UNIVERSITY OF CALIFORNIA
SANTA CRUZ

IDENTIFICATION AND MIXTURE DECONVOLUTION OF ANCIENT AND FORENSIC DNA USING POPULATION GENOMIC DATA

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Samuel H. Vohr

December 2016

The Dissertation of Samuel H. Vohr
is approved:

Professor Richard E. Green, Chair

Professor Beth Shapiro

Professor David Haussler

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Identification and Mixture Deconvolution of Ancient and Forensic DNA using Population Genomic Data

by

Samuel H. Vohr

Forensic scientists routinely use DNA for identification and to match samples with individuals. Although standard approaches are effective on a wide variety of samples in various conditions, issues such as low-template DNA samples and mixtures of DNA from multiple individuals pose significant challenges. Extreme examples of these challenges can be found in the field of ancient DNA, where DNA recovered from ancient remains is highly fragmented and marked by patterns of DNA-damage. Additionally, ancient libraries are often characterized by low endogenous DNA content and contaminating DNA from outside sources. As a result, standard forensics approaches, such as amplification of short-tandem repeats, are not effective on ancient samples. Alternatively, ancient DNA is routinely directly sequenced using high-throughput sequencing to survey the molecules that are present within a library. However, the resulting sequences are not easily compared for the purposes of identification, as each data set represents a random and, in some cases, non-overlapping, sample of the genome.

In this dissertation, I present two approaches for interpreting shotgun sequences that address two common issues in forensic and ancient DNA: extremely low nuclear genome coverage and mixtures of sequences from multiple individuals. First, I present an approach to test for a common source individual between extremely low-coverage sequence data sets that makes use

of the vast number of single-nucleotide polymorphisms (SNPs) discovered by surveys of human genetic diversity. As almost no observed SNP positions will be common to both samples, our method uses patterns of linkage disequilibrium as modeled by a panel of haplotypes to determine whether observations made across samples are consistent with originating from a single individual. I demonstrate the power of this approach using coalescent simulations, downsampled high-throughput sequencing data and published ancient DNA data. Second, I present an approach for interpreting mixtures of mitochondrial DNA sequences from multiple individuals. Mixed DNA samples are common in forensics investigations, either from the direct nature of a case (e.g., a sample containing DNA from both a victim and a perpetrator) or from outside contamination. I describe an expectation maximization approach for detecting the mitochondrial haplogroups contributing to a mixture and partitioning fragments by haplogroup to reconstruct the underlying haplotypes. I demonstrate the approach's feasibility, accuracy, and sensitivity on both *in silico* and *in vitro* sequence mixtures. Finally, I present the results of applying our mixture interpretation approach on ancient contact DNA recovered from $\sim$ 700 year old moccasin and cordage samples.

Acknowledgments

First, I would like to thank my dissertation reading committee, David Haussler, Beth Shapiro, and Richard E. Green, for their helpful comments in writing this dissertation as well as their advice and guidance over the course of my graduate work. I would also like to thank Ed and Beth for the opportunity to work with them and their research group on many exciting projects over the years.

The text of this dissertation includes a reprint of previously published material. Chapter 2 and Appendix A were previously published in the journal *BMC Genomics* under the title "A method for positive forensic identification of samples from extremely low-coverage sequence data" [112]. The research was supervised by Beth Shapiro and Richard E. Green and the methodology was conceived by myself, Beth Shapiro and Richard E. Green. Carlos Fernando Buen Abad Najar assisted in performing the experiments described in the text. I would also like to acknowledge and thank the Smithsonian Institution for funding this research.

I would also like to acknowledge and thank my collaborators on the work on mitochondrial DNA mixtures presented in Chapter 3. This chapter is the result of a collaboration with Rachel Gordon, Henry A. Erlich, and Cassandra D. Calloway at Children's Hospital Oakland Research Institute (CHORI). The research was supervised by Richard E. Green, Cassandra D. Calloway, and Henry A. Erlich. The methodology was conceived by myself, Richard E. Green, and Jordan M. Eizenga at University of California, Santa Cruz with critical input from Cassandra D. Calloway, Henry A. Erlich and Rachel Gordon. Rachel Gordon and Cassandra D. Calloway designed and performed the laborartory work and generated the sequencing data

sets used to validate the methodology. This project was funded by a grant from the National Institute of Justice.

I would also like to acknowledge my collaborators on the work presented in Chapter 4 on mixture interpretation for ancient contact DNA. I would like to thank Kristine K. Richter, who performed the original laboratory work and interpretation of amplicon sequencing data sets [85]. I would also like to thank Joshua D. Kapp, who performed additional laboratory work to generate the enriched shotgun libraries and, along with Peter D. Heintzman, contributed to the bioinformatic analyses of the sequencing data. Finally, I would like to thank Beth Shapiro and John W. Ives, who supervised this research.

I am fortunate to have been supported by funding from several organizations during my graduate work, including the Smithsonian Institution and the National Institute of Justice. I am also grateful to have received support from the Chancellor's Dissertation Year Fellowship for my final year at University of California, Santa Cruz.

In closing, I would like to acknowledge the people who may not have directly contributed to the research presented here, but still made my time in Santa Cruz memorable and enjoyable. I would like to thank the PIs, postdocs, grads, undergrads, and technicians who made the Paleogenomics Lab a scientifically exciting and fun place to work everyday. I would also like to thank the graduate students, professors and administrative staff in the Biomolecular Engineering Department for providing an inspiring and encouraging environment in which to study. Finally, I would like to thank my family and friends for all of their support through the course of my graduate studies.

Chapter 1

Introduction

1.1 DNA-based identification

A diploid human genomes contains over 6 billion base pairs and differs from the human reference sequence at 4.1 to 5.0 million variant sites on average [102]. Over $> 99.9\%$ of these variants are single-nucleotide substitutions and short insertions or deletions. For forensic scientists, this variation represents a vast resource of information that can be used to distinguish between individuals for the purposes of identification and attributing samples to individuals. However, the central challenge in forensic identification has not been a lack of underlying information but how to access it quickly, efficiently, and in a way that is reproducible across samples and replicates [53]. To this end, forensic scientists have developed identification methods based on relatively small numbers of highly polymorphic loci, which provide extremely high discriminatory power. In this framework, DNA from a sample is genotyped at specific loci, usually through amplification by PCR, to build a profile, a set of allelic observations for each

locus. The resulting profile can be compared directly with other profiles, against standardized databases, or through statistical analysis for challenging samples.

The most commonly used approach for identification makes use of polymorphic loci known as short tandem repeats (STRs) [53]. These loci contain a variable number (≤ 50) of tandemly repeated short (2–6 base pairs (bp)) sequences. An STR can be easily genotyped by PCR amplification based on the flanking sequence combined with gel or capillary electrophoresis to identify the number of repeats that are present. A profile is generated for a sample by determining the number of repeats present in the two copies of each STR carried by an individual. As these loci harbor many distinct alleles (number of repeats), high discriminating power can be obtain from a relatively small number of STR loci, particularly when STRs are not linked (e.g., 13 for Combined DNA Index System (CODIS) used in the United States [18]).

In addition to STRs, the use of single nucleotide polymorphisms (SNPs) for identification has been explored [53, 19, 55]. As the name implies, these loci contain a single nucleotide position where at least two distinct bases are observed within the population. Unlike short tandem repeat segments, which can contain a variable number of repeats and thus many possible variants, a SNP locus usually represents a binary marker. As a result, less information is obtainable per amplification and more loci are required to gain power comparable to STR-based identification [37]. As SNP loci do not vary in length, SNPs can be genotyped using much shorter target sequences than STRs (~ 50 bp vs. ~ 300 bp). This feature of SNP markers is especially advantageous in degraded samples where STRs are not likely to be found intact [19, 118].

Aside from autosomal markers, the Y chromosome and mitochondrial genome have

proved important targets for forensic identification due to their unique properties. For example, as the Y chromosome is only found in genetic male individuals, Y chromosome STRs and SNPs can be isolated from male-female mixtures of DNA [80]. Mitochondrial DNA (mtDNA) also possesses unique properties that make it a valuable target for identification assays. First, the mitochondrial genome is present in each cell in high-copy number relative to the nuclear genome. As a result, mtDNA is a more tractable target for recovery in low-template or degraded samples. Second, the mitochondrial genome is inherited only from the maternal line and variation between mtDNA copies within an individual is generally very low [76]. Third, mtDNA is highly variable between individuals compared with the nuclear genome on a per-nucleotide basis. In particular, the hyper-variable portions of the control region contain mutational hotspots that can be sequenced to detect single nucleotide and short insertion/deletion variants.

Although standard approaches for identification can be applied to a wide variety of samples in various conditions, some types of samples remain challenging for forensic analysis. Two particularly difficult problems facing forensic scientists are low-template DNA samples and samples that contain DNA from multiple individuals [108, 8]. In low-template or trace samples, DNA is present in such small quantities that informative loci, such as STRs, cannot be readily amplified for genotyping, e.g., the sample contains no molecules that completely span the locus and flanking sequences. As a result, these samples can suffer from allelic dropout, where one or more alleles carried by the source individual are not observed in the profile. In addition to low-template samples, samples collected in an investigation may represent a mixture of DNA from multiple individuals, either due to the nature of case (e.g., a sample containing DNA from a victim and perpetrator [10]) or contamination from outside sources [66]. As crime

scene samples and DNA extracted from them represent limited resources for forensic scientists, interpretation of low-template and mixed DNA samples is critical component of forensic investigations.

1.2 Ancient DNA

Extreme examples of the challenges faced in forensic DNA analysis can be found in the field of ancient DNA, where researchers work to recover and interpret sequences from samples that are hundreds to hundreds of thousands of years old [47, 94]. DNA is often described as "ancient" regardless of its age based on the biochemical condition of the molecules. Ancient DNA is highly fragmented by nature and the majority of molecules recovered from sample are less than 100 bp long. This fragment size is substantially shorter than the length required for most forensics applications ($\sim$ 300 bp) including short-amplicon ($\sim$ 200 bp) miniSTRs specifically designed for degraded samples [69]. As a result, STRs cannot be amplified from many ancient samples, as these sequences are not intact within a single molecule. Ancient DNA is also characterized by patterns of cytosine deamination that preferentially affect the 5' ends of molecules. These damaged bases result in C to T and complimentary G to A base pair misincorporations in read sequences [74, 47, 26], which must be accounted for during sequence analysis.

In addition to fragmentation and base damage, libraries generated from ancient DNA extract often suffer from compositional issues as well. First, the age [3] and the historical surrounding environmental conditions [99] of a sample dictate the extent to which endogenous

DNA is preserved. As a result, ancient DNA work often involves extensive screening to find samples containing preserved endogenous DNA. Second, in most samples, endogenous DNA makes up only a small fraction (often on the order of 1%) of the total DNA present in an extract. The remaining portion represents a mixture of environmental and microbial DNA from organisms that have colonized the sample since its deposition. Finally, ancient DNA, like forensic DNA, is highly susceptible to contamination from outside sources. As many ancient samples contain only small quantities of endogenous DNA, contaminating molecules from a modern source can easily overwhelm the endogenous content or produce a mixture of ancient and modern sequences.

1.3 High-throughput sequencing

The development of PCR [90, 70] as a tool for amplifying specific DNA sequences had profound effects on both DNA-based identification and ancient DNA research. For forensic scientists, PCR offered increased sensitivity over existing methods, which made identification from degraded and minute DNA samples possible, and eventually led to the widespread adoption of PCR-based identification methods [53, 88]. For ancient DNA researchers, the sensitivity of PCR greatly improved the technical feasibility of ancient DNA recovery [74, 47].

Similarly, the development and application of high-throughput sequencing technologies [65] has again transformed ancient DNA research over the past decade. Rather than carrying out PCR experiments to amplify specific sequences, researchers now routinely directly sequence short DNA fragments to survey the molecules present within an ancient sample. A

key advantage to shotgun sequencing of ancient DNA is that the resulting data can be used as a resource to address many scientific questions (e.g., the demographic history of a population or the variants carried within a given gene), while PCR experiments must be designed to gather information for specific questions. Given sufficient endogenous DNA content and sequencing resources, complete, high-coverage genomes can be produced for ancient individuals [67, 81, 59] but significant insights can be gained from even modest genomic coverage [41, 95, 34, 4] (see review by Slatkin and Racimo [98]).

In recent years, there has also been increasing interest in the forensic science community in applying high-throughput sequencing to forensic DNA samples [15]. Much of this research has focused on using high-throughput sequencing to sequence the PCR amplification products currently used in forensics casework. These data can provide new information to increase statistical power for identification, e.g., detecting SNP polymorphisms that occur within known STR segments. However, little work has been done to apply high-throughput shotgun sequencing strategies to forensic identification, for several reasons. First, existing PCR-based approaches are highly sensitive and require less DNA than what would be needed to generate high-coverage shotgun data sets from forensic samples. Second, high-throughput sequencing produces an enormous amount of data, which makes interpreting the results much more resource intensive. Third, the nature of shotgun sequencing leads to concerns surrounding the reproducibility of experiments. As shotgun sequencing produces a random sample of reads from across the genome, two sequencing runs from the same sample may be difficult to compare directly. In addition, technical biases in sequencing technologies and variant calling also complicate comparisons across samples (e.g., [58]). Finally, in many cases, sequencing every

fragment in a sample would be excessive as most forensic genetic questions can be answered by surveying a small number of markers. However, for highly degraded or ancient DNA, shotgun sequencing may be the only way to extract any genetic information from a sample.

1.4 Outline of dissertation

In this dissertation, I present two approaches for assessing genetic identity from high-throughput shotgun sequencing data that address two common problems in forensic and ancient DNA analysis: low nuclear genome sequence coverage and mixtures of sequences from multiple individuals. In Chapter 2, I present a method for determining whether two extremely low-coverage shotgun sequence data sets (~ 0.01-fold coverage) originate from the same individual or different individuals. In these types of samples, standard forensic identification approaches would fail, as coverage of the genome is so sparse that specific loci cannot be reliably recovered and full diploid genotypes cannot be observed. Furthermore, in a comparison of two extremely low-coverage libraries, vanishingly few polymorphic sites will be observed in both samples. To address these issues, I describe an approach that differs from existing identification methods in two important ways. First, our approach makes use of large catalogs of single-nucleotide variation to maximize the amount of information available from each sample. Second, we further reduce the amount of data required by taking advantage of patterns of linkage disequilibrium, as modeled by a reference panel of haplotypes, to indirectly compare observations at pairs of linked SNPs. I demonstrate the feasibility of our approach using coalescent simulations, down-sampled sequence data from modern individuals and ancient DNA from bronze-age humans

from Eurasia [4] compared using panels of haplotypes from the 1000 Genomes Project[101]. Our approach, implemented in the software package `tilde`, enables identification from the nuclear genome in extremely low-coverage samples that would be intractable for other types of analysis.

In addition to low-coverage and low-template DNA samples, mixed samples of DNA from multiple individuals pose challenges for both forensic and ancient DNA applications. In Chapter 3, I describe an approach for interpreting mixtures of mitochondrial DNA sequences to determine the number of contributing haplotypes, estimate their relative proportions within the mixture and assign individual sequence fragments to their most likely source contributor. To accomplish this task, we make use of the inferred mitochondrial phylogeny and haplogroup descriptions available from PhyloTree.org [109] to model the haplotypes contributing to the mixture. I employ the expectation maximization algorithm to co-estimate the overall mixture proportions and source haplogroup for each fragment. After applying heuristic filters to identify the most likely contributing haplogroups and to remove spurious signals, we assign each sequence to the most likely source contributor to build haplogroup specific read sets that can be used to reconstruct the sequence and variants for each contributor. To demonstrate the utility and evaluate the performance of the approach, I applied the method to *in silico* mixtures generated from single-source sequencing data. I found that the approach can reliably identify the haplogroups contributing to a mixed sample and provide accurate mixture proportion estimates. Furthermore, I show that fragments can be accurately assigned to their source contributor to reconstruct the constituent haplotypes. In addition to *in silico* mixtures of two individuals, I demonstrate the utility of the approach on three individual mixtures and *in vitro* mixture sam-

ples. Our approach provides the novel functionality of detecting haplogroups within a mixture directly from mapped high-throughput sequencing data and significantly improves the tractability of mixed sample interpretation.

In Chapter 4, I describe the results of applying our approach for interpreting mixtures of mtDNA sequences to ancient contact DNA extracted from leather and cordage samples. DNA transferred to objects by casual contact has been an area of great interest for forensic scientists, but only a few studies have explored recovering ancient contact DNA from the surfaces of artifacts. Accurate interpretation of DNA mixtures plays an important role in analysis of contact DNA, as the recovered DNA may represent a mixture from multiple individuals handling the object as well as the possibility of modern contamination. We demonstrate that a strategy of whole-mitochondrial sequence enrichment and high-throughput shotgun sequencing can recover ancient contact DNA that can be interpreted by our approach to detect the mitochondrial haplogroups present within a sample.

Finally, in Chapter 5, I discuss some potential applications of these methods in the fields of archaeology, forensic science, and conservation genomics.

Chapter 2

A method for positive forensic identification of samples from extremely low-coverage sequence data

2.1 Background

Recent advances in high-throughput sequencing now allow for small DNA fragments to be extracted and sequenced from ancient or heavily degraded samples [94]. DNA is now routinely extracted from bones, teeth, and hair from individuals that lived hundreds to tens of thousands of years ago and from small and poorly preserved forensic samples. Deep sequencing of well-preserved samples can produce many-fold coverage of the complete nuclear genome of an individual [67, 81, 59], but more often samples yield only small amounts of endogenous DNA, i.e., less than one-fold genome coverage [95, 34, 4]. This is due to several factors. First, the amount of intact DNA varies greatly between samples. Also, in many samples endogenous

DNA makes up only a small fraction of the DNA that can be extracted and sequenced. Deep sequencing and capture techniques [40, 7, 22] can be applied to recover more human DNA from each sample but, ultimately, the total amount of useful human DNA sequence is limited by the number of intact human DNA fragments in the sample. In many cases, no enrichment or deep sequencing strategy will increase the amount of data collected.

A common task in forensics and in the analysis of archeological and historical samples is to determine whether two samples originated from the same individual by comparing DNA extracted from each. The most widely adopted approach in forensics is to compare genotype information at a defined set of multi-allelic short tandem repeat (STR) markers [53, 20]. These genotype data can be generated by PCR amplification using primer pairs flanking the markers followed by capillary electrophoresis or, more recently, from deep sequencing of these ampliconic products [14, 114]. This approach is fully effective only when there are amplifiable template molecules for each of the two alleles of each marker. Thus, it is not uncommon for this approach to fail or generate only a partial DNA profile from minute or poorly preserved samples.

The advent of high-throughput sequencing is enabling new approaches to compare DNA between samples [12, 55]. For example, STR profiles can be genotyped from shotgun data by altering the search strategy used to map shotgun reads to a reference genome [44, 33]. These STR data can then be compared to available data generated using more traditional approaches. More recently, some researchers have begun to use single nucleotide polymorphism (SNP) markers, which are more easily typed by whole genome shotgun sequencing [37, 93]. As of July 2015, the Single Nucleotide Polymorphism Database (dbSNP) contains records for

149,735,377 SNP clusters in the human genome [72]. Despite the vast number of SNPs, SNP-based comparison methods typically employ no more than a few hundred markers, in part to avoid the effects of genetic linkage, as this would render each comparison non-independent during analysis. Nevertheless, this enormous catalog is potentially a powerful resource for sample comparison and identification, as it provides much higher probability that genotype data at overlapping markers can be learned.

Here we describe a new approach for comparing very sparse DNA sequence data from two samples that differs from existing methods in two important ways. First, we draw on large catalogs of SNP markers with the expectation that only a small fraction of these will be observed in a sample. Second, we completely avoid direct comparison of alleles between samples. In comparisons of poorly preserved samples, direct comparison is difficult since few or zero SNP positions will be observed in both samples and full genotypes are not recoverable. In addition, multiple observations of the same SNP position may be the result of incorrectly mapped reads from similar sequences in the genome. To address this, our method does not require genetic observations at the same loci. Instead, we compare pairs of observations at linked markers to assess whether the observations more likely originated from a single individual or from independent individuals. This is made possible by the large and growing catalog of known segregating genetic variation in humans and the known patterns of linkage present in human populations. Using this approach we show that it is possible to reliably determine if two samples originated from the same individual when the data available are random, non-overlapping shotgun sequence data representing less than 1% of the genome in both samples.

