose the ancestor with the target sex. Repeat the process for j.

   d) Merge the two ancestors of i and j. If the merging creates an invalid pedigree, reject.

2. Cut: detach a subpedigree.

   a) Choose a random node i. With equal probability, choose the target sex of the parent from which i will be cut. If the parent of the chosen sex is not present in the pedigree, reject.

   b) Remove the edge between i and the parent of the target sex.

3. Split: remove a subset of children from its parent pedigree.

a) Choose a random node i. Choose a random subset of i's children, where each subset has an equal probability of being chosen.

b) Remove the edges between the selected children and i. Make a new ghost parent for the selected children.

4. CutLink: combine Cut and Link in a single move.

5. SplitLink: combine Split and Link in a single move.

6. Contract: merge a child node with its parent node.

    a) Choose a random node i. If i does not have exactly one ghost parent of the same sex, reject.

    b) Remove an edge between i and its ghost parent, merging the two nodes. That is, delete the ghost parent; with equal probability, either shift the subpedigree containing i up one depth or shift the parent subpedigree down one depth; make the grandparents of i its new parents. If the merging creates an invalid pedigree, reject.

7. Stretch: Add an edge between a child and its parent.

    a) Choose a random node i. If i does not have any parent nodes, reject.

    b) Disconnect i from its parents; with equal probability, either shift the resulting subpedigree containing i down one depth or shirt the subpedigree containing its parent up one depth.

    c) Make a new ghost node that has the same sex as i and make the ghost node the new parent parent of i.

    d) Make the old parents of i new parents of the ghost node. If this results in an invalid pedigree, reject the move

8. SwapDescAnc: swap ancestor and descendent.

    a) Choose a random sampled node i.

    b) Choose a random sampled node j among is sampled ancestors who have the same sex as i.

    c) Swap i and j. If an invalid pedigree is generated, reject the move.

9. ShiftClusterLevel: shift the depth of family cluster

    a) Choose a random node i and get the nodes connected to i, including i.

    b) Choose a new target base depth d, drawn from the geometric distribution.

    c) Adjust the depth of each node belonging to i's cluster so that the lowest depth of the cluster is d. If the shift violates the depth constraint, reject the move.

10. SwitchSex: switch the sex of an individual(s).

    a) Choose a random node i. If i's sex is fixed (e.g. i is a sampled node and has a fixed sex, or has a spouse with a fixed sex), reject the move.

    b) Switch i's sex.

    c) Switch the sex of all nodes wth conflicting sexes caused by step (b). For example, if i has a spouse, the spouse must now also switch his/her sex.

11. FullSiblingsToParentOffspring: change full sibling relationship to parent-offspring relationship.

    a) Choose a random node i. If i does not have any full siblings, reject.

    b) Choose a random node j from the set of full siblings of i.

    c) Disconnect i from its parents. With equal probability, either shift i's cluster down one depth or shift j's cluster up one depth.

    d) Make j parent of i. If this results in an invalid pedigree, reject the move.

12. ParentOffspringToFullSiblings: change parent-offspring relationship to full sibling relationship.

    a) Choose a random node i. If i does not have exactly one parent, reject the move. Call this parent p.

    b) Detach i from p. With equal probability, shift either i's cluster up one depth or shift p's cluster down one depth.

    c) Form a full sibling relationship between i and p. If this results in an invalid pedigree, reject.

13. FullSiblingsToSelf: merge two full sibling nodes into one node.

    a) Choose a random node i. If i does not have any full siblings with the same sex as i, reject.

    b) Choose a random node j among the nodes who are full siblings with i and who also have the same sex as i.

    c) Merge i and j. If this results in an invalid pedigree, reject the move.

14. SelfToFullSiblings: split a single node into two nodes that form a full sibling relationship.

    a) Choose a random node i.

    b) Make a ghost node j that has the same sex as i. Set parents of j to parents of i.

    c) Assign a random set of i's children to j. If this results in an invalid pedigree, reject.