2.2 Methods

Two loci are in linkage disequilibrium (LD) if their alleles are not randomly associated [97]. As the name implies, LD can occur because alleles that are physically linked and nearby on chromosomes are often co-inherited. This non-random association implies that observing the allelic state of one locus provides some information about the state of the other. Our approach is based on the idea that observations of alleles made in one sample consistently provide information about the allelic state in another sample via LD if and only if the samples are from the same individual (or from a genetically identical individual, i.e., a monozygotic twin). On the other hand, when comparing data from samples from unrelated individuals, each sample provides no predictive information about alleles in the other sample.

Consider a pair of SNP loci in physical proximity on the same chromosome in linkage disequilibrium (Figure 2.1). Alleles (bases) are observed for each of these two positions from independent shotgun reads that overlap these SNP positions in two independent sequence samples (libraries). We compare the probabilities of observing these two bases under two models. The first model represents the case where the two reads, and thus the two alleles they contain, were drawn from independent chromosomes. Under this model, the probability of the observation is simply the product of the two allele frequencies in the population. The second model represents the case where the two bases were drawn from the same diploid individual. Under this model there is an equal chance that this second read was drawn from the same chromosome as the first read or the other chromosome in that individual. In the case that the read derives from the other chromosome, its allelic state is also not informed by the first read. When drawn

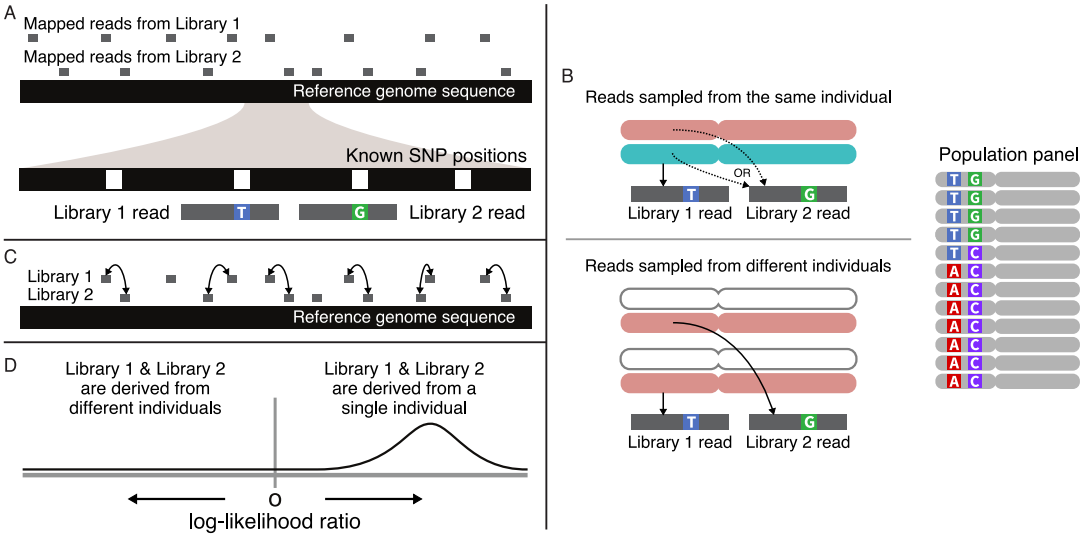

Figure 2.1: Overview of tilde method. (A) Sequence data from two extremely low-coverage shotgun libraries are mapped to a reference sequence. Reads that overlap known single nucleotide polymorphimic (SNP) sites are used to observe a single base for that position. The majority of SNP positions have no observations. Pairs of closely linked SNPs with observed bases are identified. (B) For each SNP pair, we calculate the probability of the observations under two models. The first model represents the case where the two reads orginated from a single diploid individual (top). In this model there is an equal chance that the two reads were drawn from the same chromosome or different chromosomes. This model takes the haplotype frequency as well as allele frequency into account. The second model represents the case where the two reads originated on different chromosomes (bottom). The probability of observing the two bases is the product of the allele frequencies. A reference panel of phased haplotypes is used to model the allele and haplotype frequencies for the population from which the samples were drawn. These probablilies are compared as a log-likelihood ratio. (C) Comparisons are made for any pair of SNPs occuring within a specified distance along a chromosome. D) log-likelihood values are aggregated by sampling pairs from windows across the genome to avoid confounding effects from linkage. This sampling step is repeated in a bootstrapping approach to build an empiracal distribution of the genome-wide log-likelihood ratio. Positive values indicate that the single, diploid individual model is favored while negative values indicate that the two samples are from independent individuals.

from the same chromosome as the first read, however, the probability of observing the alleles seen on the first and second reads is given by the LD between these two alleles, i.e. the haplotype frequency of these two alleles. The two models can be used to evaluate the probability of

any pair of allelic observations from two libraries to determine if it is more likely that the two reads originated from one individual or not.

To estimate these probabilities, we model allele and haplotype frequencies in an approach similar to the one employed by Rodriguez et al. in the identity by descent detection method Parente2 [87]. Allele and haplotype frequencies are explicitly modeled using a reference panel of phased haplotypes. Since this reference panel is used to discover segregating polymorphic positions, their allele frequencies, and the patterns of linkage of these sites, our approach benefits when the population reference panel and samples are drawn from genetically similar populations. In practice, these panels can be constructed from statistically phased genotypes from surveys of human genetic diversity, such as the 1000 Genomes Project [101].

For every SNP pair within a specified distance, the probabilities under the single individual and two individual models are calculated and compared as a log-likelihood ratio (LLR). A positive value indicates that the probability of observing these bases under the single individual model is higher than that under the two individual model. A negative value indicates that the data are more likely to derive from two different individuals. A single comparison of a pair of SNPs offers little information to distinguish between these two models on its own. However, the two models can become more distinguishable by combining observations from many pairs of SNPs across the genome. To achieve this, we aggregate log-likelihood ratios for pairs of SNPs by sampling pairs from sliding windows across each chromosome. We chose a wide window size (500 kb) and pick a suitable pair of alleles within each window. The human genome, at roughly 3 gigabases, contains approximately 6,000 independent windows for these comparisons. This approach allows us to treat each window as an independent assessment of the two

models. Then, we generate an overall LLR by summing the values across all windows where a suitable comparison can be made. Note that in any particular window, it is only necessary that at least one read in each library contains allelic information for at least one SNP position. As described below, this allows analysis of datasets of extremely low overall coverage.

We repeat this process in a bootstrapping approach, selecting random pairs of alleles across each window, to construct an empirical distribution of the genome-wide aggregated log-likelihood ratio. From this distribution, we can assess the mean and variance of the LLR from the underlying data.

As described here, this method is applied to sets of base observations made from shotgun data from two samples. In addition to this, the method can be applied to a set of base observations made from a single sample. When done in this way, the analysis tests if pairs of observations from a single sample are internally consistent. In this case, positive LLR values indicate that the sample data is consistent with a single, diploid individual. Negative LLR values or values not significantly different from zero indicate that the sample is a mixture of fragments from more than one individual or that the reference panel is too distantly related from the sample individual.

2.3 Results

2.3.1 Results from simulations

To examine the feasibility and power of this method, we simulated sets of single allele observations from diploid individuals along with panels of reference haplotypes using

coalescent simulations and tested our method under various demographic scenarios, free from observation errors. We first simulated observations under a simple demographic model of a single population of constant size ($N_e = 10,000$). Since this model lacks population bottlenecks that increase genome-wide LD, we consider this to be a conservative model. In each round of simulation, we generated two diploid individuals by drawing two haplotypes each from the simulations results. Alleles from these individuals were sampled at a rate of 0.02 to produce sets of allelic observations similar to what could be achieved in genome sequencing from libraries that represent 0.02-fold genome coverage. At this level of coverage, traditional methods like STR typing could be expected to fail. The remaining haplotypes, i.e., those not drawn for comparison, were used as the reference panel to model allele and haplotype frequencies. From 100 rounds of simulation, we found our method can consistently distinguish between two single allele observation sets that originate from the same or different individuals (Figure 2.2A) when the reference panel and diploid individuals are drawn from the same population. Within sample comparisons produced results that were nearly identical to across sample comparisons from the same simulated individual. In addition to this, we found that simulated parent-child and sibling-sibling comparisons produce intermediate results centered around zero (see Appendix A, Figure A.2).

In many cases, an ideal reference panel may not be available to describe the variation present in the population from which the samples originated. However, a related population may provide sufficiently similar allele frequencies and patterns of LD to power the approach. To test this, we extended our simulations to model a population split, producing two equally sized populations with no subsequent migration, for various times in the past. We drew the

haplotypes for the diploid individuals from one population and the haplotypes for the reference panel from the other. In this way, genetic drift and independent recombination histories will alter allele frequencies and LD in a way that reduces power in our method. Nevertheless, we found that a reference panel separated by hundreds of generations still provides information to differentiate between samples that come from one individual or two (Figure 2.2B). However, discrimination becomes increasingly difficult as the number of generations since the split of these populations increases.

To test our method on a realistic demographic model for humans, we simulated a demographic history based on parameters inferred from SNP frequency data [43]. We simulated diploid individuals and reference population panels for three subpopulations representing populations in Africa, East Asia and Europe. Importantly, this model features historical bottlenecks that increase and restructure LD in the affected subpopulations. Comparisons made within each subpopulation can differentiate between samples that come from one or two individuals (Figure 2.2C). We also noted that the separation between single individual and two individual comparisons varied between populations due to their distinct demographic histories. The subpopulations that have undergone more recent bottlenecks (Europe and Asia) and thus have higher LD consequently have much clearer separation between models. Strikingly, the subpopulation with the greatest separation (Asia) is the one that has undergone the strongest recent bottleneck. In contrast, the simulated African subpopulation, with the largest N_e and thus the least LD, has the least power to distinguish individuals. Notably, comparisons where the sample data and reference variation were drawn from different populations produced results that were more difficult to interpret (see Appendix A, Figure A.1). These observations highlight

the importance of choosing a suitable reference panel.

2.3.2 Results from shotgun sequencing data

To test our method on extremely low-coverage shotgun sequencing data, we obtained read data from a European male (NA12891) and a European female (NA12892) sequenced as part of Illumina's Platinum Genomes [50]. We sampled 3,200, 32,000, 160,000, 320,000 and 1,600,000 reads without replacement to approximate 0.01%, 0.1%, 0.5%, 1.0%, and 5.0%-fold coverage of the nuclear genome to generate two sets at each coverage level for both individuals. Only the forward reads (101 bp) from each pair were used, to approximate the effect of having short fragments that must be reliably mapped without mated reads. Reads were mapped to the human reference genome (hg19) using BWA [62]. Only biallelic single base substitutions were included in the panels. Base observations were made from the mapped reads using `samtools mpileup` after coverage, map quality and base quality filtering [64]. We constructed reference population panels using the statistically phased haplotypes from the 1000 Genomes Project phase 1 data [101].

We compared each sample to itself and to the other three at the same coverage level using a reference panel of 170 haplotypes from the CEU population (4,444,573 SNPs in total). Pairs of SNPs were separated by 1 to 50 Kb and aggregated LLRs were calculated by sampling pairs in 500 Kb windows (Figure 2.3). At the lowest coverage level, 0.01%, comparisons were not possible as we found very few pairs in close proximity due to the low observation density. At 0.1%$\times$ coverage, the model distributions can be seen separating away from 0, with the comparisons made between the same individuals and comparisons between different individuals

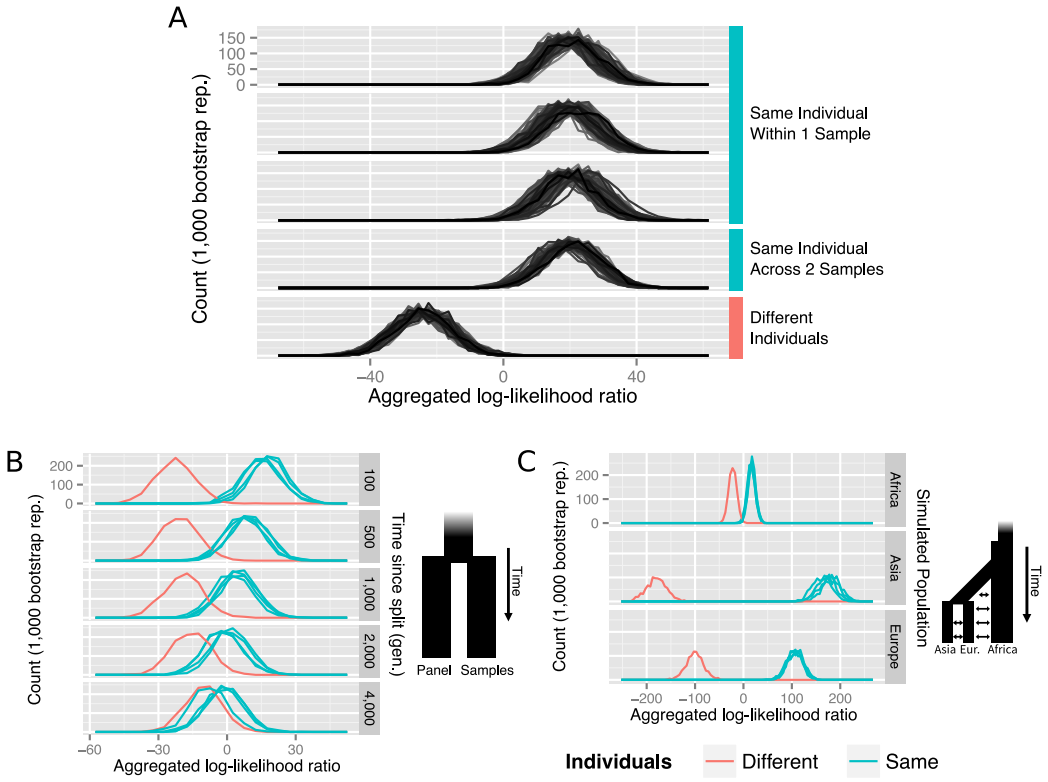

Figure 2.2: Summary of results from coalescent simulations. (A) Results from 100 replicates of comparisons between 3 low-coverage subsamples from 2 simulated diploid individuals with a reference panel of haplotypes drawn from the same population. The first three panels report the results of comparisons made with alleles observed within the same sample. The aggregated log-likelihood ratio values are largely positive, indicating that the observations are consistent with originating from a single diploid individual. Similarly, the fourth panel shows positive values for the comparison made between two independent subsamplings of the same diploid individual. In the last panel, comparisons of allele observations made from two independent individuals produce negative aggregated log-likelihood values, indicating the two samples are from different individuals. (B) A variation of the comparisons made in Panel A where samples from diploid individuals are compared using a panel of reference haplotypes from a related, but diverged population (see diagram). Each panel shows the results of 5 comparisons; 4 made between samples from the same individual (blue) and 1 made between different individuals (red). From these results, a closely related population retains power to differentiate between one or two individuals over many generations. (C) Results from a simulated model of recent human population history (see diagram and Appendix A). In each subpopulation, comparisons between samples from diploid individuals using a panel of haplotypes and differentiate between the 4 comparisons made against the same individual and the 1 comparison made between different individuals. Increased linkage disequilibrium in subpopulations that have undergone recent population bottlenecks (Europe and Asia) increases the power of the method to differentiate between one and two individuals.

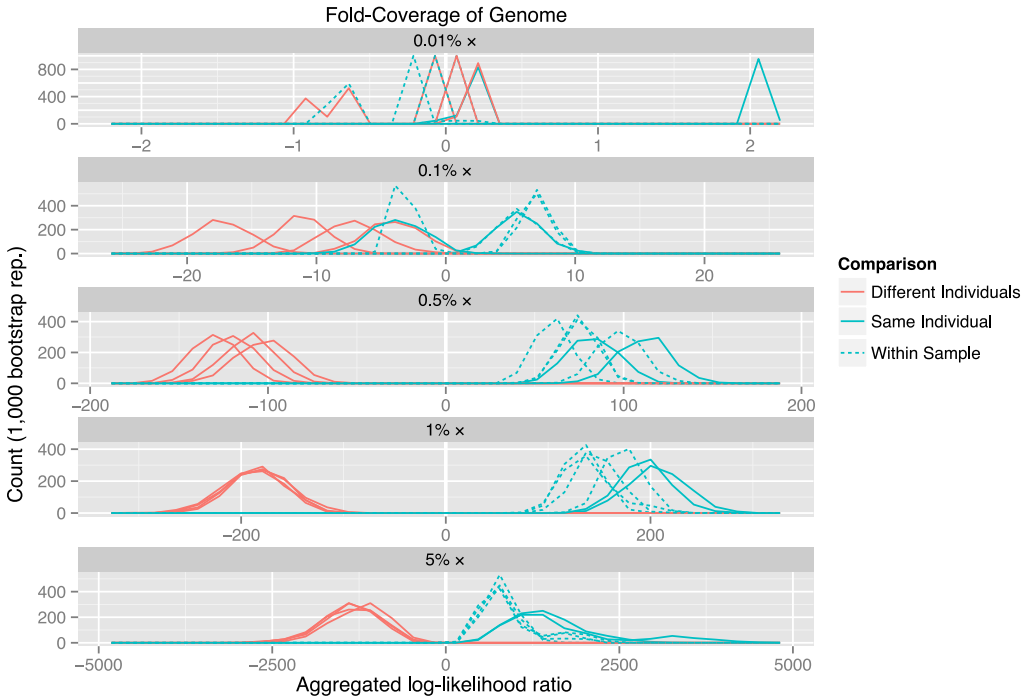

Figure 2.3: Results from comparisons made with single base observations from extremely low-coverage sequencing data. Each panel shows the results of 10 comparisons made between 4 low-coverage subsamples of reads sequenced from 2 human individuals (NA12891 and NA12892 from the CEU pedigree) using a reference panel of haplotypes from the 1000 Genomes CEU subpopulation. Comparisons made from samples from the same individual are shown in blue, with comparisons made within the sample shown as a dashed line. Comparisons made from samples from different individuals are show in red. At the extremely low coverage of 0.01%-fold, very few usable pairs of SNPs can be found. As coverage increases above 0.1%-fold, the two groups of comparisons are easily separable. Comparisons made within a sample have less extreme LLR values as they were made with less read coverage than across-sample comparisons.

trending towards positive and negative respectively. With coverage levels 0.5% and higher, comparisons between the same and different individuals are distinct and easily differentiable. We found similar patterns when using reference panels from other European populations (see Appendix A, Figures A.3–A.6).

2.3.3 Results from ancient DNA sequencing data

Finally, to test our method on genuine ancient DNA sequencing data, we selected 12 samples from 11 Bronze Age Eurasian humans sequenced by Allentoft et al. [4]. After independent DNA extraction and sequencing, the authors discovered that two of these samples (RISE507 and RISE508) were from the same individual through genetic comparison and consultation of the museum records. Sequencing of DNA extracted from these tooth samples produced 0.13 and 0.26-fold coverage of the genome. In addition to these two, we randomly selected 10 other samples sequenced from bones and teeth with similar coverage levels from the data set. We applied our method to all 78 pairwise comparisons between the 12 samples using the same parameters and CEU reference panel of present day humans described in the previous section. We limited our comparison to SNPs that represented transversion mutations to avoid erroneous observations caused by the C to T and G to A substitutions associated with ancient DNA damage [26].

All 12 within-sample comparisons produced positive aggregated LLR distributions (Figure 2.4) indicating that each sample is consistent with originating in a single diploid individual. In contrast, comparisons made between data from different samples are strongly negative, indicating that the two samples most likely originated in different individuals. The exception to this is the comparison between samples RISE507 and RISE508 where the aggregated LLR distribution is strongly positive, correctly indicating that they originated from the same individual. Taken along with mitochondrial result from the original work, this result provides confirmation that the two samples are indeed from the same individual. Variation between results appears to

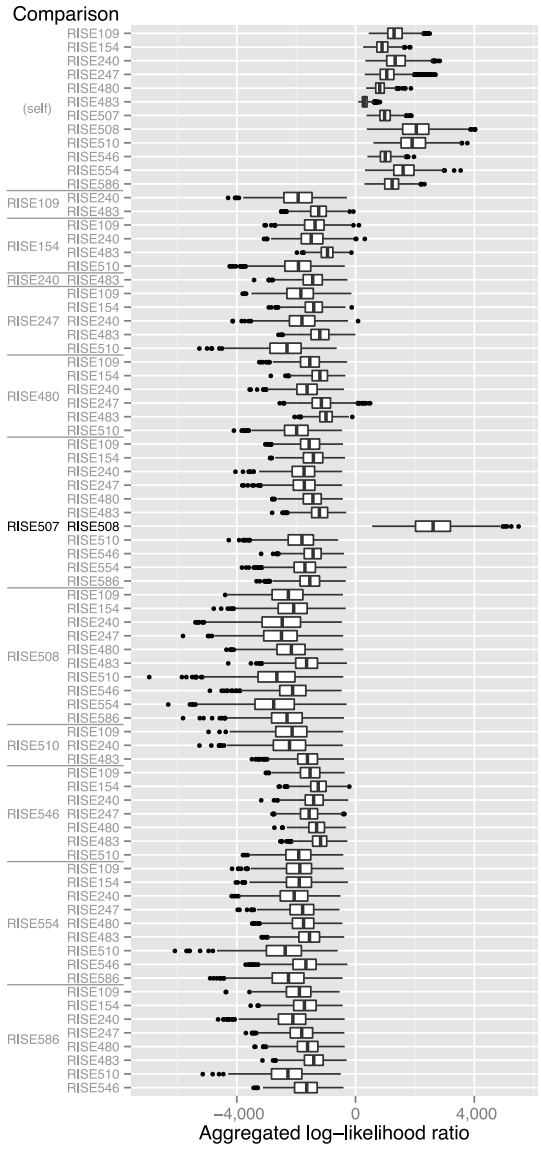

Figure 2.4: Results from comparisons of low-coverage ancient DNA sequencing. Low-coverage (0.1–0.26-fold) sequence data from 12 ancient DNA samples from 11 Bronze Age Eurasian individuals were compared to each other using a panel of haplotypes from the 1000 Genomes CEU population. All comparisons made within a sample produce positive aggregated log-likelihood ratio values indicating that they are consistent with originating in a single diploid individual. Comparisons made between samples produce negative aggregated log-likelihood values, confirming that the two samples are from are from different individuals. The exception to this is the comparison between samples RISE507 and RISE508, the two samples reportedly made from the same individual, indicating that they are consistent with originating in the same individual.

be primarily due to differences in the coverage of the genome in each comparison (see Appendix A, Figure A.10).

2.4 Discussion

We present a method for comparing sparse shotgun read data from two samples to determine whether the two samples originated from the same individual. Our method assumes that observations of alleles from one sample provide information via LD about the allelic state in another sample if both samples derive from the same individual. In addition, because ancestry within an individual and patterns of LD in human populations are local within the genome, we provide quantitative assessment of model fit by considering genomic regions individually and aggregating information across the genome. In this way, our approach provides a robust statistical assessment of whether two samples derive from the same individual or from more than one individual.

Our approach performs well even with as little as 0.01-fold genomic coverage each from the two samples. The quantity of sample necessary to achieve this level of coverage will vary from sample to sample but is attainable in many cases where traditional forensics assays would fail. This amount of genomic data is equivalent to 30 fg of human DNA recovered and sequenced. While the efficiency of DNA recovery and library generation is not perfect [25, 35] and poorly preserved samples are often mixed with environmental DNA [94], our approach should allow useful comparative analysis for samples that were heretofore intractable for forensic analysis.

Central to our approach is the use of a large panel of phased SNP haplotypes from a reference population. For optimal discriminatory power, the reference should be of the same or a closely related population. To select the most appropriate reference panel, one might perform principal components analysis or some other classifier on the sparse data set prior to analysis. However, because haplotype variation across the human genome is often shared among human populations [32, 52], our approach has power even in cases of a poorly chosen reference panel. In many cases, a suitable reference population may not be available, especially for individuals who draw ancestry from multiple distinct populations. In these cases, comparing results from within samples comparisons using different haplotype panels may be useful in identifying the most appropriate panel for comparisons between samples. More work will be needed to understand how admixture affects the results of comparisons to make the method fully applicable to present-day individuals and populations.

As described, our approach is designed to test the relative fit of two models: that the samples originated from the same individual or that the samples are derived from unrelated individuals from the same population. If the two individuals are unrelated and from different populations, then the structure of the test would predict even more strongly that they are two unrelated individuals. One obvious extension of this approach is to test other models of relatedness. For example, our method as described has some power to detect parent-child and sibling-sibling relationships (see Appendix A, Figure A.2). Our approach could be further extended to detect more distant relationships between individuals. Another useful application of this method is to compare pairs of observations made from within a single sample. Using this approach, it is possible to assess the relatedness of the sample to itself, which is functionally

equivalent to a test for contamination or a mixture of samples. Further work will be required to explore the limits of sensitivity for contamination, the impact of complex mixtures in samples, and the limits of sensitivity in real world forensics, historical, and ancient DNA.

Through the method described here, we demonstrate that by using a large collection of SNP markers and patterns of linkage disequilibrium modeled by a panel of haplotypes, sets of non-overlapping, low-coverage sequencing data can be compared to determine if the two samples originated from the same individual.

2.5 Additional methods

2.5.1 Description of algorithm

We begin with a SNP reference panel of haplotypes from a population chosen based on *a priori* knowledge or previous analyses to best represent the population from which the samples originated. For each sample, we examine reads that overlap SNP positions from the panel to identify the base at that position. Bases that do not match one of the two alleles from the reference panel are discarded. Observations for positions where multiple reads map are omitted. The majority of positions will have no observation.

Next, we find all pairs of SNPs between the two samples within a specified distance on the chromosome. A minimal distance is also enforced to ensure that the two base observations are never made from the same fragment in within-library comparisons. For each pair of base observations, denoted as A and B, we calculate the probabilities of that observation under two models. The first represents the probability of observing this combination of bases when the

observations were made from independent chromosomes, i.e. two unrelated individuals. In this case, the two observations are independent and based solely on the frequencies of each allele in the population:

$$P_2(A \wedge B) = f(A)f(B) \tag{2.1}$$

Where $f(A)$ and $f(B)$ are the frequencies of alleles A and B in the population.

In the second model, the observations are made from a diploid individual, where there is an equal chance of the two observations originating from the same chromosome or from different chromosomes.

$$P_1(A \wedge B) = \frac{1}{2}f(AB) + \frac{1}{2}f(A)f(B) \tag{2.2}$$

In the case where both observations are made from the same chromosome, the probability of observing alleles A and B is the frequency that A and B appear on the same chromosome in the population, i.e. the haplotype frequency, $f(AB)$. Otherwise, the probability is the same as independently observing A and B on different chromosomes.

These two models are compared as a log-likelihood ratio, which is calculated as:

$$\gamma(A,B) = \log_2 \frac{P_1(A \wedge B)}{P_2(A \wedge B)} \tag{2.3}$$

Log-likelihood ratios are aggregated across the entire genome through summation of γ for pairs of SNPs in a set S of SNP pairs sampled from windows a set size across the genome.

$$\Lambda(S) = \sum_{(A,B)\in S} \gamma(A,B) \tag{2.4}$$

This step can be repeated as a bootstrapping approach to estimate the empirical distribution of the genome-wide aggregated log-likelihood ratio.

2.5.2 Simulations

We performed coalescent simulations to test our method free of base errors and under various demographic scenarios. We used the coalescent simulator ms [49] to simulate diploid individuals and population reference panels of haplotypes for comparison. For each replicate, we simulated 3,000 independent segments of 500 Kb in size for a total of 1.5 Gb. Segregating sites with minor allele frequencies lower than 10% were removed. Reference panels consisted of 200 haplotypes. For diploid individuals, we simulated base observations from low-coverage sequencing by randomly drawing an allele from segregating sites at a rate of 0.01. This was done separately for each chromosome and sites where both alleles were observed were discarded, resulting in ~ 0.02-fold coverage. This process was repeated to construct multiple observation sets per individual.

The single population model used a constant effective population size of 10,000. The second model, representing the reference population and samples originating in distinct populations, simulated an ancestral population of 10,000 that split 100, 500, 1,000, 2,000, and 4,000 generations ago into two equal sized populations of 10,000 each. The model of recent human history was based off of parameters inferred by Gutenkunst et al. [43] (see Appendix A).

2.5.3 Haplotype reference panels

All human sequence and reference panel data used in this study were downloaded from public sources (accession details listed below). Institutional review and ethical approval were not required for this research.

We constructed reference panels of single nucleotide polymorphisms (SNPs) using

the 1000 Genomes Project Phase 1 data set [101]. We filtered for biallelic SNPs that were polymorphic in the target population (CEU, GBR, etc) with a minimum minor allele count of 10. To avoid errors from mismapped reads, we restricted our panels to sites where all overlapping 35mers are unique across hg19 according to the Duke Uniq 35 track from the Mappability tracks on the UCSC Genome Browser [89].

2.5.4 Modern and ancient human sequence data

We obtained Illumina sequencing data from a European male (NA12891) and a European female (NA12892) sequenced as part of Platinum Genomes by Illumina, Inc [50] from the National Center for Biotechnology Information Sequence Read Archive (accession IDs ERR194160 and ERR194161) [71].

Illumina sequencing data from DNA extracted from 12 samples from 11 Bronze Age Eurasian humans [4] were downloaded from the European Nucleotide Archive (project accession ID PRJEB9021) [29]. We downloaded mapped reads in BAM format for samples RISE109, RISE154, RISE240, RISE247, RISE480, RISE483, RISE507, RISE508, RISE510, RISE546, RISE554, and RISE586.

2.5.5 Availability of supporting data

Software written for this manuscript is available at `http://github.com/svohr/tilde`.

Chapter 3

A fast method for mitochondrial haplotype analysis of sequence data from complex DNA mixtures

3.1 Introduction

Mixtures of DNA from more than one individual are commonly found in forensic samples [66]. For example, a stain found at a crime scene may contain DNA from both a victim and a perpetrator. Mixed samples can also be the result of factors not directly related to a forensics case. A sample from a single individual may contain unrelated environmental DNA or inadvertent contamination from handling prior to laboratory processing. In addition to the complication of mixed samples, many forensic samples represent a limited resource for DNA extraction and the available DNA may exist in small quantities and suffer from degradation. As a result, the interpretation of mixed DNA samples is a critical challenge for forensic analysis.

The interpretation of DNA from mixed samples has been the subject of research since the early days of DNA-based identification [30, 115]. Many distinct approaches have been proposed for different genotyping technologies using various genomic loci. For short-tandem repeat profiles (STRs), methods have been proposed that examine electrophoretogram peaks associated with amplification products of different lengths [39, 23]. While these methods are able to detect if a sample of DNA represents a mixture, it is not possible to reconstruct the individual contributors from a mixed profile, as no linkage information exists between loci. In addition, determining the number of contributors to a mixture can be difficult and is often complicated by allelic dropout [10, 38, 17].

Mitochondrial DNA (mtDNA) represents a promising target locus for interpreting mixed DNA samples. Unlike the nuclear genome, where each cell carries two distinct copies inherited from each parent, the mitochondrial genome is only passed from mother to child. As a result, each distinct haplotype found in a mixed sample can be attributed to a single individual. Each cell contains many identical or nearly identical copies of the mitochondrial genome, making it a more tractable target for recovery and amplification in trace samples [117]. The mitochondrial genome also does not undergo recombination during meiosis. As a result, patterns of variation persist in populations across time and geography. Human genetic variation within mtDNA is extremely well-sampled and catalogued [111, 105, 106]. These data have yielded a comprehensive human mtDNA phylogeny with reconstructed mutation history and a detailed and standardized haplogroup nomenclature [109].

The highly polymorphic mtDNA hypervariable regions (HVRI, HVRII) have been used to examine DNA mixtures through polymerase-chain reaction (PCR) amplification and

Sanger sequencing [113, 68]. Through this approach, it is possible to observe the variants that are present in mixed samples. However, estimating individual contributions is difficult as trace peak areas or heights do not necessarily correlate with mixture proportions [84, 39, 27]. Sanger data from PCR amplification of mixtures alone cannot be used to reconstruct haplotypes from each contributor since the primary signal provides no linkage information.

Recently, studies have focused on applying high-throughput, clonal sequencing technologies, also known as next-generation sequencing, to sequence PCR amplification products from mtDNA hyper-variable regions [48, 75, 57]. Unlike Sanger sequencing, which provides a summary of the bases present at each position, massively parallel clonal sequencing is a survey of individual DNA molecules in the sample and the linkage information they contain. Therefore, it is possible to get a quantitative estimate of the haplotypes that are present in a mixture sample, including novel or low-frequency variants and the haplotype background on which they appear. In addition to the sequencing of PCR products, mitochondrial DNA can be sequenced from shotgun libraries and assembled into a complete mtDNA sequence. This approach avoids many of the biases introduced by multiple runs of PCR and allows for the molecules to be observed in a way that more closely resembles how they appear in the input sample. This approach has been used to sequence and reconstruct the mitochondrial genomes of Neandertals [42], ancient humans [36], and other species [79] from highly-fragmented and damaged DNA [94]. Target-capture enrichment strategies have also been employed with this approach to recover mtDNA from samples more efficiently [16, 40, 7, 22, 100].

Directly sequencing the molecules present in a sample poses its own unique set of challenges for interpretation. When examining PCR amplified sequences, each sequence rep-

resents the same genomic region and therefore every distinct sequence represents a distinct contributing haplotype. Sequences obtained from a shotgun library do not necessarily represent the same regions and cannot always be directly compared against each other to identify the haplotypes that are present. Complete mtDNA sequences can easily be reconstructed when samples contain DNA predominantly or exclusively from a single individual. Mixed samples, however, are not easily interpreted. The human mtDNA locus is over 16 kilobases long while sequence read lengths of high-throughput sequencers are a few hundred base pairs. Thus, individual reads span only a subset of the variants that are present within an individual contributor's mtDNA genome. For this reason, standard assembly strategies cannot reconstruct the complete constituent mtDNA genomes from mixed sequence data.

We present a fast and powerful approach for interpreting shotgun mtDNA data from mixed or unmixed samples, implemented in the software package mixemt. We address two fundamental questions in mixture interpretation. First, how many individuals (contributors) are present in a potentially mixed sample and at what relative proportions? Second, what is the sequence of each contributing haplotype? To answer these questions, we make use of the large catalog of defined mitochondrial haplogroups to identify the distinct haplotypes that contribute to the mixture. We then reconstruct the haplotypes of each contributor by assigning reads to the haplogroup from which it most likely originated. Through this strategy, reads carrying novel variants, which provide the most power to discriminate between individuals, are partitioned by contributor using common, well-described variants. We demonstrate with *in silico* and *in vitro* mixture samples that our method can reliably detect the haplogroups present in mixed samples and estimate their relative mixture proportions. We also show that our method can

reliably assign reads to the correct contributor in mixtures of 2 and 3 individuals, allowing for the reconstruction of the constituent haplotype sequences.

3.2 Materials and methods

In a mixture sample, each sequenced DNA fragment, i.e., a single read or read pair, represents a partial observation of a complete mtDNA haplotype. Our goal is to infer from these data the haplogroups that are present in a mixture, their relative proportions and, ultimately, the sequence for each contributor to the extent it is possible. Exhaustively searching the space of all possible 2, 3, 4, etc. contributor mixtures and proportions to fit the observed data is computationally infeasible. A confounding factor is that individual fragments contain little information on their own. Many fragments overlap conserved regions where few variants occur in the population and often the variants that fall within a fragment are shared by several haplogroups. A robust approach must consider each fragment in the context of all other fragments in a mixture. We propose an approach in which a sample of sequenced fragments is treated as a mixture of all known mtDNA haplogroups. We employ the expectation maximization algorithm to co-estimate mixture proportions and the probabilities of each fragment originating from each haplogroup. We then reduce the pool of possible contributors by applying heuristic filters to identify the most likely haplogroup contributors and remove any spurious haplogroup signals. Finally, the fragments are assigned to one of the identified contributors or none if the haplogroup of origin is ambiguous or cannot be discerned.

Our approach for interpreting mixtures of mtDNA sequences is based on existing

knowledge of the broad patterns of diversity found in the human population. These variants are used to identify haplogroups, estimate mixture proportions, and form the basis of our haplotype assemblies. We use the mtDNA phylogenetic tree from PhyloTree.org, which describes the defining mutations for over 5,000 haplogroups spanning the full range of mtDNA diversity in humans [109]. We use the information from this phylogeny to build our set of candidate contributing haplotypes, each defined by its associated variants as well as variants inherited from its ancestral haplogroups. Although we cannot represent all haplotypes in our set of potential contributors, all mtDNA haplotypes can be described in relation to the haplogroup tree. With this reasoning, we also consider the internal haplogroup nodes to detect haplotypes that are only partially described by the tree. A key feature of our approach is its power to disentangle mixtures with contributors present at near-equal proportions. In these cases, methods that reconstruct haplotypes based on variant frequency in a mixture would fail.

Sequenced fragments are aligned to a reference sequence (e.g., the Reconstructed Sapiens Reference Sequence (RSRS) [11] or revised Cambridge Reference Sequence (rCRS) [6]) to observe variants and haplotypes. Only single-nucleotide substitutions are considered at the haplogroup detection stage, as they represent the vast majority of variants and are easier to observe correctly in single reads. For each fragment, we identify any known variant sites that fall within the fragment and note the bases that are observed. The base observations from a fragment are compared with the expected bases for each haplogroup to calculate its probability of originating from each of the haplogroups. To capture the haplotype information, we obtain this probability by taking the product of the individual probabilities of each observed base under each haplogroup, while allowing for mismatches due to sequencing error. Through this

construction there is a non-zero probability of observing any fragment from any haplogroup.

Using the matrix describing the probabilities for each fragment originating from each haplogroup, we apply expectation maximization (EM) to estimate the relative contribution of each haplogroup to the mixture. Under the EM algorithm, mixture proportion estimates are initialized at random and incrementally improved until convergence. In each round of EM, we perform the expectation step by calculating the conditional probability that each fragment originated from each haplogroup given the current estimation of the mixture proportions. Intuitively, we improve our estimate of the source haplogroup for an individual fragment by considering the mixture proportions estimated from all fragments. In the maximization step, we obtain a new estimate of the mixture proportions by finding the proportions that maximize the likelihood of the conditional probabilities. The expectation and maximization steps are iterated until convergence. The final result is a maximum-likelihood estimate of the haplogroup mixture proportions and the corresponding read/haplogroup conditional probability matrix.

We apply two filtering steps to the results from expectation maximization to reduce the set of potential contributors from all Phylotree haplogroups to only the most likely contributors. First, we remove haplogroups that have no support from individual fragments. We examine each row (fragment) in the conditional probability matrix to identify the column (haplogroup) associated with the highest probability, i.e., the haplogroup that best explains that fragment. Haplogroups with little or no support from individual reads are removed from consideration. Next, we apply a second filter to remove haplogroup signals that are likely driven by private mutations. For each of the haplogroups in the reduced set, we verify that the majority of diagnostic variants unique to that haplogroup, as defined by Phylotree, are observed

within the mixture sample. Haplogroups lacking broad support from multiple variant sites are removed from consideration. Optionally, the EM algorithm can be run again using only the set of high-confidence haplogroups to improve the mixture proportion estimates.

To reconstruct the sequences of contributing haplotypes, we implemented a strategy of assigning aligned fragments to contributors based on the final estimated conditional probabilities from EM. For each read, we examine the conditional probabilities of originating from each of the putative contributing haplogroups. A read is assigned to the contributor with the highest probability if the odds ratio between the first and second highest probabilities exceeds a user-specified threshold, after accounting for any difference in estimated proportions. Finally, we perform an assembly finishing step by iterative calling consensus bases for each contributor, identifying variants unique to a single contributor and assigning reads that could not be assigned based on Phylotree variants alone.

The effectiveness of our haplotype reconstruction strategy depends on the haplotypes in the underlying mixture (Figure 3.1). In a sample with only one identified haplogroup, all fragments can be trivially assigned to the single contributor. In a mixture of two haplogroups, the number of differences that exist between the two haplogroups dictates how much information is available to trace fragments back to their source haplotype. In complex mixtures of more than two haplotypes, fewer variant bases will be unique to a single haplogroup, further reducing the information available for assignment to a specific contributor.