15. HalfSiblingsToParentOffspring: change half sibling relationship to parent-offspring relationship.

    a) Choose a random node i. If i does not have any half siblings, reject.

    b) Choose a random node j from the set of half siblings of i.

    c) Detach i from the parent whose sex is the same as j.

    d) With equal probability, either shift i's cluster down one depth or shift j's cluster up one depth.

    e) Make j parent of i. If this results in an invalid pedigree, reject.

16. ParentOffspringToHalfSiblings: change parent-offspring relationship to half sibling relationship.

    a) Choose a random node i. If i does not have any parents, reject the move.

    b) Choose one of i's parent at random. Call this node j.

    c) Detach i from j. With equal probability, shift either i's cluster up a depth or shift j's cluster down a depth.

    d) Form a half sibling relationship between i and j. If this results in an invalid pedigree, reject.

17. ParentOffspringToOffspringParent: change parent-offspring relationship to offspring-parent.

    a) Choose a random sampled node i.

    b) Choose a random node j among i's children.

    c) Disconnect i from its children and shift i down two depths.

    d) Make j new parent of i. If this results in an invalid pedigree, reject.

18. OffspringParentToParentOffspring change offspring-parent relationship to parent-offspring .

    a) Choose a random sampled node i.

    b) Choose a random parent of i. Call the chosen j.

    c) Disconnect i from its parents and shift it two depths up.

    d) Make i new parent of j. If this results in an invalid pedigree, reject.

19. UncleToNephew: change uncle-nephew relationship to nephew-uncle relationship.

    a) Choose a random node i.

    b) Choose a random node j among the nephews/nieces of i.

c) Disconnect i from its parents. With equal probability, either shift i's cluster down two depths, shift j's cluster up two depths, or shift i's cluster down one depth and j's cluster up one depth.

d) Make a new parent (ghost node) for i. Make this ghost parent and j full siblings. If this results in an invalid pedigree, reject.

20. NephewToUncle: change nephew-uncle relationship to uncle-nephew relationship.

a) Choose a random node i.

b) Choose a random node j among the uncles/aunts of i.

c) Disconnect i from its parents. With equal probability, either shift i's cluster up two depths, shift j's cluster down two depths, or shift i's cluster up one depth and j's cluster down one depth.

d) Choose a random parent of j. Make this parent and j full siblings. If this results in an invalid pedigree, reject.

21. HalfSiblingsToGP: change half sibling relationship to grandparent-grandchild relationship.

a) Choose a random node i.

b) Choose a random node j among the half siblings of i.

c) Disconnect i from j. With equal probability, either shift i's cluster down two depths, shift j's cluster up two depths, or shift i's cluster down one depth and j's cluster up one depth.

d) Make a new parent for j (sex of parent is randomly chosen). Call this new parent p. Make j parent of p. If this results in an invalid pedigree, reject.

22. GPtoHalfSiblings: change grandparent-grandchild relationship to half sibling relationship.

a) Choose a random node i.

b) Choose at random parent node j from which i will be cut.

c) Choose a random node k among the parents of j.

d) Disconnect i from j. With equal probability, either shift i's cluster up two depths, shift j's cluster down two depths, or shift i's cluster up one depth and j's cluster down one depth.

e) Make i and k half siblings via a common parent whose sex is randomly chosen. If this results in an invalid pedigree, reject.

Appendix B

Command Lines for Running External Softwares

1. To run PLINK to estimate population allele frequencies:
   ```
   plink –tfile 100founders –freq –out 100founders,
   ```
   where100founders is the name of tped and tfam files containing 100 unrelated diploid individuals. This generates 100founders.frq that contains allele frequency estimates.

2. To run PLINK to estimate IBD proportions:
   ```
   plink –tfile testFile –genome –read-freq 100founders.frq –out testFile,
   ```
   where testFile is the name of tped and tfam files containing test samples and 100founders.frq are allele frequency estimates from above. This generates testFile.genome containing IBD proportion estimates.