To test our approach, we generated shotgun mitochondrial genome sequence data from single-source and contrived *in vitro* mixture samples using probe capture for mtDNA enrichment and the Illumina MiSeq for massively parallel sequencing. Previously extracted blood-

derived DNA from 18 individuals representing 4 population groups (African-American, US Caucasian, US Hispanic, and Japanese) [84] was quantified, sheared (250 bp average length), made into libraries, indexed, and pooled for probe capture enrichment targeting the entire mitochondrial genome. Single source samples and *in vitro* mixtures were prepared using estimates of mtDNA copy number of total amount of nuclear DNA with a target of 1 million mtDNA copies total (1–2 ng nuclear DNA).

The shotgun mitochondrial genome sequence data from 18 individuals (3 African-American, 4 US Caucasian, 5 US Hispanic, and 6 Japanese) were mapped to the Reconstructed Sapiens Reference Sequence (RSRS) and PCR duplicates were collapsed to avoid any differences in complexity between libraries. We generated *in silico* mixtures by randomly sampling reads without replacement from single-individual samples. This strategy allows us to know the exact proportions for every mixture and the true source contributor for every fragment in the sample. In addition to the *in silico* mixtures, we generated sequence data by sequencing three mixtures of DNA, each prepared *in vitro* from two source individuals. Two of these mixtures were made from the same contributors (US Hispanic H015: Japanese J167) but prepared with different mixture proportions (50/50 and 75/25). A third *in silico* mixture prepared from two other contributors was also sequenced (Caucasian C163: US Hispanic H104).

3.3 Results

3.3.1 Sequences from a single individual

We first applied our method to sets of 10,000 reads sampled from a single individual. For each set of reads, mixemt correctly identified the single contributing haplogroup and estimated the percent contributions between 86–99% (Table 3.1). The initial contribution estimates are less than 100% because the proportion estimates were made considering all haplogroups. As a result, some of the mixture proportion was attributed to haplogroups that share similar patterns in variation. For comparison, we independently aligned the reads from each sample to the Reconstructed Sapiens Reference Sequence, called a consensus sequence, and used MitoTool [31] to determine the haplogroup of each individual. We found that mixemt was able to identify the correct haplogroup for each individual. In the case of sample C035, MitoTool identified H1, H1i, H1an, H1bb, and H1bc as possible haplogroup designations based on Phylotree build 16 of the mtDNA phylogeny. Our program identifies the haplogroup as the subgroup of H1 with no haplogroup identifier (here referred to as H1[2] as the second unnamed subgroup of H1) which is the parent haplogroup of H1i, H1an, and H1bb. Our designation agrees with MitoTools results, as the haplotype of H035 carries the back-mutation (T152C) shared by H1i, H1an, H1bb, and H1bc that differentiates them from their parent haplogroup H1, but none of the derived alleles that characterize these other haplogroups.

We also found that mixemt was able to avoid cases where additional haplotypes may be implicated by private mutations. For example, in the mtDNA haplotype for C017, the variant G7853A appears on the background haplogroup T2b. In Phylotree, this variant is not known

Sample ID	Population	Haplogroup (MitoTool)	Haplogroup (mixemt)	Estimated Contribution (mixemt) (%)
B009	African American	L1b1a3	L1b1a3	99.22
B015	African American	L2e1a	L2e1a	98.27
B037	African American	L1c2b1c	L1c2b1c	99.87
C017	US Caucasian	T2b	T2b	89.86
C035	US Caucasian	H1, H1i, H1an, H1bb, H1bc	H1[2]$*$	98.74
C125	US Caucasian	H1n, H1n1, H1n3, H1n4	H1n[1]$*$	99.34
C163	US Caucasian	K2a6	K2a6	99.63
H006	US Hispanic	B2d	B2d	89.05
H015	US Hispanic	C1b2	C1b2	99.82
H053	US Hispanic	C1b2	C1b2	96.85
H086	US Hispanic	L1b1a7a	L1b1a7a	99.61
H104	US Hispanic	C1b11	C1b11	78.28
J166	Japanese	D4a1b	D4a1b	99.19
J167	Japanese	D4j	D4j	96.81
J171	Japanese	F2a	F2a	87.11
J180	Japanese	M7b1a1a1	M7b1a1a1	99.08
J213	Japanese	D4b2b1	D4b2b1	98.90
J217	Japanese	C1a	C1a	88.64

Table 3.1: Samples used in single-individual and *in silico* mixture analyses. The last two columns contain the results from a mixemt analysis of 3,000 fragments from each sample. Haplogroup designations from MitoTool [31] match the haplogroup identified by our program, mixemt. $*$ H1[2] and H1n[1] are our own notation for haplogroups without names in Phylotree. H1[2] is the unnamed sub-haplogroup of H1 and "parent" of haplogroups H1i, H1an, and H1bb in Phylotree builds 16 and 17. H1n[1] is the unnamed sub-haplogroup of H1n and "parent" of haplogroups H1n1, H1n3 and H1n4.

to occur on the haplotype background of T2b, but it is known to occur in several other haplogroups, including the closely related T1a2. Because the private variant occurs in the coding region where variant density is low, the fragments in the sample that carry this variant are unlikely to overlap any other variant sites that could aid in interpretation. As a result, the EM algorithm takes this observation as evidence for the presence of T1a2 when estimating the mixture proportions. However, the subsequent filtering steps, namely, requiring observations of all diagnostic mutations in a called haplotype, are able to identify and remove this spurious signal.

3.3.2 Mixtures of sequences from two individuals

We generated a panel of 336 *in silico* mixture samples of two individuals by sampling fragments without replacement from pairs of our single-individual sequence data sets. Each mixture sample consists of 3,000 fragments, after PCR duplication removal, representing approximately 50-fold coverage of the mitochondrial reference. We generated six different mixture proportions (50/50, 60/40, 70/30, 80/20, 90/10, 95/5) for each pair of individuals. The equal proportion (50/50) and low-frequency (95/5) mixtures represent the most challenging cases. When mixture proportions are equal, the frequency of a variant provides no information on the contributor from which it originated. The 95/5 mixtures represent the other extreme, where the minor haplotype may be difficult to detect due to random coverage fluctuations along stretches of the reference sequence.

We ran mixemt on each of 336 *in silico* mixture samples of two contributors (Table 3.2). We found that our program was able to detect the correct two contributing haplogroups with no additional haplogroups in 263 mixtures across various coverage levels. This total in-

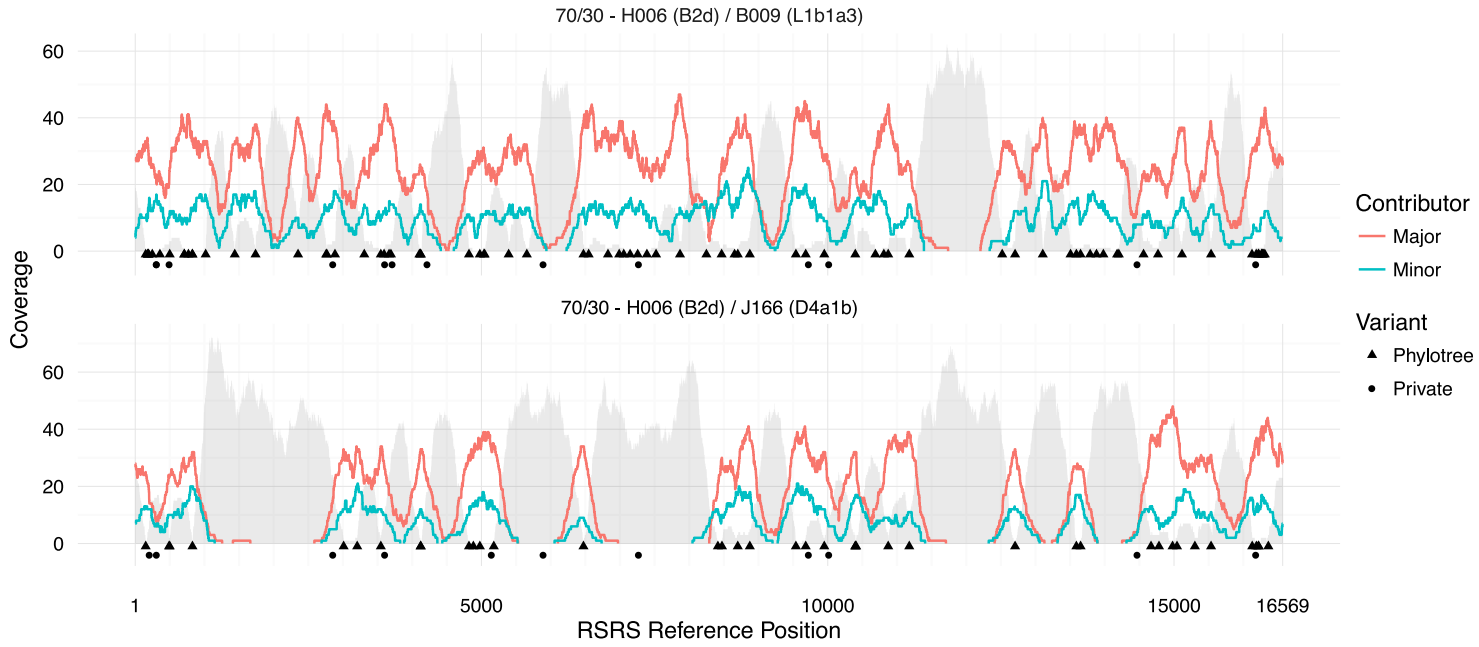

Figure 3.1: Coverage of mtDNA reference sequence from fragments assigned to detected mixture contributors for two mixture samples. Regions surrounding known variant positions from Phylotree (triangles) are recovered using information from assigned reads including private variants (circles). The amount of the underlying haplotypes that can be recovered depends on the number of differences that are known to exist between the two haplogroups and the size of the fragments in the sample. With distantly related haplogroups (top) nearly all of the mtDNA sequence is recoverable, while mixtures of more closely related haplogroups produce more fragments that cannot be assigned to either contributor (gray area). Lines represent coverage for read fragments assigned to the detected major contributor (sample H006, haplogroup B2d) and minor contributor (sample B009, haplogroup L1b1a3 or J166, haplogroup D4a1b), while the gray area represents the coverage for fragments that were unable to be assigned to either contributor. Triangles indicate positions where the two haplogroups differ by a single base substitution according to Phylotree. Circles mark positions where a variant is not known to exist but two distinct bases are observed within the mixture.

Mixture	2 Exact Matches Only	Major Haplogroup Exact Match	Major Haplogroup Close Match	Major Haplogroup Not Detected	Minor Haplogroup Exact Match	Minor Haplogroup Close Match	Minor Haplogroup Not Detected	Additional Haplogroups
50/50	54/56	55	1	0	56	0	0	2
60/40	52/56	56	0	0	54	2	0	3
70/30	54/56	56	0	0	55	1	0	2
80/20	51/56	56	0	0	52	4	0	2
90/10	47/56	56	0	0	49	7	0	3
95/5	5/56	56	0	0	5	2	49	0
Total:	263/336	335	1	0	271	16	49	12

Table 3.2: Results from analysis of *in silico* mixtures of 3,000 fragments from two source individuals.

creases to 275 if we include mixtures where a closely related haplogroup was detected (e.g., F2 was detected rather than the correct F2a). The major haplogroup was identified correctly in all mixture samples, while the minor haplogroup or a closely-related haplogroup was found in 287 samples. The only cases where the minor haplogroup was not detected occurred in the 95/5 mixtures, where the minor haplogroup was detected in only 7 of 56 mixtures. The failure to detect minor contributors at low frequency likely represents a limitation of the data sets rather than the method itself. Since 5% of 3,000 fragments with an average length of 300 bp represent less than 3-fold coverage of the reference sequence, it is likely that observations of the defining mutations of the minor haplogroup have dropped out from the mixture sample entirely. This limitation can be addressed by deeper sequence coverage of the mixture sample.

In 12 of the mixture samples, additional haplogroups were detected alongside those of the contributing haplotypes. Each of these cases arose from the true contributing haplotypes differing from the haplotypes known from Phylotree. For example, the haplotype of sample H006 contains a mutation (A7271G) that is not known to exist on its haplogroup, B2a, but is found elsewhere in the tree. The post-EM filtering steps often remove haplogroups that were

erroneously detected from private mutations. However, these haplogroups can occasionally elude filtering when the minor haplogroup is closely related or present at low frequency in the mixture. 8 of the 12 cases where an additional haplogroup was reported occur in mixtures from the individuals J171 and C035. In these cases, the J171 haplotype carries 2 novel coding-region variants (G8572A and T11152) that are shared with two haplogroups (H1s and H26) that differ from the true haplogroup for C035 (H1[2]) by only 2 and 3 SNP bases. As a result, the EM algorithm has difficulty distinguishing these closely-related haplogroups and more than one may pass the detection requirements. In these cases, false haplogroup signals can be identified by manual inspection of the read coverage of the assigned read sets.

To evaluate our program's effectiveness at estimating mixture proportions, we compared the estimated haplogroup contributions from our set of two individual mixtures with the actual proportion or read fragments used to generate the mixtures. We found that the estimated contributions closely follow the actual fraction of reads used to build each mixture sample (Figure 3.2A). In general, we found that our estimates tend to be slightly lower than the true contribution proportion. As our mixture proportions are estimated by considering all haplogroups, some proportion of the mixture is attributed to similar or extremely low frequency haplogroups. These haplogroups are removed from analysis during the subsequent filtering steps, but the fraction of the mixture attributed to them is not redistributed. The missing proportion in each mixture report provides some information on the uncertainty of the result. Similarly, we found that the mixtures whose estimated proportions are least accurate are often the same cases that generate false assignments in the EM step. For example, in mixtures that include C017, we found that the EM algorithm attributed a small but considerable component

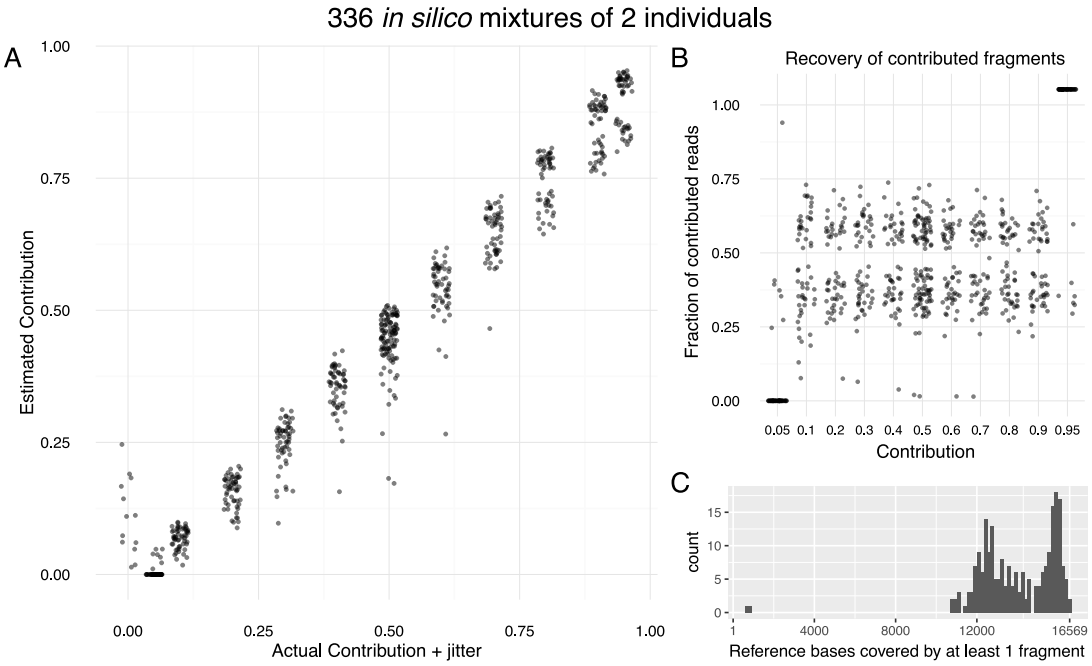

Figure 3.2: Summary of results from analysis of *in silico* mixtures of sequences from two source individuals. (A) Estimated contributions from mixemt compared with the actual contributions used to generate the mixture. Points are randomly shifted along the x-axis to avoid overplotting. (B) Fraction of reads assigned to a detected contributor out of the total number of reads from an individual, compared across contribution levels. The distribution of the fraction of assigned fragments is bimodal due to the number of differences between the two haplogroups in each mixture, i.e., mixtures of deeply diverged haplogroups contain more SNP differences that can be used to assign reads. The fraction of recovered reads remains consistent across contribution levels. (C) Distribution of the number of reference bases covered by at least 1 fragment in the sets of reads recovered for each contributor. Although, fewer than half of the reads contributed from an individual can be identified and assigned, the majority of the mtDNA sequence can be observed through long range information from paired reads (mean insert size between 200–300 base pairs). Samples where the minor contributor was not detected are excluded.

to haplogroup T1a2, at the expense of the true haplogroup T2b. The greatest discrepancies in proportion estimation occur in the mixtures of J171 and C035, where additional haplogroups are falsely detected. As a result, the probability associated with C035 is split across several closely related haplogroups. Even when considering these cases, our method provides accurate contribution estimates that follow the actual proportions in a predictable and interpretable way.

Additionally, contribution estimates can be improved by running the EM algorithm again on only the haplogroups that pass all filters (Additional Figure B.1).

3.3.3 Assigning sequence reads to haplogroups

In addition to detecting haplogroups and estimating their proportions, mixemt also assigns sequence reads to the contributing haplogroup when possible. The completeness of this assignment depends on both the number of differences between the contributing haplotypes and the distribution of fragment lengths of the DNA library. We examined the output read sets in our 2 contributor mixture experiments, partitioned by contributor, to measure the completeness of the assignments. Because our sequence data was generated independently for each individual, we can determine the true origin of each read by examining its identifier. On average, 45.9% of the initial number of contributing sequence reads (22.4–68.7% middle 95% interval) could be assigned to a specific haplotype, after excluding mixtures where only 1 haplotype was identified. Across our comparisons, the observed distribution of percentage of fragments recovered is bimodal (Figure 3.2B). The component with a higher percentage of assigned reads is made up of mixtures that include at least 1 of the deeply diverged L haplogroups.

Although many fragments cannot be assigned to a single contributor, we found that large portions of the contributing haplotypes are recoverable through long fragments (Figure 3.1). Figure 3.2C shows the distribution of the number of reference positions that were covered by at least 1 fragment in the read sets partitioned by contributor. Again, we found that this distribution is bimodal with the mixtures including an L haplogroup yielding more complete haplotype reconstructions. In all but 2 mixtures, our method was able to recover fragments covering over

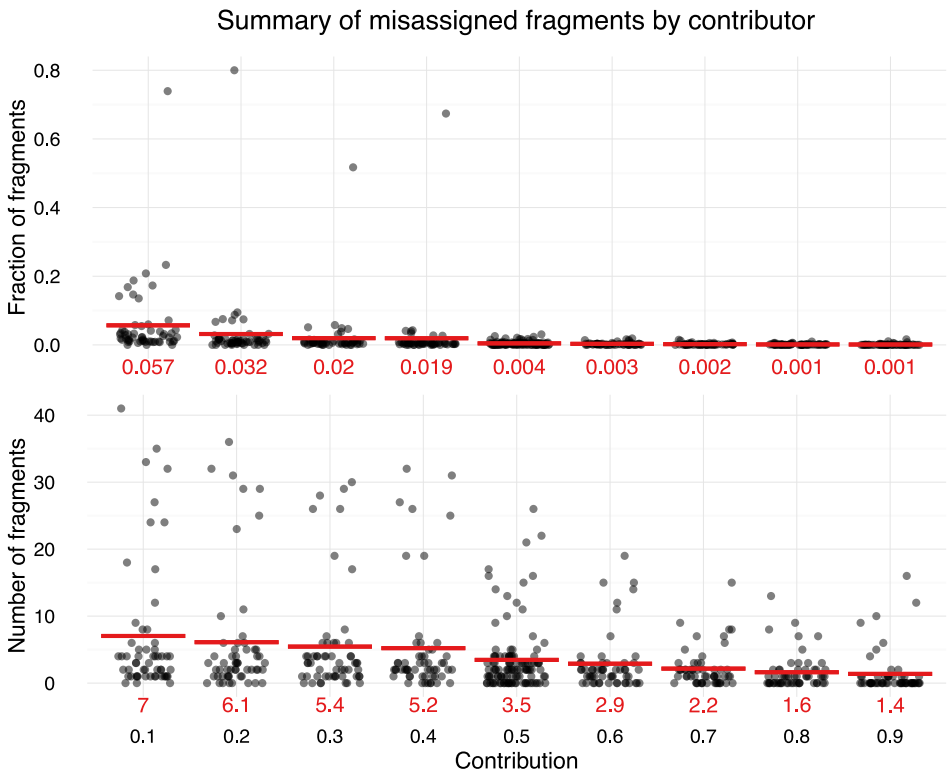

Figure 3.3: Fraction (top) and count (bottom) of fragments that were incorrectly assigned to each detected contributor (i.e., the haplogroup of the true source contributor did not match the detected contributor). The rate and count of incorrectly assigned fragments is low across all contribution levels, although slightly increased with lower contributions. 95/5 mixtures are not shown as the minor contributor was not detected in majority of mixtures.