3. To run RELATE to estimate IBD proportions and IBD segments:
   ```
   relateHMM -g geno -p pos -c chr -o options -d indiv -post postout -k kout,
   ```
   where geno is the genotype file containing test individuals and unrelated founder individuals. The unrelated founders are only used to estimate the allele frequencies and haplotype frequencies. Details for the remaining files can be found in the RELATE manual (see main text for citation). The parameters in the option file are shown below:
   ```
   1 #1=allpairs 0=normal run
   0 #pair[0]
   1 #pair[1]
   0 #double recombination 0 #LD=0=rsq2 LD=1=D
   0.001 # alim[0]
   5 # alim[1]
   0 # doParameter calculation (pars)
   0 # par[0] = a this is only used if doParameter is set to 1
   0 # par[1] = k2 this is only used if doParameter is set to 1
   0 # par[2] = k1 this is only used if doParameter is set to 1
   ```

```
1 # ld_adj
0.01 # epsilon
1 # back
0 # doPrune
0 # prune_value
0 # fixA
0.0 # fixA_value
0 # fixK2
0.0 # fixk2_value
0 # calculateA
0.013 # phi_value
0.1 # convergence_tolerance
5 # times_to_converge
10 # times_to_run
2 # back2
```

4. To run PRIMUS:
   ```
   run_PRIMUS.pl -p IBDfile –sex_file sexFile –plink_ex myPlinkPath –no_PCA_plot –no_IMUS -o myOutPath –max_gens 5,
   ```
   where IBDfile contains IBD proportion estimates from either PLINK or RELATE; sexFile contains sex information.

5. To run ERSA:
   ```
   ersa –control_files=controlFile –segment_files=segmentFile –model_output_file=ersaModelFile –output_file=ersaResultFile–confidence_level .95 –rec_per_meiosis 39,
   ```
   where controlFile contains IBD segment estimates by RELATE for 100 founder individuals; segmentFile contains IBD segment estimates for the test samples.

6. To run PADRE:
   ```
   run_PADRE.pl –ersa_model_output ersaModelFile –ersa_results ersaResultFile –project_summary summaryFile –degree_rel_cutoff 3 –output_dir,
   ```
   where the file paths point to appropriate output directories for ERSA and PRIMUS.

7. To run COLONY for 10K SNPs:
   ```
   'filename' !Output file name
   50 ! Number of offspring in the sample
   10000 ! Number of loci
   1234 ! Seed for random number generator
   1 ! 0/1=Not updating/updating allele frequency
   2 ! 2/1=Dioecious/Monoecious species
   0 ! 0/1=No inbreeding/inbreeding
   0 ! 0/1=Diploid species/HaploDiploid species
   ```

```
0 0 ! 0/1=Polygamy/Monogamy for males & females
0 ! 0/1=Clone inference =No/Yes
0 ! 0/1=Full sibship size scaling =No/Yes
0 !  0,1,2,3=No,weak,medium,strong sibship size prior; mean paternal & meteral sibship size
0 ! 0/1=Unknown/Known population allele frequency
1 ! Number of runs
2 ! Length of run
0 ! 0/1=Monitor method by Iterate#/Time in second
100000 ! Monitor interval in Iterate# / in seconds
0 ! non-Windows version
1 ! 0/1 pairwise/Fulllikelihood
2 ! 1/2/3=low/medium/high Precision for Fulllikelihood
m@ !Marker names
0@ !Marker types, 0/1 = codominant/dominant
0@ !Allelic dropout rate
0.01@ !false allele rate
0 0 !prob. of dad/mum included in the candidates
0 0 !numbers of candiadte males & females
0 0 !#known fater-offspring dyads, paternity exclusion threshold
0 0 !#known moter-offspring dyads, maternity exclusion threshold
0 !#known paternal sibship with unknown fathers
0 !#known maternal sibship with unknown mothers
0 !#known paternity exclusions
0 !#known maternity exclusions
0 !#known paternal sibship exclusions
0 !#known maternal sibship exclusions
```

8. To run COLONY for 20 microsatellites:
   The parameters were the same as they were for Simulation A, except the number of loci was set to 20 and the allele dropout rate was set to 0.05.

Appendix C

MCMC Updates

1. Link: join two pedigrees

   a) Choose a random pair of nodes i and j. Choose target depth k, drawn from a geometric distribution with $p = .5$. k is the depth at which i and j will share a common ancestor. If k is larger than the maximum depth of the pedigree, reject the move.

   b) Choose sex s, male or female, of the would-be common ancestor with equal probability. Take a random path from node i up to the ancestor of sex s and depth k, choosing either the mother or the father at each step with equal probability. Do the same for node j.

   c) At depth k, merge the two ancestors of i and j.

2. Cut: detach a child and its subpedigree from a parent

   a) Choose node i at random. Choose sex s, male or female, at random, which is the sex of the parent from which i will be cut.

   b) Delete the edge between i and the parent of sex s.

3. Split: detach a set of children and its subpedigrees from a parent

   a) Choose node i at random. Choose a random set of i's children, where each set has an equal probability of being chosen.

   b) Delete the edges between i and the chosen children. Make a new parent node j and connect j and the chosen children.

4. Switch Sex: switch the sex of a node

   a) Choose a random node i. Reject the move if i's sex cannot be changed (i.e. i is sampled and has a known sex; or i has a spouse with a fixed sex).

b) If i is female, switch its sex to male. Vice versa if i is male. If this sex change conflicts with the sex of other nodes, switch the sex of the other nodes as well. (e.g. if i has a spouse, then the spouse must switch its sex as well).

5. Full Sibs to Self: merge a pair of full sibling nodes into one node

   a) Choose a random node i. Choose at random node j among full siblings of i whose sex is the same as that of i. If no such node exists or if both i and j are sampled nodes, reject the move.

   b) Merge i and j.

6. Self to Full Sibs: split a node into a pair of full siblings

   a) Choose a random node i. Make a new node j, where the sex is the same as that of i.

   b) Choose a random set of i's children, where each set has an equal probability of being chosen. Remove edges between the chosen children and i, and insert edges between the children and j.

   c) Make i and j full siblings (i.e. make them share the same mother and father).

7. Self to Parents: split one parent into a pair of parents

   a) Choose a random node i. If i does not have exactly one parent, reject the move. Let $p1$ be the parent of i.

   b) Make a new node $p2$ and set its sex to the opposite of $p1$'s sex. Set $p2$ to be the other parent of i.

   c) Choose a random set of $p1$'s parents, where each set has an equal probability of being chosen. Remove the edges between the chosen nodes and $p1$, and insert edges between the nodes and $p2$ (i.e. transfer the chosen nodes from $p1$ to $p2$).

8. Parents to Self: merge two parents into one node

   a) Choose a random node i. Reject if it doesn't have exactly 2 parents: $p1$ and $p2$.

   b) Choose sex s–male or female–at random.

   c) Merge $p1$ and $p2$ and set the sex to s.

9. MaternalHC to PaternalHC: change maternal half cousins into paternal half cousins, and vice versa.

   a) Choose a random node i. Choose a random of set of maternal (or paternal) half cousins of i. If no such nodes exist, reject the move.

   b) Modify the graph so that the set of maternal (or paternal) half cousins become paternal (or maternal) half cousins of i.

10. MaternalHS to PaternalHC: change maternal half sibs into paternal half sibs, and vice versa.

    a) Choose a random node i. Choose at random node j among the maternal (or paternal) half siblings of i. If no such node exists, reject the move.

    b) Modify the graph so that i becomes a paternal (or maternal) half sibling of j.

11. Update α

    a) Given $\alpha_{current}$ and for some fixed variance σ_α^2, draw α_{new} from $N(\alpha_{current}, \sigma_\alpha^2)$.

    b) If α_{new} does not lie between the pre-specified bounds $[\alpha_{min}, \alpha_{max}]$, reject the move.

12. Update β

    a) Given $\beta_{current}$ and for some fixed variance σ_β^2, draw β_{new} from $N(\beta_{current}, \sigma_\beta^2)$.

    b) If β_{new} does not lie between the pre-specified bounds $[\beta_{min}, \beta_{max}]$, reject the move.