60% of reference positions for both contributors.

To test the accuracy of our method's assignment of reads to contributors, we compared the sets of reads our program generated for each identified contributor with the original read data from each sample (Figure 3.3 (top)). We found that on average 96.2% of fragments assigned to a contributor were from the source individual that matched the detected haplogroup. In each detected contributor, including additional haplogroups not present in the source individuals, we found an average of 15.5 fragments from a source individual that did not match the haplogroup

assignment. This average drops to 4.3 fragments when we exclude the cases where the minor haplotype was not detected in the 95/5 mixtures. When we compared assignment error rates by contribution level, we found that the minor contributors had slightly elevated rates of fragment misassignment compared with the major contributors (Figure 3.3 (bottom)). However, the increased rate may be the result of the larger population of "incorrect" fragments available for minor contributors compared with major contributors. As the overall coverage level for these samples is $\sim 50\times$, we do not expect the consensus sequence to be affected greatly at these levels of read misassignment, unless all incorrect reads appear in the same region of the reference sequence.

To test the effectiveness of our method at capturing the variants present in a contributor's haplotype and the effect of misassigned fragments on the overall haplotype reconstruction for a contributor, we compared variants called from the single-individual contributions with variants called from sets of reads partitioned by contributor (Additional Figure B.3). After excluding the 95/5 mixtures, where the minor haplogroup was often not detected, we found that on average 87.5% of the variants for each contributor were recovered from our sets of partitioned reads. The fraction of recovered variants for minor contributors (84.8%) was only slightly lower than that of the major contributors (90.1%). The majority of variants that were detected in the single individual sample but not from reads assigned to a contributor were due to lack of coverage in the reconstructed set (76.4%). The remaining missed variants were covered in the reconstructed set but called as a match to the reference. We observed few cases of novel variants appearing from the reconstructed set. Of the total 27,768 variants found in the contributor read sets, only 57 were not found in the single-individual variant calls. The majority of these

Mixture	Number of Fragments	2 Exact Matches Only	Major Haplogroup Exact Match	Major Haplogroup Close Match	Major Haplogroup Not Detected	Minor Haplogroup Exact Match	Minor Haplogroup Close Match	Minor Haplogroup Not Detected	Additional Haplogroups
95/5	10,000	39/50	50	0	0	41	7	2	3
95/5	20,000	41/50	50	0	0	45	4	1	6
95/5	40,000	42/50	50	0	0	48	2	0	7
99/1	10,000	0/50	50	0	0	0	1	49	1
99/1	20,000	19/50	50	0	0	20	5	25	4
99/1	40,000	25/50	50	0	0	31	12	7	14
Total:		166/300	300	0	0	185	31	84	35

Table 3.3: Results from analysis of *in silico* mixtures of two source individuals with a minor contribution of 1% and 5% at 3 coverage levels.

(48) were novel variants. We observed only 9 cases of a variant from one contributor appearing in the variant calls for the other, i.e., wrong, contributor in the mixture.

3.3.4 Sensitive detection of haplogroups representing 1–5% of a mixture

In the initial evaluation of our approach, we found that haplogroups that represent less than 10% of a mixture were occasionally not detected at a modest sequencing depth (3,000 fragments, $\sim$ 50-fold coverage). However, these contributors may be visible with more sequence data. To test the sensitivity of our method on mixtures representing higher reference coverage, we generated 300 additional 2-contributor mixtures representing mixture proportions of 95/5 and 99/1 and at 3 sequencing depths (10,000, 20,000, and 40,000 fragments). For each combination of mixture proportion and sequencing depth, we generated 50 mixtures by randomly drawing pairs of individuals from our panel. The mixtures were analyzed with the same parameters used in the previous experiment.

In mixtures where the minor contributor represented 5% of the total number of sequences, we found that the additional sequence coverage helped resolve the minor contributor (Table 3.3). From 10,000 fragments, the exact minor haplogroup was detected in 41 out of 50

mixtures. The number of correctly identified haplogroups increased with 20,000 and 40,000 fragments. In the 1% mixtures, we found that the minor haplogroup was not resolvable with 10,000 fragments (dropout in 49 out of 50 mixtures). With 20,000 and 40,000 fragments, the number of mixtures where the minor haplogroups at 1% was recovered increased to 20 and 31. However, in both the 1% and 5% mixtures, we found that the addition of more sequence data also led to an increase in the number of additional haplogroups passing the filtering steps. This may be the result of an increased number of base-call errors introduced with the increased sequence coverage. Adjusting the parameters for the contributor filtering steps may alleviate the effects of sequencing error. Although more effort in interpretation may be required, mixemt is able to resolve haplogroups present between 1–5% in a mixture given increased sequence coverage.

3.3.5 Mixtures of sequences from three individuals

Our approach is not limited by the number of contributors to a sample. However, more complex mixtures are potentially more difficult to interpret. Minor haplotypes at low frequency within a mixture may be harder to detect in a complex mixture. Added complexity may also lead to incorrect haplogroup identification and errors in estimating contributions and the addition of more haplotypes increases the possibility of erroneous additional haplogroup calls caused by interference between variants. To demonstrate the utility of our approach on complex mixtures of 3 individuals, we generated 132 mixtures of 10,000 fragments from 12 individuals in our panel (3 from each population group). Each mixture was generated by sampling from a “major”, “minor”, and “trace” contributor at mixtures proportions 75/20/5.

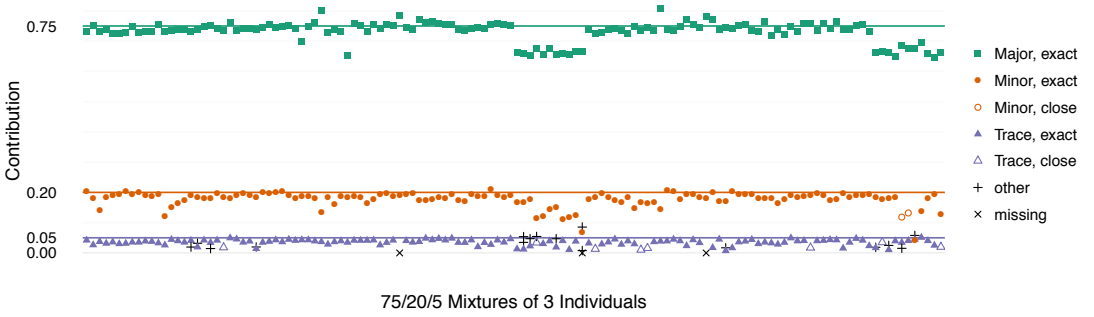

Figure 3.4: Contribution estimates from in silico mixtures of three individuals. Each mixture of 10,000 fragments contains fragments from a major (75%), minor (20%) and trace (5%) contributor. Filled shapes (e.g, •) indicate where mixemt identified the exact haplogroup for the contributor and outlined shapes (e.g., ○) indicate that a closely related haplogroup was identified. Trace contributors that were not detected are indicated with an "×" and additional haplogroups are shown as "+". Mixtures are sorted by the major contributor and samples where the contribution estimate for the major haplogroup drops share the same major contributor (H006 and J217).

Out of the 132 three individual mixtures we generated, mixemt detected the three contributing haplogroups with no additional haplogroups in 108. The haplogroup of the major contributor (75%) was identified correctly in every mixture while the haplogroup of the minor (25%) was found in 130 mixtures. The two mixtures where the exact minor haplogroup was not found involved mixtures of samples C035 and J217. In both cases, interference from a private allele in J217 (G13590A) caused a closely related haplogroup to be identified (H1e1b1 instead of H1[2]). The trace (5%) haplogroup was detected in 129 mixtures and correctly identified in 121. Additional haplotypes passed through filters in only 14 of the mixtures (12 with a single additional haplogroup, 2 mixtures where 2 additional haplogroups were found). We found that the contribution estimates closely followed the true mixture proportions (Figure 3.4). The mixtures with the greatest discrepancy between the actual and estimate contributions were found when samples J217 and H006 represented the major contributor. Both of these haplotypes con-

Contributor	Exact Match	Close Match	Missing	
Major (75%)	132	0	0	-
Minor (20%)	130	2	0	-
Trace (5%)	121	8	3	-
Additional	-	-	-	14

Table 3.4: Results from analysis of *in silico* mixtures of 10,000 fragments from three individuals. Additional haplogroups were erroneously detected in 14 mixtures.

		Haplogroup		Proportions		
Mixture	Contributor	Single source (mitotool)	Found (mixemt)	Actual	Estimated (GeneMarker)	Estimated (mixemt)
H015-J167	H015 (-)	C1b2	C1b2	50	51.79	52.2
	J167 (-)	D4j	D4j	50	47.46	47.8
H015-J167	H015 (Major)	C1b2	C1b2	75	75.96	76.3
	J167 (Minor)	D4j	D4j	25	23.20	23.9
C163-H104	C163 (Major)	K2a6	K2a6	80	79.58	82.2
	H104 (Minor)	C1b11	C1b11	20	18.68	17.8

Table 3.5: Results from mixture proportion analysis of 3 *in vitro* mixtures of two individuals. Estimated mixture proportions made using the detected contributors closely match the expected proportions and independent estimates from GeneMarker.

tain private alleles that occur elsewhere in Phylotree. We also found that reads were generally assigned to the correct haplogroup although the "trace" contributor was affected the most by fragment misassignment (Additional Figure B.4).

3.3.6 *In vitro* mixtures

Finally, to demonstrate the utility of our approach on *in vitro* mixture samples, we analyzed data generated from directly sequencing mixtures of DNA from two contributors. We analyzed sequences from three mixtures, two of which were made up of DNA from the same two contributors (US Hispanic H015: Japanese J167) with two different mixture proportions (50/50 and 75/25). We found that mixemt detected the correct constituent haplogroups for all

Mixture	Contributor	Mixture Prop.	Variants					
			Single Source	Found	Dropped	Missed	Novel	Bled
H015-J167	H015 (-)	50	54	48	5	1	2	0
	J167 (-)	50	47	41	6	0	1	0
H015-J167	H015 (Major)	75	54	49	5	0	0	0
	J167 (Minor)	25	47	42	5	0	0	0
C163-H104	C163 (Major)	82	52	50	2	0	1	0
	H104 (Minor)	17	56	50	5	1	4	0

Table 3.6: Results from analysis of variant calls made from haplogroups separated by mixemt from 3 *in vitro* mixtures of two individuals. Variants detected from the sequences partition by contributor were compared with variants called from data generated from a single source. On average, 90% of variants were recovered correctly. Missing variants were due to insufficient ("Missed") or no sequence coverage ("Dropped") in the sequences partitioned by contributor. Few novel variants (i.e., not found in the single source data) were found in the reconstructed sets and we observed no cases of a variant from one contributor appearing in the variant calls for the other contributor ("Bled").

three mixtures with no additional haplogroups found. Next, we compared the estimated mixture proportions, generated from a second pass of the EM algorithm made on only the detected haplogroups, with the expected proportions. We found that our estimates closely followed the expected contributions in all mixtures as well as independent mixture proportions estimated by GeneMarker (Table 3.5).

To evaluate the effectiveness of our program to reconstruct haplotypes from mixtures, we generated variant calls for each set of reads partitioned by haplogroup. For comparison, we generated variant calls for each individual using reads sequenced from a single source sample (Table 3.6). Unlike in the comparison made for *in silico* mixtures, the variant calls made for the single source and reconstructed sets are not made from the same sequences and may vary between mixture and control. We found that on average 90% (280 of 310 total) of the variants identified in the single source sample were recovered from the haplogroup read sets. In the cases where a variant was not detected, 18 were missing because those positions were not covered by

any assigned reads in the partitioned set and 2 variant positions were observed but not called due to low sequence coverage. Eight novel variants were detected in the reconstructed sets that were not observed in single source variant calls. Upon closer inspection, the novel variants all represent indels either surrounding the inserted reference 'N' positions (523, 524 and 3107) or in control region homopolymer runs ($\sim$ 285 and 310). In all three *in vitro* mixtures, we found no cases where a variant from one contributor appearing in the variant calls for the other.

3.4 Discussion

We describe an approach for deconvoluting mixtures of shotgun mitochondrial sequence data that makes use of the large knowledge base of mitochondrial diversity. We have demonstrated that our method can accurately determine the haplogroups that are present in a sample and estimate the proportions of the mixture from high-throughput sequencing data. Furthermore, we have demonstrated that our method can separate fragments in a mixed sample by their likely originating contributor and reconstruct the constituent haplotypes of a mixture. Our approach has power in difficult cases where contributors are present in equal proportions and where the minor contributor makes up 5% of the total sequences.

Our program is the first, to our knowledge, to identify the haplogroups present in a mixed sample directly from high-throughput sequence data mapped to a reference sequence. Many tools have been developed to determine haplogroups based on Phylotree and other mtDNA databases. Among these are MitoTool [31], HaploGrep [116], HaploFind [110], EMMA [86], Phy-Mer [73], and MToolBox [21]. Each operates on different types of input data with dif-

ferent strategies for identifying haplogroups. EMMA and HaploGrep take variants as input, HAPLOFIND works on near complete mtDNA sequences, and MitoTool accepts both variants and sequences. Phy-Mer and MToolBox both operate on high-throughput sequencing reads. With the exception of the unique, alignment-free, K-mer based approach implemented by Phy-Mer, all of these tools require significant processing of high-throughput sequencing reads to generate consensus sequences or variants needed for haplogroup identification. This makes detection of minor contributors difficult or even impossible as low-frequency variants are discarded prior to haplogroup analysis.

In addition to identifying the haplogroups contributing to a mixed sample, our approach also partitions the input reads by the most likely contributor. Similarly, in the field of ancient DNA, methods for separating endogenous DNA in a contaminated sample have been developed and applied [96, 82]. These methods rely on the characteristic fragmentation and base pair misincorporations of ancient DNA to separate ancient fragments from contaminating modern sequence. Our approach operates in a similar way, but on an orthogonal feature of the data. Under our approach, distinct contributors can be extracted from a sample where patterns of DNA damage are not found or in cases where damage patterns appear on fragments from multiple contributors (e.g. an environmental DNA sample).

The power of our method to recover full haplotypes of contributors depends on two aspects of the input data: the number of differences that exist between the contributing haplotypes and the distribution of fragment lengths in the DNA library. As our method relies on common variants to assign fragments to contributors, mixtures of similar haplotypes inherently contain less information to determine the haplogroup of origin for each fragment. In complex

mixtures of 3 or more individuals, even less information is available as variants are often shared by more than one contributor. If a fragment can be assigned to a contributor, the average fragment length of a sequenced library determines how many base pairs can be observed upstream and downstream of an informative position. In highly fragmented samples where the fragment lengths rarely exceed 100 bp, such as those from ancient sources, it is unlikely that the complete haplotype could be recovered. Long read data is advantageous in our approach for both identifying haplogroups and assembling haplotypes. However, our results demonstrate that the majority of reference positions can be observed with an average insert size of just 200–300 bp.

Although our approach is able to identify the haplogroups present in a mixture, occasionally other haplogroups are identified by our software in addition to the true contributors. These haplogroup signals are often driven by private mutations that occur at known variant positions. While the heuristic filters we employ are effective at removing most erroneous signals, a filtering scheme that works correctly for every mixture case is unrealistic. Instead, improvements in accuracy may come through deeper understanding of haplotype variation and the mitochondrial phylogeny. Our implementation assigns less weight to sites where multiple mutations occur in Phylotree, but more comprehensive estimates of site-specific instability from additional mtDNA databases, such as those used in EMMA [86], may also help to improve accuracy.

It has been noted that the process for determining the haplogroup of a single mtDNA sequence cannot be completely automated with accuracy [9, 86]. These challenges are multiplied when considering a mixed sample of short mtDNA sequences. We have demonstrated that our approach can yield accurate haplogroup assignments and mixture proportion estimates, as well as accurate assignment of fragments to identified contributors. Furthermore, our ap-

proach and algorithm work in a transparent way that allows for results to be easily interpreted. Through this combination of automation and human intervention, we believe that the approach implemented in mixemt can greatly improve the tractability of mixed sample interpretation.

3.5 Additional methods

3.5.1 Description of algorithm

We employ the Expectation Maximization (EM) algorithm to estimate the haplogroup mixture proportions in samples of DNA fragments and to estimate the conditional probabilities of each fragment originating from each haplogroup (see primer by Do and Batzoglou [28]). First, we define $\delta_{i,v,g}$, a matrix of indicator variables for whether the observation of variant $v \in \{A,C,G,T\}$ at reference position i is consistent with haplotype g, which we set to represent the probability of observing (i,v) in haplotype g. For each sequenced fragment j, we define $\{V_j\}$ as the set of variant tuples (i,v) observed from read j. From these, we obtain the probability of observing the variants in fragment j from haplotype g:

$$\tilde{\delta}_{j,g} = \prod_{i,v \in V_j} \delta_{i,v,g} \tag{3.1}$$

Given a set of fragments N and a set of haplotypes G, we estimate the mixture proportions $\{\theta_g\}$ by the following algorithm. First, $\{\theta_g\}$ is initialized randomly by drawing from a Dirichlet distribution. In the expectation step, we obtain the conditional probability of observing the variants in fragment $j \in N$ in haplotype $g \in G$ given mixture proportions $\{\theta_g\}$:

$$z_{j \in N, g \in G} = \frac{\theta_g \tilde{\delta}_{j,g}}{\sum_{g'=1}^{G} \theta_{g'} \tilde{\delta}_{j,g'}} \tag{3.2}$$

We perform the maximization step by obtaining a new estimate of $\{\theta_g\}$ from z:

$$\theta_{g \in G} = \frac{\sum_{j=1}^{N} z_{j,g}}{\sum_{g'=1}^{G} \sum_{j=1}^{N} z_{j,g'}} \tag{3.3}$$

The expectation and maximization steps are iterated until convergence. The mixture proportions and the conditional probabilities that achieved the highest likelihood are used to identify and assign fragments to contributing haplotypes.

To reduce the number of rows in the matrices, we collapse fragments that represent the same sub-haplotype and keep track of the number of times each sub-haplotype has been observed.

3.5.2 Identifying contributors

We employ two heuristic filters to reduce the number of haplotypes under consideration (Figure 3.5). First, we find the haplotypes associated with the highest conditional probability for each fragment and remove haplotypes that represent the highest probability for fewer than n fragments (10 by default). This step removes candidate haplotypes that differ from the true contributing haplogroups by only a few variants. For a haplotype to be considered as a true contributor, there must be fragments that are best explained by the presence of this haplotype.