13. Update N

    a) Given $N_{current}$ and for some fixed variance σ_N^2, draw N_{new} from $N(N_{current}, \sigma_N^2)$. Round $N_{current}$ to be an integer.

    b) If N_{new} does not lie between the pre-specified bounds $[N_{min}, N_{max}]$, reject the move.

For more details of each move and how the Hastings ratios are computed, refer to the source code provided on https://github.com/amyko/mcmcPed.

Bibliography

[1] Gonçalo R Abecasis et al. "Merlin–rapid analysis of dense genetic maps using sparse gene flow trees". In: *Nat Genet* 30.1 (Jan. 2002), pp. 97–101. DOI: 10.1038/ng786.

[2] Jakris Eu-ahsunthornwattana et al. "Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data". In: *PLoS Genet.* 10.7 (July 2014). ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1004445.

[3] Anders Albrechtsen et al. "Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium". In: *Genet Epidemiol* 33.3 (Apr. 2009), pp. 266–74. DOI: 10.1002/gepi.20378.

[4] A. Almudevar. "A simulated annealing algorithm for maximum likelihood pedigree reconstruction". In: *Theor. Popul. Biol.* 63.2 (Mar. 2003), pp. 63–75. ISSN: 0040-5809. DOI: 10.1016/S0040-5809(02)00048-5.

[5] Anthony Almudevar and Eric C. Anderson. "A new version of PRT software for sibling groups reconstruction with comments regarding several issues in the sibling reconstruction problem". In: *Mol. Ecol. Resour.* 12.1 (Jan. 2012), pp. 164–178. ISSN: 1755-098X. DOI: 10.1111/j.1755-0998.2011.03061.x.

[6] Amy D Anderson and Bruce S Weir. "A maximum-likelihood method for the estimation of pairwise relatedness in structured populations". In: *Genetics* 176.1 (2007), pp. 421–440.

[7] Eric C. Anderson and Thomas C. Ng. "Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation". In: *Theor. Popul. Biol.* 107 (Feb. 2016), pp. 39–51. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2015.09.005.

[8] M. S. Blouin. "DNA-based methods for pedigree reconstruction and kinship analysis in natural populations". In: *Trends Ecol. Evol.* 18.10 (Oct. 2003), pp. 503–511. ISSN: 0169-5347. DOI: 10.1016/S0169-5347(03)00225-8.

[9] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4 (2015), p. 7. DOI: 10.1186/s13742-015-0047-8.

[10] NH Chapman and EA Thompson. "A model for the length of tracts of identity by descent in finite random mating populations". In: *Theoretical population biology* 64.2 (2003), pp. 141–150.

[11] 1000 Genomes Project Consortium et al. "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422 (2012), p. 56.

[12] Robert G Cowell. "A simple greedy algorithm for reconstructing pedigrees". In: *Theor Popul Biol* 83 (Feb. 2013), pp. 55–63. DOI: `10.1016/j.tpb.2012.11.002`.

[13] Robert G Cowell. "Efficient maximum likelihood pedigree reconstruction". In: *Theor Popul Biol* 76.4 (Dec. 2009), pp. 285–91. DOI: `10.1016/j.tpb.2009.09.002`.

[14] James Cussens et al. "Maximum likelihood pedigree reconstruction using integer linear programming". In: *Genet Epidemiol* 37.1 (Jan. 2013), pp. 69–83. DOI: `10.1002/gepi.21686`.

[15] Kevin P Donnelly. "The probability that related individuals share some section of genome identical by descent". In: *Theoretical population biology* 23.1 (1983), pp. 34–63.

[16] R C Elston and J Stewart. "A general model for the genetic analysis of pedigree data". In: *Hum Hered* 21.6 (1971), pp. 523–42.

[17] Dario Gasbarra, Mikko J Sillanpää, and Elja Arjas. "Backward simulation of ancestors of sampled individuals". In: *Theoretical Population Biology* 67.2 (2005), pp. 75–83.

[18] Dario Gasbarra et al. "Estimating genealogies from unlinked marker data: a Bayesian approach". In: *Theoretical population biology* 72.3 (2007), pp. 305–322.

[19] Steven Gazal et al. "High level of inbreeding in final phase of 1000 Genomes Project". In: *Scientific reports* 5 (2015).

[20] Andrew Gelman, Donald B Rubin, et al. "Inference from iterative simulation using multiple sequences". In: *Statistical science* 7.4 (1992), pp. 457–472.

[21] Marta Girdea et al. "PhenoTips: patient phenotyping software for clinical and research use". In: *Human mutation* 34.8 (2013), pp. 1057–1065.

[22] Simon Gravel. "Population genetics models of local ancestry". In: *Genetics* (2012), genetics–112.

[23] J D Hadfield, D S Richardson, and T Burke. "Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework". In: *Mol Ecol* 15.12 (Oct. 2006), pp. 3715–30. DOI: `10.1111/j.1365-294X.2006.03050.x`.

[24] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970).

[25] Dan He et al. "IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data". In: *J. Comput. Biol.* 20.10 (Oct. 2013), pp. 780–791. ISSN: 1066-5277. DOI: `10.1089/cmb.2013.0080`.

[26] Garrett Hellenthal and Matthew Stephens. “msHOT: modifying Hudson’s ms simulator to incorporate crossover and gene conversion hotspots.” In: *Bioinformatics* 23.4 (2007).

[27] Tom Higham et al. “The timing and spatiotemporal patterning of Neanderthal disappearance”. In: *Nature* 512.7514 (2014), p. 306.

[28] W G Hill and B S Weir. “Variation in actual relationship as a consequence of Mendelian sampling and linkage”. In: *Genet Res (Camb)* 93.1 (Feb. 2011), pp. 47–64. DOI: `10.1017/S0016672310000480`.

[29] Chad D Huff et al. “Maximum-likelihood estimation of recent shared ancestry (ERSA)”. In: *Genome research* 21.5 (2011), pp. 768–774.

[30] Albert Jacquard. *The genetic structure of populations*. Vol. v. 5. Biomathematics. Berlin: Springer-Verlag, 1974. ISBN: 0387063293 (New York).

[31] Owen R Jones and Jinliang Wang. “COLONY: a program for parentage and sibship inference from multilocus genotype data”. In: *Molecular ecology resources* 10.3 (2010), pp. 551–555.

[32] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. “Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes”. In: *PLoS Comput Biol* 12.5 (May 2016), e1004842. DOI: `10.1371/journal.pcbi.1004842`.

[33] JFC Kingman. “Exchangeability and the evolution of large populations”. In: (1982).

[34] John FC Kingman. “On the genealogy of large populations”. In: *Journal of Applied Probability* (1982), pp. 27–43.

[35] John Frank Charles Kingman. “The coalescent”. In: *Stochastic processes and their applications* 13.3 (1982), pp. 235–248.

[36] Vecchi MP. Kirkpatrick S Gelatt CD. “Optimization by Simulated Annealing”. In: *Science* 220 (1983), pp. 671–680.

[37] Bonnie Kirkpatrick. “Non-identifiable pedigrees and a bayesian solution”. In: *International Symposium on Bioinformatics Research and Applications*. Springer. 2012, pp. 139–152.

[38] Bonnie Kirkpatrick et al. “Pedigree Reconstruction Using Identity by Descent”. In: *J. Comput. Biol.* 18.11 (Nov. 2011), pp. 1481–1493. ISSN: 1066-5277. DOI: `10.1089/cmb.2011.0156`.

[39] Amy Ko and Rasmus Nielsen. “Composite likelihood method for inferring local pedigrees”. In: *PLoS genetics* 13.8 (2017), e1006963.

[40] E S Lander and P Green. “Construction of multilocus genetic linkage maps in humans”. In: *Proc Natl Acad Sci U S A* 84.8 (Apr. 1987), pp. 2363–7.

[41] Anne-Louise Leutenegger et al. “Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us&amp;quest”. In: *European Journal of Human Genetics* 19.5 (2011), pp. 583–587.