We apply a second filter based on the diagnostic variants described by Phylotree (Figure 3.5, right panel). As described above, our algorithm does not require fragments from a haplotype to be evenly distributed across the reference sequence. As a result, a variant present in a contributor's haplotype that is unknown to that haplogroup background may be taken as evidence for the presence of another haplotype. To remove these cases, we examine the counts

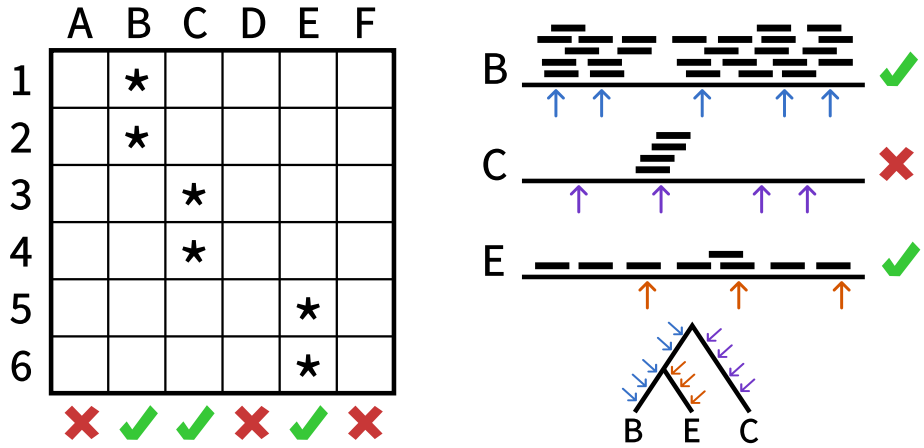

Figure 3.5: Schematic of filtering steps to identify likely contributors. For a haplogroup to be considered a contributor, we require there to be read fragments that are best explained by the presence of that haplogroup (left). Using the conditional probability matrix produced by the expectation maximization algorithm, we find the haplogroup (column) associated with the highest probability for every read (row) (marked with a $\star$) and remove haplogroups with little or no read support from consideration. Next, haplogroups are filtered based on the diagnostic SNP positions from Phylotree (right). Haplogroups signals that are driven by private mutations are removed, while haplogroups with broad support across the majority of unique diagnostic positions are reported and passed on to the assembly steps.

of base observations for unique diagnostic markers for each haplogroup. In descending order of estimated proportion, we identify the defining variants for the contributing haplogroup relative to the Reconstructed Sapiens Reference Sequence from Phylotree. If the fraction of observed diagnostic variants is lower than a specified cutoff (50% by default), the haplogroup is removed from consideration. Otherwise, the haplogroup is retained and the observed variant bases are removed from consideration in subsequent contributors. This is repeated for each of the remaining haplogroups.

3.5.3 Assigning reads to contributors

Fragments are assigned to the remaining contributors using the conditional probabilities obtained from the EM algorithm. The conditional probabilities are normalized by the final mixture proportions and the top two probabilities are compared as a odds ratio. If the ratio exceeds a user-specified cutoff, the read is assigned to the contributor associated with the highest probability. Otherwise, the read is reported as unassigned.

3.5.4 Implementation

Our method is implemented in the software package mixemt in the scripting language Python. It makes use of pysam (`https://github.com/pysam-developers/pysam`), a Python interface for HTSlib [64], to read and write BAM formatted files and NumPy [107] for vector and matrix operations. The source is available online at `https://github.com/svohr/mixemt`.

3.5.5 Laboratory processing

To test our approach, we generated shotgun mitochondrial genome sequence data from single source and contrived mixture samples using probe capture for mtDNA enrichment and the Illumina MiSeq for massively parallel sequencing. The mtDNA copy number and nuclear DNA amount was estimated using a duplex real-time qPCR assay [104] from previously extracted blood-derived DNA from 18 individuals representing 4 population groups (African-American, US Caucasian, US Hispanic, and Japanese) [84]. These estimates were used to normalize the target input DNA amounts of the single source samples (1–2 ng nuclear DNA

and 500,000–2 million mtDNA copies) and to prepare the two-person contrived mixtures by varying the proportions based on mitochondrial copy number with a target of 1 million mtDNA copies total (1–2 ng nuclear DNA). DNA libraries were prepared following the KAPA Hyper Prep Kit TDS protocol (Roche, Indianapolis, Indiana) following the manufactures guidelines for the Illumina platform. All samples were sheared to an average of 250 bp following the Covaris M220 manufacturer's recommended protocol (50 W, 20% duty factor, 200 cycles per burst, Treatment Time (s) 375, Temperature 20° C, 50 µL). After each DNA sample was fragmented a unique dual matched index adapter is ligated to the sheared DNA sample. 24 sets of IDT's HT Dual Matched Index adapters (IDT, San Diego, CA, USA) were used in order to be able to pool samples prior to probe capture. Dual matched indexes ensure no sample mix-up due to possible template switching. The dual matched adapter concentration varied based on DNA amount (300 nM–3 µM). The PCR cycle number (15–18 cycles) was varied depending on the DNA amount. The Agencourt AMPure XP reagent (Beckman Coulter, Brea, CA, USA) was used for small fragment removal following the ratios recommended in Nimblegen SeqCap EZ Library SR protocol (Roche, Indianapolis, IN, USA) and was repeated for a subset of samples which showed incomplete removal of dimer artifacts after the initial size selection step.

Up to 24 dual-indexed samples were normalized and pooled for probe capture using a custom Roche SeqCap EZ Library Developer Library kit (Roche, Indianapolis, IN, USA) probe set targeting the entire mitochondrial genome with high tiling redundancy developed by Calloway et al. (unpublished data). The NimbleGen SeqCap EZ Library SR protocol was followed to enrich for mtDNA and hybridization was carried out overnight (16–18 h). Small fragment removal was carried out two times using the Agencourt AMPure XP reagent at the

recommended ratio. The amount, quality, and size distribution of the capture product pool was determined using the Bioanlayzer DNA 7500 kit (Agilent Technologies, Santa Clara, CA, USA). A dual size selection was carried out following the Hyper Prep Kit TDS protocol for the single source population sample capture product pool to remove large fragments as well as small fragments. The probe capture product pool was quantified using the Quant-iT^{TM} PicoGreen® dsDNA Reagent (ThermoFisher Scientific, Waltham, MA, USA) and diluted to 12.5 pM and sequenced using the MiSeq Reagent PE Kit V2 500 cycles on the Illlumina Miseq (Illumina, San Diego, CA, USA). Sequencing data was analyzed using Nextgene, GeneMarker Softwares (Softgenetics, State College, PA, USA) and mixemt.

3.5.6 Sequence data processing

Reads sequenced from each library were processed with SeqPrep (`https://github.com/jstjohn/SeqPrep`) to remove adapter sequence and merge overlapping reads. Reads were mapped to the Reconstructed Sapiens Reference Sequence (RSRS) [109] using BWA (`bwa aln` with default parameters) [63]. Mapping was done with a linear reference sequence and circularized reference to capture fragments overlapping both the start and end of the reference sequence. Custom scripts were used to merge alignments to both reference sequences and to correct alignments that spanned the ends of the sequence (available from `https://github.com/svohr/circ_aln`)

Variants relative to the RSRS reference sequenced were called using the `mpileup` utility in SAMtools and the `call` function of BCFtools [64]. All variants were called as haploid for both single-individual samples and contributor sets from mixtures.

3.5.7 Haplogroup assignments

To determine haplogroup assignments for each individual independently of mixemt, we generated complete mtDNA sequence assemblies from mapped reads using `mpileup` in SAMtools [64] and a custom utility to determine the majority base for each reference position (`pu2fa`, available from `https://github.com/Paleogenomics/Chrom-Compare`). Haplogroup assignments were made by running MitoTool [31] on the complete mtDNA sequence for each individual using Build 16 of the Phylotree mtDNA phylogeny.

Chapter 4

Detecting mitochondrial haplogroups in ancient contact DNA

4.1 Introduction

DNA from an individual can easily be transferred to an object through casual contact. Forensic scientist have developed methods to collect and amplify DNA from objects that have been in direct contact with skin or bodily fluids [108]. Contact DNA presents a unique set of challenges compared with direct sources of DNA. First, contact DNA is present in small quantities. Second, contact DNA is especially susceptible to contaminating DNA from outside sources such as handling during collection or laboratory preparation. Third, the target DNA itself may represent a mixture from multiple individuals in contact with the object. As a result, the interpretation of DNA mixtures is a fundamental necessity for the correct analysis of contact DNA.

Although contact DNA has been used extensively by forensic scientists, very few archaeological studies have examined DNA transferred to artifacts [46, 60, 85], despite the fact that ancient DNA can be routinely extracted from bones, teeth, and hair [94]. Ancient contact DNA is susceptible to the same confounding factors as modern contact DNA with additional complications unique to ancient samples. First, the preservation of ancient DNA is largely dependent on the surrounding environment, with cooler and dryer conditions being more amenable to DNA preservation [99, 61]. As contact DNA occurs on the surface of an object, it is not protected by outer layers and may be more affected by the surrounding temperature and humidity [2]. In addition, ancient DNA is highly fragmented, with the majority of ancient molecules being shorter than 100 bp. Ancient DNA is also marked by damage due to cytosine deamination near the 5' ends of molecules. These damaged bases manifest as C to T and complementary G to A transitions in sequenced reads, which may confound downstream analysis [26].

As discussed in Chapter 3, the mitochondrial genome is a promising target for the analysis of DNA mixtures, including contact DNA. Compared with the nuclear genome, the mitochondrial genome is highly variable per base pair and present in high-copy number within each cell. These properties allow for highly informative mtDNA sequences to be more easily recovered from the surface of an artifact. In addition, mitochondrial haplotype variation has been studied in depth to characterize human matrilineal population history, including migrations and population expansions and bottlenecks [105, 106]. Unlike forensic identification, where power is gained through the detection of private alleles and precise haplotype reconstruction, significant insights can be gained from archaeological samples by assigning recovered sequences to haplogroups based on common alleles.

One strategy that has been used to identify haplogroups from ancient contact DNA is to amplify specific diagnostic regions of the mitochondrial genome and sequence the products to determine the bases present at known variant sites [60, 85]. Although this approach has proved effective, it is not without its limitations. First, the number of diagnostic region amplifications that can be performed limits the number of haplogroups that can be screened. Therefore, the design of an experiment may sacrifice power to identify haplogroups across the entire mitochondrial tree in order to identify more specific haplogroups on one branch, or vice versa. Second, the use of PCR limits amplification only to molecules that span the entire length of the target region. As ancient DNA is highly fragmented by nature, this excludes a large population of potentially informative molecules and biases the results towards longer fragments that are more likely the result of modern sequence contamination.

We propose an alternative approach to identifying haplogroups from ancient contact DNA samples that makes use of high-throughput shotgun sequencing to generate an accurate survey of the molecules present on an artifact. In addition, we employ whole mitochondrial sequence enrichment to recover informative sequences more efficiently [40, 100]. To identify haplogroups that are present in the recovered DNA, we employ the software tool mixemt, a program developed to identify the haplogroups present in a mixture sample and estimate their relative abundances (See Chapter 3). To demonstrate our approach, we generated and analyzed sequence data from DNA extracted from moccasins, leather, and cordage material that was previously assessed for contact DNA using high-throughput sequencing of PCR amplification products [85]. We performed both automated haplogroup mixture detection and inspection of diagnostic positions to verify the results. In addition, we examined sequences for evidence of

haplogroup A2a, a distinct haplotype that has been inferred to have undergone a range expansion from Alaska to the American southwest in the past 1–2 thousand years [1].

4.2 Materials and methods

4.2.1 Moccasins, leather, and cordage samples

Moccasins, Leather and Cordage material were collected from a cave on the north shore of the Great Salt Lake. The site was excavated originally in 1930s and again in 2010 and 2011. Original excavations revealed signs of occupation around 700 years ago [13, 51]. 38 samples were selected for ancient DNA analysis, and 9 of these samples were selected for whole-mitochondrial sequence enrichment. The 9 samples consisted of 7 moccasins, 1 cordage sample, and 1 leather sample that was possibly a mitten or the lining of a mitten. The moccasins and mitten all showed signs of wear including holes and were therefore excellent candidates for containing DNA transferred by contact with skin during use. Since cordage is made by rubbing plant material between the hands and thighs and occasionally applying saliva, the cordage sample may contain DNA transferred during its manufacture.

4.2.2 Sequence data set

DNA was extracted from 9 samples in a clean room dedicated to ancient DNA work. Illumina sequencing libraries were prepared from the DNA extract. We enriched libraries for human mitochondrial DNA using biotinylated RNA baits that were based on the Reconstructed Sapiens Reference Sequence (RSRS) [11] (MYbaits v3; MYcroarray), following the manufac-

turer's instructions. The enriched libraries were sequenced on an Illumina MiSeq sequencer using v3 75-bp paired-end chemistry.

4.2.3 Sequence data processing

Prior to haplogroup analysis, sequenced reads for each moccasin library were independently processed to remove potentially confounding artifacts. First, read pairs were adapter trimmed and overlapping pairs were merged using SeqPrep (`https://github.com/jstjohn/SeqPrep`). Merged and unmerged fragments were filtered to remove low-complexity sequences using PRINSEQ-lite [91]. The sets of merged and unmerged fragments were combined and mapped to the Recontructed Sapiens Reference Sequence (RSRS) using `bwa aln` [63] and custom scripts to correct alignments that occurred at the ends of the circular genome (available from `https://github.com/svohr/circ_aln`). Duplicate sequences from PCR were removed using `samtools rmdup` [64] (Table 4.2.3).

While the baits used in the capture protocol were specifically designed to capture human mtDNA sequences, it is possible that other mammalian mtDNA sequences were enriched along with the human sequences. Conserved regions of the mitochondrial genome are especially susceptible to these off target fragments. To remove these fragments from consideration in haplogroup analysis, we implemented a strategy for filtering non-human fragments using results from BLAST [5]. For each read, we ran BLAST searches against a database of non-human sequences and a database that included only human sequences. We removed reads for which a search hit was found in the non-human database but not in the human database. If a hit was found in both databases, we removed reads where the E-value for the top hit in the non-human

Museum/ Field ID	Extract ID	Type	Raw Reads	Merged	Unmerged	% merged	Complexity Merged	Complexity Unmerged	Mapped (mq20)	rmdup	% Unique	BLAST Filter	Cov.
42 B01 FS801	JK332	Moccasin	501,217	454,326	13,573	90.6%	452,787	13,453	71,768	9,401	13.1%	9,005	35.79
42 B01 FS225	JK333	Leather	470,066	448,895	9,305	95.5%	448,525	9,275	13,622	2,653	19.5%	2,387	8.66
42 B01 FS235	JK334	Moccasin	512,127	487,565	10,462	95.2%	482,085	10,120	25,342	4,822	19.0%	4,575	18.39
42 B01 FS303	JK335	Plant fibre cordage	476,958	441,169	16,233	92.5%	440,649	16,171	35,605	4,379	12.3%	4,110	16.79
FS168	JK336	Moccasin	610,880	486,571	5,313	79.7%	485,801	5,280	4,527	684	15.1%	561	1.87
FS173	JK337	Moccasin	497,676	451,616	4,441	90.7%	448,989	4,367	14,801	2,826	19.1%	2,476	9.31
FS291	JK338	Moccasin	524,016	454,282	1,056	86.7%	452,326	1,010	1,291	980	75.9%	878	2.55
Cave#1 Moc#1	JK339	Moccasin	475,227	451,815	12,596	95.1%	449,798	11,561	56,755	5,462	9.6%	5,316	22.16
Cave#1 Moc#3	JK340	Moccasin	411,474	375,218	3,515	91.2%	374,282	3,458	30,556	1,788	5.9%	1,722	6.04
(Control)	JK-EC51	Extraction Control	358,599	298,270	44,437	83.2%	297,868	44,387	278,528	477	0.2%	475	2.15
(Control)	JK-EC52	Extraction Control	387,313	320,936	45,215	82.9%	319,749	45,005	203,107	1,049	0.5%	1,048	4.91

Table 4.1: Summary of sequence data processing for 9 samples and 2 extraction controls.

database was greater than the E-value for top hit in the human database. After filtering by BLAST search results, we found elevated sequence coverage around reference position 3,060, within the 16S RNA gene, in all moccasin samples. As a precaution, we excluded the region surrounding this peak (3,000–3,100) from variant analysis.

Sequenced DNA from ancient sources often contains excess C to T substitutions due to cytosine deamination near the 5' end of the sequenced fragment [26]. This process also results in complementary G to A substitutions near the 3' end. These errors can lead to problems in haplogroup analysis as the overwhelming majority of mtDNA SNP variation is due to transition mutations (C$\leftrightarrow$T, A$\leftrightarrow$G). Our software, mixemt, does not explicitly consider ancient DNA damage patterns when examining a fragment for any haplogroup informative markers. Instead, we masked the first 5 base pairs and last 5 base pairs of every read (marked as soft clipped), where base errors due to damage are most likely to appear.

When running mixemt on the moccasin reads, we chose parameters to fit the nature of the data sets. First, the number of fragments available in each moccasin set (472–8,921) was similar to the number of fragments used to test mixemt (see Chapter 3). However, the average length of each fragment in the moccasin data sets was considerably lower than that of modern sequence data. As a result, fewer sequenced base pairs are available for analysis and less haplotype information can be extracted from individual fragments. To avoid stochastic affects that may arise from the expectation maximization (EM) algorithm with the reduced amount of data available, we ran the EM step to convergence from multiple random starting points ($N = 10$) and took the mean estimated proportions and conditional probabilities. Second, since the sequence coverage from each moccasin is relatively low, it is possible that some diagnostic

alleles may not be observed in our read sets. This is especially a concern for our data sets, which could be complex mixtures of multiple haplotypes. As a result, the heuristic filtering step in mixemt that removes haplogroups for which the majority of diagnostic positions are not observed may be over aggressive in these cases. We ran mixemt with this step both enabled and disabled to compare results.

4.3 Results

4.3.1 Moccasin fragments exhibit patterns of ancient DNA damage

To verify that the libraries made from the DNA extracted from the moccasins contain ancient DNA, we used MapDamage [54] to examine each set of aligned sequences for the characteristics of ancient DNA. The mean fragment lengths in the moccasin samples (between 53 and 70 bp) were lower compared with the fragment lengths in the extraction control samples (76 and 79 bp). In each of the moccasin samples, we found elevated C to T substitutions at the 5' ends of sequences, and complementary G to A substitutions at the 3' ends (Figure 4.1). This increase in base substitutions was not found in either of the extraction controls.

4.3.2 Haplogroups detected in moccasin samples

We ran mixemt on the filtered reads for each moccasin with the parameters discussed in Section 4.2.3 to identify any haplogroups that were present. We found that the haplogroups that were detected were predominantly related to haplogroup B2 (Figure 4.2). The most common detected haplogroups were B2a, found in 7 moccasins, and B2y1, found in 6 moccasins.

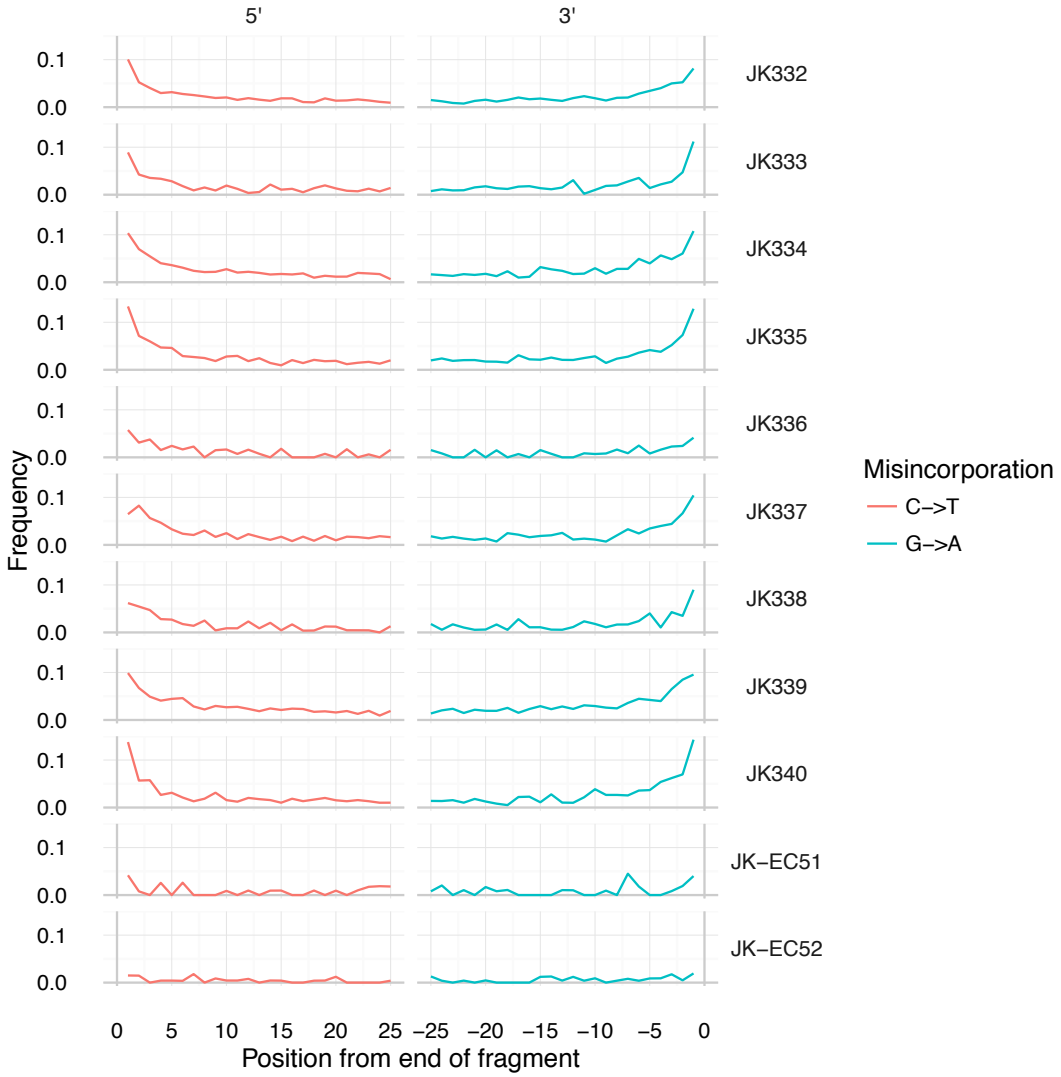

Figure 4.1: Ancient DNA base misincorporations in leather and cordage samples. In the moccasin, leather, and cordage samples, C to T transitions occur at higher rates near the 5' end of fragments mapped to the Reconstructed Sapiens Reference Sequence. Complementary G to A transitions are also observed at higher rate near 3' end of mapped fragments. These patterns in base misincorporation are not found in either of the extraction controls (JK-EC51 and JK-EC52).