[42] Mason Liang and Rasmus Nielsen. “The lengths of admixture tracts”. In: *Genetics* (2014), genetics–114.

[43] Oren E. Livne et al. “PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population”. In: *PLoS Comput. Biol.* 11.3 (Mar. 2015). ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1004139.

[44] Sarah L Lundregan et al. “Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis”. In: *Molecular ecology* 27.17 (2018), pp. 3498–3514.

[45] Ani Manichaikul et al. “Robust relationship inference in genome-wide association studies”. In: *Bioinformatics* 26.22 (2010), pp. 2867–2873.

[46] Mary Sara McPeek and Lei Sun. “Statistical tests for detection of misspecified relationships by use of genome-screen data”. In: *The American Journal of Human Genetics* 66.3 (2000), pp. 1076–1094.

[47] Philipp W Messer. “SLiM: simulating evolution with selection and linkage”. In: *Genetics* 194.4 (2013), pp. 1037–1039.

[48] Brook G Milligan. “Maximum-likelihood estimation of relatedness”. In: *Genetics* 163.3 (2003), pp. 1153–1167.

[49] Ida Moltke et al. “Uncovering the genetic history of the present-day Greenlandic population”. In: *Am J Hum Genet* 96.1 (Jan. 2015), pp. 54–69. DOI: 10.1016/j.ajhg.2014.11.012.

[50] Jurg Ott, Yoichiro Kamatani, and Mark Lathrop. “Family-based designs for genome-wide association studies”. In: *Nat. Rev. Genet.* 12.7 (July 2011), pp. 465–474. ISSN: 1471-0056. DOI: 10.1038/nrg2989.

[51] Trevor J Pemberton et al. “Inference of unexpected genetic relatedness among individuals in HapMap Phase III”. In: *The American Journal of Human Genetics* 87.4 (2010), pp. 457–464.

[52] Alkes L Price et al. “Sensitive detection of chromosomal segments of distinct ancestry in admixed populations”. In: *PLoS genetics* 5.6 (2009), e1000519.

[53] Kay Prüfer et al. “The complete genome sequence of a Neanderthal from the Altai Mountains”. In: *Nature* 505.7481 (2014), p. 43.

[54] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[55] Monica D Ramstetter et al. “Inferring Identical-by-Descent Sharing of Sample Ancestors Promotes High-Resolution Relative Detection”. In: *The American Journal of Human Genetics* (2018).

[56] Markus Riester, Peter F Stadler, and Konstantin Klemm. “FRANz: reconstruction of wild multi-generation pedigrees”. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2134–9. DOI: 10.1093/bioinformatics/btp064.

[57] Rori V Rohlfs, Stephanie Malia Fullerton, and Bruce S Weir. “Familial identification: population structure and relationship distinguishability”. In: *PLoS Genet* 8.2 (2012), e1002469.

[58] Sriram Sankararaman et al. “The genomic landscape of Neanderthal ancestry in present-day humans”. In: *Nature* 507.7492 (2014), p. 354.

[59] Andaine Seguin-Orlando et al. “Genomic structure in Europeans dating back at least 36,200 years”. In: *Science* 346.6213 (2014), pp. 1113–1118.

[60] Martin Sikora et al. “Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers”. In: *Science* (2017), eaao1807.

[61] B. R. Smith, C. M. Herbinger, and H. R. Merry. “Accurate partition of individuals into full-sib families from genetic data without parental information”. In: *Genetics* 158.3 (July 2001), pp. 1329–1338. ISSN: 0016-6731.

[62] Jeffrey Staples et al. “PADRE: Pedigree-Aware Distant-Relationship Estimation”. In: *The American Journal of Human Genetics* 99.1 (2016), pp. 154–162.

[63] Jeffrey Staples et al. “PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent”. In: *Am. J. Hum. Genet.* 95.5 (Nov. 2014), pp. 553–564. ISSN: 0002-9297. DOI: `10.1016/j.ajhg.2014.10.005`.

[64] Mike Steel and Jotun Hein. “Reconstructing pedigrees: a combinatorial perspective”. In: *Journal of theoretical biology* 240.3 (2006), pp. 360–367.

[65] Lei Sun, Mark Abney, and Mary Sara McPeek. “Detection of misspecified relationships in inbred and outbred pedigrees”. In: *Genetic epidemiology* 21.S1 (2001), S36–S41.

[66] Lei Sun and Apostolos Dimitromanolakis. “PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data”. In: *BMC proceedings*. Vol. 8. 1. BioMed Central. 2014, S23.

[67] M Sun et al. “On the use of dense SNP marker data for the identification of distant relative pairs”. In: *Theor Popul Biol* 107 (Feb. 2016), pp. 14–25. DOI: `10.1016/j.tpb.2015.10.002`.

[68] Bhalchandra D Thatte and Mike Steel. “Reconstructing pedigrees: a stochastic perspective”. In: *J Theor Biol* 251.3 (Apr. 2008), pp. 440–9. DOI: `10.1016/j.jtbi.2007.12.004`.

[69] S. C. Thomas and W. G. Hill. “Estimating quantitative genetic parameters using sibships reconstructed from marker data”. In: *Genetics* 155.4 (Aug. 2000), pp. 1961–1972. ISSN: 0016-6731.

[70] EA Thompson. “The estimation of pairwise relationships”. In: *Annals of human genetics* 39.2 (1975), pp. 173–188.

[71] Timothy Thornton et al. “Estimating kinship in admixed populations”. In: *The American Journal of Human Genetics* 91.1 (2012), pp. 122–138.

[72] Anna A E Vinkhuyzen et al. "Estimation and partition of heritability in human populations using whole-genome analysis methods". In: *Annu Rev Genet* 47 (2013), pp. 75–95. DOI: 10.1146/annurev-genet-111212-133258.

[73] Andrew Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE transactions on Information Theory* 13.2 (1967), pp. 260–269.

[74] Benjamin F Voight and Jonathan K Pritchard. "Confounding from cryptic relatedness in case-control association studies". In: *PLoS Genet* 1.3 (Sept. 2005), e32. DOI: 10.1371/journal.pgen.0010032.

[75] John Wakeley, Léandra King, and Peter R Wilton. "Effects of the population pedigree on genetic signatures of historical demographic events". In: *Proceedings of the National Academy of Sciences* 113.29 (2016), pp. 7994–8001.

[76] John Wakeley et al. "Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent". In: *Genetics* 190.4 (Apr. 2012), pp. 1433–1445. ISSN: 0016-6731. DOI: 10.1534/genetics.111.135574.

[77] Jeffrey D Wall et al. "Higher levels of Neanderthal ancestry in East Asians than in Europeans". In: *Genetics* (2013), genetics–112.

[78] J. L. Wang. "Sibship reconstruction from genetic data with typing errors". In: *Genetics* 166.4 (Apr. 2004), pp. 1963–1979. ISSN: 0016-6731.

[79] J Wang, E Santiago, and Armando Caballero. "Prediction and estimation of effective population size". In: *Heredity* 117.4 (2016), p. 193.

[80] J. Wang and A. W. Santure. "Parentage and Sibship Inference From Multilocus Genotype Data Under Polygamy". In: *Genetics* 181.4 (Apr. 2009), pp. 1579–1594. ISSN: 0016-6731. DOI: 10.1534/genetics.108.100214.

[81] Jinliang Wang. "A new method for estimating effective population sizes from a single sample of multilocus genotypes". In: *Molecular Ecology* 18.10 (2009), pp. 2148–2164.

[82] Jinliang Wang. "Computationally Efficient Sibship and Parentage Assignment from Multilocus Marker Data". In: *Genetics* 191.1 (May 2012), pp. 183–194. ISSN: 0016-6731. DOI: 10.1534/genetics.111.138149.

[83] Jinliang Wang. "Unbiased relatedness estimation in structured populations". In: *Genetics* 187.3 (2011), pp. 887–901.

[84] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. "Genetic relatedness analysis: modern data and new challenges". In: *Nature Reviews Genetics* 7.10 (2006), pp. 771–780.