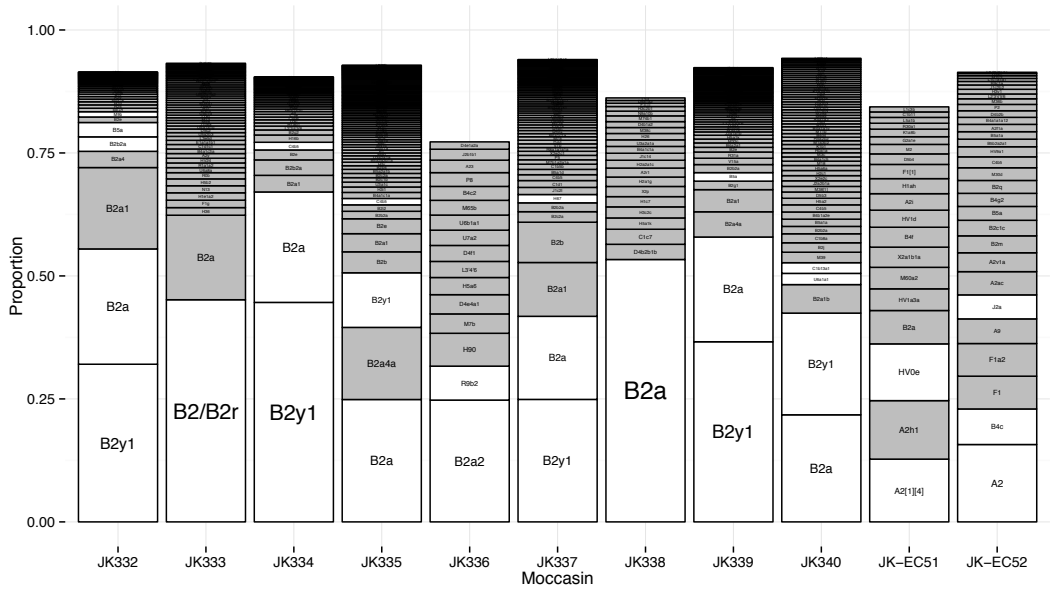

Figure 4.2: Haplogroup mixture proportions estimated by expectation maximization from sequenced whole mtDNA enriched libraries from moccasin and cordage samples (JK332–JK340) compared with two extraction controls (JK-EC51 and JK-EC52). Each column represents a single library. Haplogroups supported by at least 1 fragment are shown as a box scaled by the estimated contribution to the mixture and haplogroups with support from fewer than 10 fragments are shown in gray. Due to the short fragment sizes, many fragments do not overlap informative positions, leaving the haplogroup of origin ambiguous and increasing the overall uncertainty of the mixture estimates. However, haplogroups B2a and B2y1 are found consistently across samples but not in extraction controls. Note: Haplogroups B2 and B2r are collapsed into a single haplotype since B2r is defined by a deletion and only single nucleotide substitutions are considered.

In 8 out of the 9 moccasins, B2a or B2y1 was identified with the highest estimated proportion and in 6 moccasins, B2a and B2y1 were identified as the two highest contributors. The one moccasin where both B2a and B2y1 were not found (JK333) was one of the smaller data sets (2,227 fragments) and haplogroup B2, the parent haplogroup of both, was the only detected contributor. Several other haplogroups passed the detection requirements for mixemt, but each was estimated to make up less than 10% of the mixture individually. In the extraction controls,

neither B2a nor B2y1 were found in the detected haplogroups. In one of the extraction controls, B2a was associated with the fourth highest contribution (6.8%) of the mixture proportions. However, B2 was identified as the most likely contributor for only 2 fragments and the overall number of unique fragments in the extraction control was very low (447 total fragments).

4.3.3 Evidence for presence of haplogroup B2

To verify the presence of haplogroup B2 in the moccasin samples, we examined the 51 SNP positions that define haplogroup B2 in relation to the Reconstructed Sapiens Reference Sequence according to Phylotree [109]. Table 4.3.3 shows the number of times the derived allele (the version carried by haplogroup B2) and the inferred ancestral allele are observed within each sample. We found that all diagnostic bases for haplogroup B2 were observed in samples JK332 and JK334, all but 1 were observed in samples JK333, JK334, and JK339, and sample JK337 had only 2 missing bases. The missing observations are primarily due to lack of coverage of the reference sequence within a sample. The samples with the lowest amount of sequence data available (JK336, JK338, and JK340) had the highest number of unobserved diagnostic bases.

As many of the defining variants of B2 are shared with other haplogroups, we narrowed our focus to the 15 variant positions where B2 differs from our other haplogroup of interest, A2a (Table 4.3.3. In samples JK332, JK333, and JK334, derived bases were observed at all 15 diagnostic positions and only 1 derived base was not found in each of samples JK335, JK337, and JK339. In the samples with the lowest coverage and highest number of unobserved diagnostic positions (JK336, JK338, and JK340), we observed the derived base at each diagnostic position at the same rate.

			JK332		JK333		JK334		JK335		JK336		JK337		JK338		JK339		JK340		JK-EC51		JK-EC52	
Pos.	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc
152	T	C	18	1	8	0	12	0	8	0	0	0	5	0	1	0	15	0	8	0	0	0	1	1
182	C	T	23	0	5	0	15	0	15	0	3	0	2	0	0	0	16	0	7	0	2	0	3	0
195	T	C	18	7	2	1	8	4	10	4	3	0	0	0	0	0	9	8	1	3	2	0	2	1
247	G	A	31	0	5	0	14	0	12	0	1	0	4	0	0	0	14	1	6	0	3	0	3	0
769	G	A	21	0	9	0	8	0	14	0	0	0	3	0	0	0	14	1	2	0	0	0	5	0
825	T	A	14	0	2	0	4	1	11	0	2	0	2	0	0	0	11	0	0	0	1	0	6	0
1018	G	A	42	0	9	0	19	1	14	1	0	0	12	0	2	0	29	0	10	0	1	0	4	0
2758	G	A	16	1	1	0	8	0	14	0	1	0	4	0	0	0	9	1	6	0	1	0	2	0
2885	T	C	30	0	5	0	10	0	11	0	0	0	11	0	1	0	21	0	6	0	6	0	6	0
3594	C	T	20	1	6	0	11	0	9	0	0	0	8	0	1	0	9	0	0	0	0	0	0	0
4104	A	G	30	0	7	0	17	0	16	0	1	0	6	0	2	0	18	0	11	0	4	0	7	0
4312	C	T	26	1	4	1	8	0	9	0	0	0	8	0	0	0	11	0	3	0	3	0	5	0
7146	A	G	35	0	13	0	31	0	13	0	2	0	11	0	4	0	30	1	10	0	2	0	5	0
7256	C	T	30	0	6	1	12	2	16	0	0	1	5	0	1	0	17	2	5	0	1	0	1	0
7521	G	A	36	1	2	1	13	1	16	0	0	0	9	0	1	0	15	0	5	0	2	0	2	0
8468	C	T	15	0	0	0	7	0	8	0	0	0	2	0	1	0	12	0	4	0	0	0	1	0
8655	C	T	32	0	8	0	19	0	20	1	1	0	15	0	4	0	19	2	6	0	3	0	7	1
8701	A	G	29	1	4	0	17	0	20	1	0	0	10	2	2	0	20	1	6	0	1	0	10	1
9540	T	C	8	0	2	0	2	0	5	0	0	0	1	1	0	0	7	0	2	1	3	0	1	0
10398	A	G	31	1	7	0	19	1	8	0	0	0	6	0	1	0	18	0	6	0	0	0	4	1
10664	C	T	34	0	9	0	12	0	15	0	2	0	6	0	2	0	24	0	2	0	2	0	6	0
10688	G	A	27	0	10	0	9	1	13	0	0	0	6	0	0	0	12	3	2	0	2	0	3	0
10810	T	C	36	0	8	0	9	0	12	0	0	0	5	0	3	0	32	0	5	0	4	0	5	0
10873	T	C	24	0	4	1	12	1	11	0	2	0	5	0	1	0	17	0	2	2	3	0	3	0
10915	T	C	34	0	7	0	26	0	20	0	3	0	12	0	2	0	14	0	5	0	3	0	4	0
11914	G	A	29	1	8	1	21	0	15	1	6	0	12	0	4	0	21	3	2	0	1	0	5	0
13105	A	G	41	1	6	0	22	0	16	0	1	0	9	0	3	0	15	0	4	0	3	0	8	0
13276	A	G	33	0	3	0	20	0	19	0	2	0	3	0	2	0	19	0	7	0	0	0	6	0
13506	C	T	30	3	7	0	21	1	22	0	0	0	8	0	3	0	31	0	11	0	3	0	4	0
13650	C	T	27	0	7	0	15	0	11	0	2	0	4	0	4	0	11	0	5	0	0	0	2	1
15301	G	A	27	2	12	1	18	0	18	1	1	1	9	0	2	0	20	2	6	0	0	0	4	0
16129	G	A	33	0	8	0	13	0	9	0	3	0	11	0	2	0	14	0	5	0	3	0	8	0
16187	C	T	4	0	2	0	2	0	3	0	0	0	1	0	0	0	1	0	0	0	1	0	2	0
16230	A	G	5	0	3	0	5	0	5	0	1	0	1	0	2	0	4	0	1	0	2	0	2	0
16278	C	T	18	0	4	0	3	0	7	0	1	0	5	0	0	1	11	0	4	0	1	0	0	0
16311	T	C	22	0	7	0	10	0	11	0	1	0	5	0	0	0	11	2	5	0	0	0	2	0

Table 4.2: Base observations from moccasin samples at diagnostic positions for haplogroup B2 shared by haplogroup A.

			JK332		JK333		JK334		JK335		JK336		JK337		JK338		JK339		JK340		JK-EC51		JK-EC52	
Pos.	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc
146	T	C	14	3	8	0	11	0	6	0	0	0	6	0	1	0	12	0	5	0	0	0	0	2
499	A	G	11	1	7	0	10	1	4	2	3	2	10	1	1	0	9	0	2	2	0	5	0	12
827	G	A	12	2	2	0	3	1	10	1	1	1	2	0	0	0	10	1	0	0	0	1	1	5
3547	G	A	21	2	5	1	8	0	10	3	1	0	4	1	1	0	13	0	5	1	0	0	0	1
4820	A	G	24	4	6	1	17	0	10	1	0	0	5	2	4	0	12	1	2	0	0	4	0	6
4977	C	T	39	2	7	2	18	3	19	3	0	1	15	0	4	0	27	3	3	0	0	0	0	0
6473	T	C	33	4	7	0	13	2	16	0	2	1	9	2	1	0	18	0	6	2	2	0	2	6
9950	C	T	29	5	4	0	17	2	12	2	0	0	5	0	0	0	11	2	2	1	0	2	2	2
11177	T	C	34	0	3	0	19	1	11	2	1	0	4	1	2	0	19	1	3	1	0	2	0	9
12705	C	T	17	0	6	2	12	1	14	1	3	1	7	0	0	0	22	2	0	1	0	0	0	0
13590	A	G	21	3	7	3	16	2	13	3	1	1	7	1	2	0	13	3	3	1	0	2	0	3
15535	T	C	13	0	2	1	13	1	11	1	1	1	1	2	0	0	12	1	5	2	0	3	1	3
16189	C	T	3	1	1	1	1	1	0	3	0	0	0	1	0	0	0	1	0	0	1	0	0	2
16217	C	T	4	1	3	0	2	2	2	3	0	0	1	0	1	0	1	1	1	0	0	1	0	2
16223	C	T	2	2	3	0	2	2	4	1	1	0	1	0	1	0	1	1	1	0	0	2	0	2

Table 4.3: Base observations from moccasin samples at diagnostic positions for haplogroup B2 not shared by haplogroup A.

4.3.4 Evidence for absence of haplogroup A2a

To verify that haplogroup A2a cannot be detected in the moccasin samples, we examined the bases present at the haplogroup defining positions that are not shared with haplogroup B2 (Table 4.3.4). In all moccasin samples, we saw that fragments that overlapped unique diagnostic positions for A2a predominantly carried the inferred ancestral bases rather than the derived bases that define the A2a haplogroup. The counts of derived bases were similar to what was found in the extraction controls. The exception to this is variant C16111T where the counts for the derived base were elevated in several moccasins. However, this variant is also shared with haplogroup B2a and the elevated derived base count may be due to the presence of this sub-haplogroup of B2 in the mixture. We conclude that haplogroup A2a is not present in the DNA recovered from these moccasin samples.

4.3.5 Evidence for mixture of haplogroups B2a and B2y1

Our software, mixemt, detected haplogroups both B2a and B2y1 in 6 out of the 9 moccasin samples (JK332, JK334, JK335, JK337, JK339, and JK340). Since these two haplotypes differ by only 4 single base substitutions, we examined each position for evidence that these sets of fragments represent a mixture of these two haplogroups (Table 4.3.5). Haplogroup B2a is differentiated from B2 by variants C16111T and G16483A while B2y1 is defined by variants A3480G and C16261T. A mixture of these two haplogroups should contain both the derived and ancestral bases at each position. In the samples where B2a and B2y1 were detected, we observed all 8 variants in 4 out of the 6 samples and all but the derived variant at position 16,261 in the remaining samples.

			JK332		JK333		JK334		JK335		JK336		JK337		JK338		JK339		JK340		JK-EC51		JK-EC52	
Pos.	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc
146	C	T	3	14	0	8	0	11	0	6	0	0	0	6	0	1	0	12	0	5	0	0	2	0
153	G	A	2	18	0	8	0	12	0	8	0	0	0	5	0	1	0	14	0	8	0	0	2	0
235	G	A	2	31	0	5	0	13	0	12	0	0	0	3	0	0	0	19	0	8	2	0	2	0
663	G	A	1	26	0	9	0	16	0	17	0	3	0	9	0	3	0	25	0	5	0	2	1	2
1736	G	A	0	31	0	5	0	14	0	7	0	0	0	7	0	1	0	19	0	11	1	0	0	3
3330	T	C	1	29	0	9	0	15	0	15	0	3	0	12	0	3	0	19	1	2	0	3	0	4
4248	C	T	0	34	0	8	0	14	0	12	0	1	0	8	0	3	0	16	0	6	0	0	3	8
4824	G	A	0	28	0	7	0	15	0	11	0	0	0	6	0	4	0	13	0	2	0	4	3	4
8027	A	G	1	30	0	5	0	16	1	17	0	1	0	9	0	0	0	25	0	2	0	0	0	2
8794	T	C	0	37	0	12	0	25	0	17	0	1	0	15	0	6	0	23	0	7	1	0	2	4
12007	A	G	0	33	0	5	0	16	0	20	0	0	0	14	0	2	0	26	0	6	0	2	1	7
16111	T	C	25	27	3	14	7	15	6	10	1	2	6	13	4	2	8	17	2	2	3	0	8	3
16189	T	C	1	3	1	1	1	1	3	0	0	0	1	0	0	0	1	0	0	0	0	1	2	0
16192	T	C	0	4	0	1	0	2	0	3	0	1	0	0	0	0	0	1	0	0	0	1	0	2
16290	T	C	0	14	0	3	0	6	0	7	0	0	0	5	0	1	0	11	0	5	0	0	0	1
16319	A	G	0	25	0	8	0	12	0	13	0	1	0	4	0	0	0	16	0	5	0	0	1	1
16362	C	T	0	20	0	6	0	13	0	9	0	1	0	7	0	0	0	11	0	4	0	0	1	3

Table 4.4: Base observations from moccasin samples at diagnostic positions for haplogroup A2a not shared by B2.

				JK332		JK333		JK334		JK335		JK336		JK337		JK338		JK339		JK340		JK-EC51		JK-EC52	
Haplogroup	Pos.	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc	Der	Anc
B2a	16111	T	C	25	27	3	14	7	15	6	10	1	2	6	13	4	2	8	17	2	2	3	0	8	3
B2a	16483	A	G	7	8	1	0	4	9	7	4	1	0	2	5	0	0	7	4	3	7	0	0	0	2
B2y1	3480	G	A	11	21	0	5	8	7	2	12	0	0	1	2	0	0	10	14	2	1	0	2	0	5
B2y1	16261	T	C	4	6	1	6	0	2	2	6	0	1	1	3	0	2	7	6	0	3	0	2	0	3

Table 4.5: Base observations from moccasin samples at positions where haplogroups B2a and B2y1 differ.

The observation of a derived allele at a variant position cannot simply be interpreted as evidence for a haplogroup's presence. Some variants appear independently in different haplogroups due to back mutations. The observation of a derived base at position 16,111 provides little support for the presence of B2a; the derived base was observed in the extraction controls and the mtDNA phylogeny from Phylotree.org infers that mutations have occurred at the position 29 times throughout the phylogeny. However, the variant G16483A is only known from haplogroup B2a and the derived base was not observed in the extraction controls. For the variants defining haplogroup B2y1, a mutation producing A3480G is inferred to have occurred 3 times while 33 mutations were inferred at position C16261T. Neither derived base occurs in the extraction controls.

It is difficult to conclude that these samples represent genuine mixtures of distinct haplotypes from these observations. However, the fact the B2a and B2y1 were independently detected in nearly all samples suggests that haplogroup assignments below B2 are not the result of random patterns of DNA damage. Furthermore, it is unlikely that the mixture signal is driven by incomplete data as the mixtures were detected in the samples with the highest coverage.

4.4 Discussion

We have demonstrated that a strategy of whole mitochondrial sequence enrichment and high-throughput shotgun sequencing can be used to recover ancient human DNA transferred by contact with artifacts. Central to this approach is the ability to identify haplogroups present in a mixture of mtDNA sequences in a fast and automated way. Unlike PCR-based experiments,

which must be designed to focus on a set of haplogroups and diagnostic regions, our approach allows for the detection of haplogroups across the entire mitochondrial phylogeny [109]. This feature is especially useful in identifying unexpected haplogroups including contamination from modern human DNA.

Ancient DNA presents unique challenges to the interpretation of mixtures with our software mixemt. Since ancient DNA is highly fragmented many sequences may not contain any information that can be used to assign them to a haplogroup. This can be seen in our results as much of inferred mixture proportions were attributed to many low-contribution haplogroups. Another consequence of the short fragment size is that there is little haplotype information present within individual reads that can be used to discern the underlying haplotypes. Although mixemt was not specifically designed to handle ancient DNA, the program still produces consistent and interpretable results on these data sets.

A recurring question in the field of archaeology is whether changes in material culture observed in the archaeological record represent the spread of cultures or the movements of populations ("pots versus people") [83]. Ancient contact DNA has the potential to aid in answering these questions as DNA evidence can be obtained directly from artifacts and from sites where no human remains are present. Since artifacts may have been in contact with mulitple individuals, and therefore contain a mixture of DNA, mitochondrial mixture analysis plays an important role in extracting the most information from these sources.

Chapter 5

Conclusions

In this dissertation I have presented two approaches for forensic identification using high-throughput sequencing data and demonstrated their use on real and simulated data sets. Both approaches make use of descriptions of human genetic diversity to interpret the heterogeneous data produced by shotgun sequencing. In Chapter 2, I presented an approach for determining whether two extremely low-coverage DNA libraries share a common source individual, using patterns of linkage disequilibrium rather than direct comparisons between alleles. In Chapter 3, I described an approach to deconvolve mixtures of mitochondrial sequences based on haplogroups described by the inferred human mitochondrial phylogeny [109] and in Chapter 4, I demonstrate its utility on ancient contact DNA extracted from $\sim$ 700 year old leather and cordage samples.

The methodologies described in this dissertation have enormous potential for application in a variety of fields. Some of the most obvious applications can be found in archaeological work with ancient DNA where high-throughput shotgun sequencing is already routinely em-

ployed. For most ancient samples, high-coverage genomes are not obtainable, either because of low endogenous DNA content or limited sequencing resources. As a result, tools that are specifically designed for low-coverage data are required to learn the most from samples that are limited by the amount of material available for extraction or by the complexity of the recovered molecules. Low-coverage methods are also important for large ancient DNA sequencing projects, where generating high-coverage genomes for many samples is not feasible [34, 4, 45].

Although studying ancient DNA has produced insights into large-scale human population history, there is also great potential for ancient DNA to help reveal the local history of an archaeological site [78]. For example, our method for testing if two low-coverage libraries share a common source individual could be applied to sparse sequencing of DNA extracted from remains found in comingled burial contexts to determine the number of individuals present and associate remains with specific individuals. Furthermore, an extension of our method to detect first-order relationships would be useful for detecting kinship from excavated burials. While these types of questions have been previously addressed with PCR-based methods (e.g, [92, 56, 24]), a shotgun-based approach may be better suited for degraded or minute samples and for assessing ancient DNA authenticity. In addition, we have demonstrated that our method for mixture interpretation can be used as to determine mitochondrial haplogroups from ancient contact DNA, but our approach could be also applied to other sources of DNA mixtures (e.g., environmental DNA [103]), or used in quality control to identify sources of contamination.

There is also great potential for the application of our methods in forensics casework. First, our approach for interpreting mixtures of mitochondrial sequences addresses a critical issue for many forensics samples and can be used simply to detect mixed samples or to recon-

struct near complete haplotypes for each contributor. The method can be equally applied to known mixture samples related to a case or used to detect contamination from outside sources. Although I evaluated the method using relatively short, shotgun sequences, our approach can also be applied to read pairs with longer inserts, long reads, or sequenced amplification products from one or more loci. Second, our approach for comparing extremely low-coverage sequencing libraries could also be applied to highly degraded or low-template DNA samples. Our method, which relies on linkage information to determine whether two sparse, non-overlapping sequence data sets originated from the same individual, is fundamentally different from the standard forensics approach of generating a profile and comparing the results to identify individuals. Although these standard approaches can produce highly sensitive results very quickly, PCR-based approaches fail when the template molecules required for amplification simply do not exist within the sample. Shotgun sequencing may be the only way to extract any nuclear genetic information from some degraded and low-template DNA samples. The linkage-based identification method described here is uniquely suited to these types of data sets.

Finally, the methods I have described could also be generalized to address similar problems not related to human identification. For example, given a reference haplotype data set and sufficient linkage disequilibrium within the population, our approach for low-coverage identification could be applied to non-human species. Our method could be particularly useful in conservation genomics for noninvasive monitoring of individuals in threatened populations [77]. Similarly, given descriptions of haplotypes, our approach to mixture decovolution could be applied to mixtures of non-human mitochondrial sequences or generalized to any linkage group that does not regularly undergo recombination (e.g., the Y chromosome). As mixtures of

DNA from distinct haplotypes and low quanitities of recoverable DNA are recurring issues in many fields related to genomics, there are likely many additional applications for the methods I have presented in this disseration.

Appendix A

Additional materials for "A method for positive forensic identification of samples from extremely low-coverage sequence data"

A.1 Coalescent simulations

We performed coalescent simulations to test our method free of base errors and under various demographic scenarios. We used the coalescent simulator `ms` [49] to simulate diploid individuals and population reference panels of haplotypes for comparison. For each replicate, we simulated 3,000 independent segments of 500 Kb in size for a total of 1.5 Gb. Segregating sites with minor allele frequencies lower than 10% were removed. Reference panels consisted of 200 haplotypes. For diploid individuals, we simulated base observations from low-coverage sequencing by randomly drawn an allele from segregating sites at a rate of 0.01. This was done separately for each chromosome and sites where both alleles were observed were discarded,

resulting in ~ 0.02-fold coverage. This process was repeated to construct multiple observation sets per individual.

The following `ms` commands were used to simulate samples from diploid individuals and reference panels of haplotypes.

A.1.1 Simulation 1: Single constant-size population

```
ms 204 3000 -t 200 -r 200 500000
```

A.1.2 Simulation 2: Individuals and reference panel drawn from different populations

Populations split 100 generations ago.

```
ms 204 3000 -t 200 -r 200 500000 -I 2 4 200 -ej 0.0025 1 2 -n 1 1 -n 2 1
```

Populations split 500 generations ago.

```
ms 204 3000 -t 200 -r 200 500000 -I 2 4 200 -ej 0.0125 1 2 -n 1 1 -n 2 1
```

Populations split 1,000 generations ago.

```
ms 204 3000 -t 200 -r 200 500000 -I 2 4 200 -ej 0.025 1 2 -n 1 1 -n 2 1
```

Populations split 2,000 generations ago.

```
ms 204 3000 -t 200 -r 200 500000 -I 2 4 200 -ej 0.05 1 2 -n 1 1 -n 2 1
```

Populations split 4,000 generations ago.

```
ms 204 3000 -t 200 -r 200 500000 -I 2 4 200 -ej 0.1 1 2 -n 1 1 -n 2 1
```

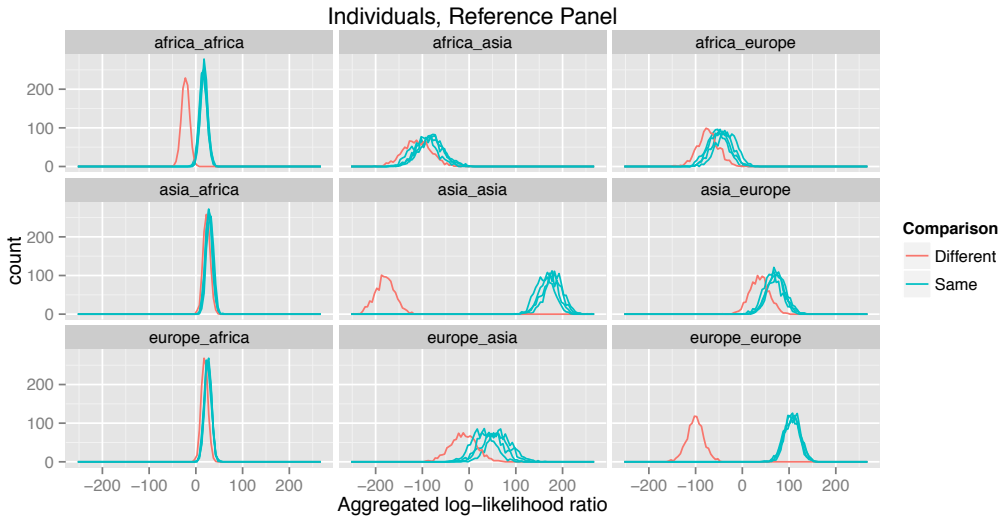

Figure A.1: Comparisons between samples from diploid individuals using different reference panels (window labels indicate populations for "*Individuals_ReferencePanel*").

A.1.3 Simulation 3: Recent human demographic history

We tested our method on simulated data based on parameters inferred from allele frequency data [43]. This `ms` command below describes the population model for the 3 populations (Africa, Europe, Asia). We repeated the command for each population comparison, changing the number of chromosomes sampled from each population.

```
ms 204 3000 -t 470 -r 200 500000 -I 3 204 0 0
    -n 1 1.23 -n 2 0.1 -n 3 0.051
    -en 0.0189 2 0.21 -en 0.1964 1 0.730
    -m 3 2 1.92 -m 2 3 1.92 -m 3 1 0.38 -m 1 3 0.38
    -m 2 1 0.6 -m 1 2 0.6
    -em 0.0189 2 1 5 -em 0.0189 1 2 5
    -ej 0.0189 3 2 -ej 0.125 2 1
```

In addition to the within population comparisons described in the paper, we also compared samples from diploid individuals from one population using a reference panel drawn from another population. We found that power to distinguish between different individuals is dimin-

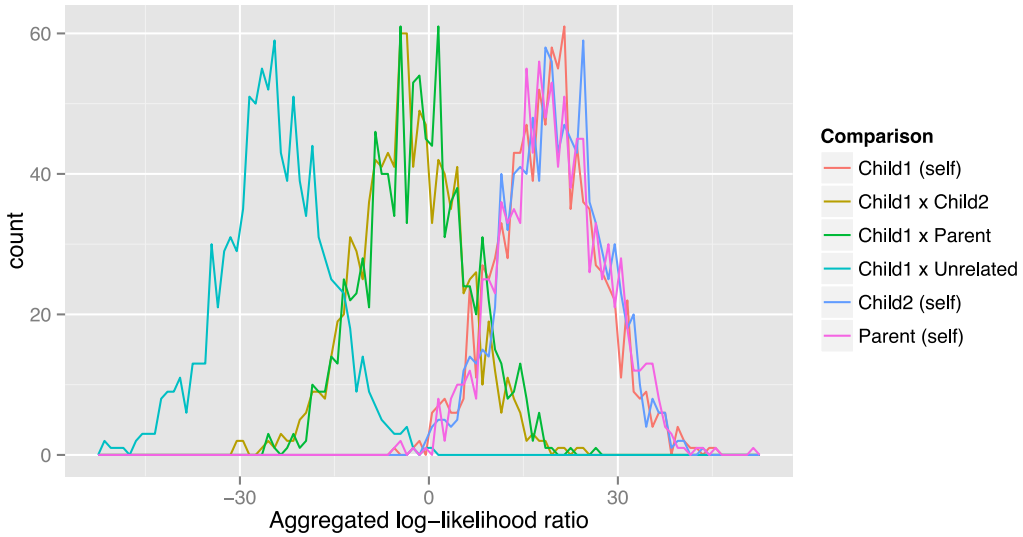

Figure A.2: Results of comparisons of simulated first-order relatives are distinct when compared with results from unrelated individuals and the same individual.

ished when the reference panel drawn from a distantly related population.

A.1.4 Simulation 4: Related individuals

We tested the utility of our method to detect close familiar relationships between individuals using low-coverage samples. Under the same population model as Simulation 1, we simulated diploid individuals that would be "parents" for two child individuals. We generated 2 children by drawing 1 chromosome at random from each diploid parent for each simulated segment. This produced two children that shared $\sim 50\%$ of the genome with each other and with their parents. Allele observations were sampled from the diploid children using the same strategy as before. We compared the two children with samples from one parent, and a third, unrelated diploid individual drawn directly from the simulated haplotypes. As expected, the aggregated log-likelihood values for the parent-child and sibling-sibling comparisons fall directly

between the same individual comparisons and the unrelated individual comparisons (Figure A.2).

A.2 Experiments with sequencing data

We constructed reference panels of single nucleotide polymorphisms (SNPs) using the 1000 Genomes Project Phase 1 data set [101]. We filtered for biallelic SNPs that were polymorphic in the target population with a minimum minor allele count of 10. To avoid errors from mismapped reads, we restricted our panels to sites where all overlapping 35mers are unique across hg19 according to the Duke Uniq 35 track from the Mapability tracks on the UCSC Genome Browser. We constructed 7 reference panels in total; 5 from Europe (CEU, FIN, GBR, IBS, and TSI), 1 from Asia (CHB) and 1 from Africa (YRI).

We obtained Illumina sequencing data from a European male (NA12891) and a European female (NA12892) sequenced as part of Platinum Genomes by Illumina, Inc (`http://www.illumina.com/platinumgenomes/`) from the National Center for Biotechnology Information Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra/`, accession ERR194160 and ERR194161). We sampled 3,200, 32,000, 160,000, 320,000 and 1,600,000 reads without replacement to approximate 0.01%, 0.1%, 0.5%, 1.0%, and 5.0%$\times$ coverage of the nuclear genome to generate two sets at each coverage level for both individuals. Reads were mapped to the human reference sequence (GRCh37/hg19) using `bwa mem` [62] with default parameters.

Single base observations were made from mapped reads using output from `samtools mpileup`. Reads were required to have a map quality greater than 30 and bases were required

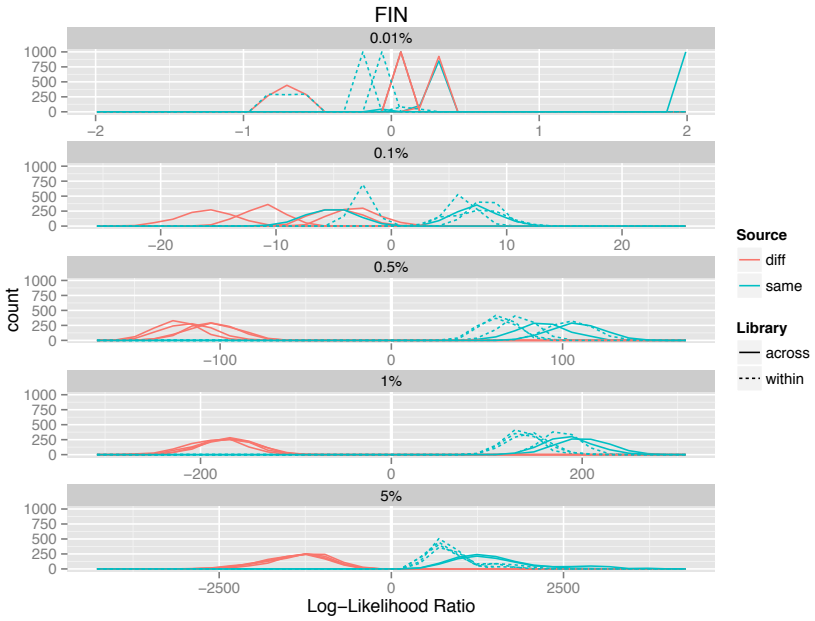

Figure A.3: Results from low-coverage sequencing comparison using Finnish in Finland (FIN) population as the reference panel.

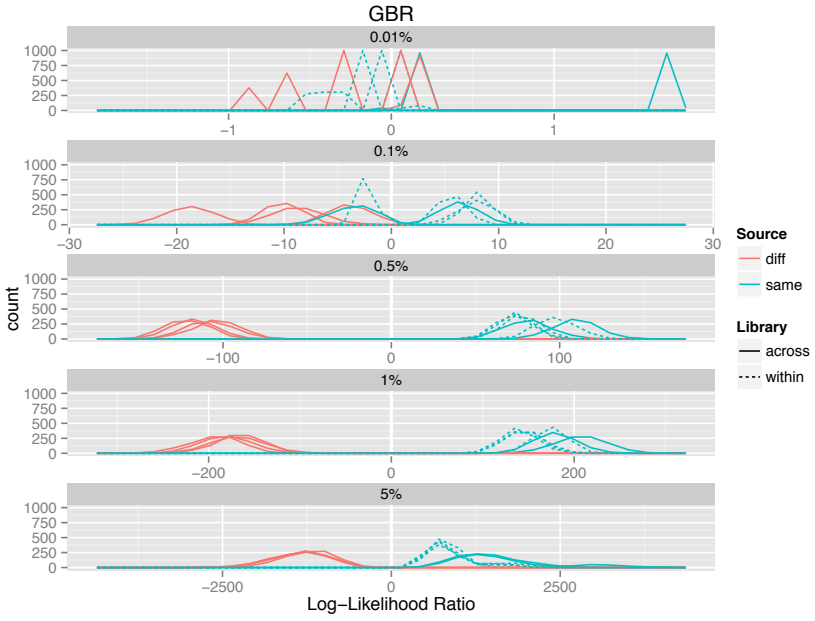

Figure A.4: Results from low-coverage sequencing comparison using British in England and Scotland (GBR) population as the reference panel.

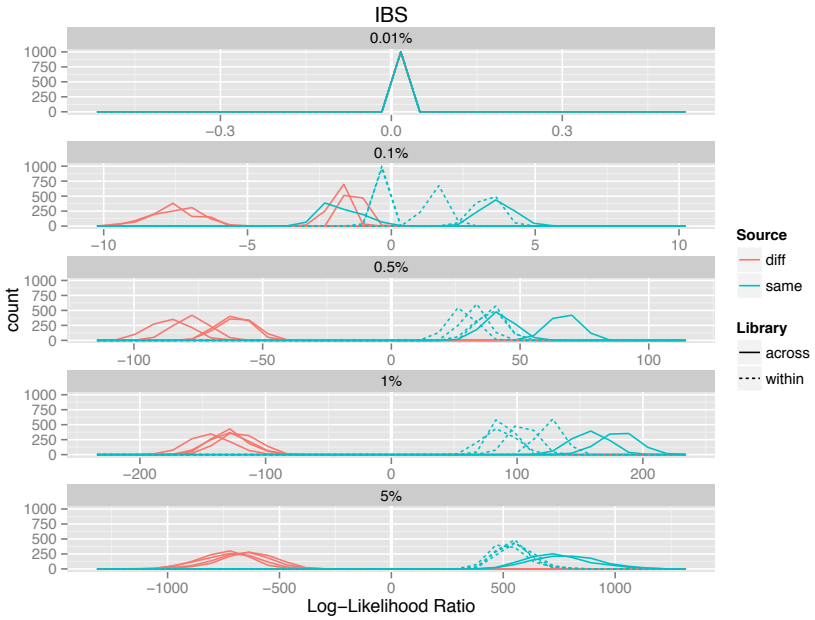

Figure A.5: Results from low-coverage sequencing comparison using Iberian populations in Spain (IBS) population as the reference panel.

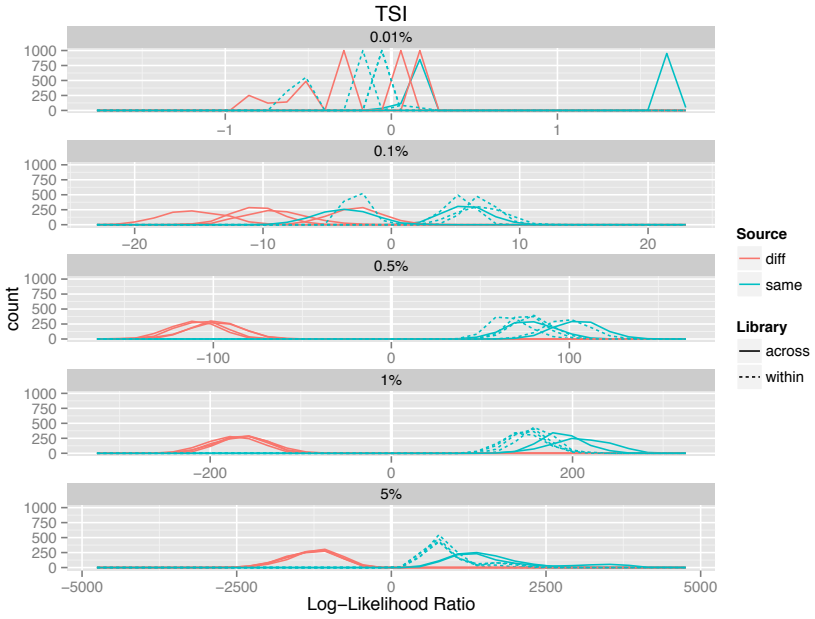

Figure A.6: Results from low-coverage sequencing comparison using Toscani in Italia (TSI) population as the reference panel.

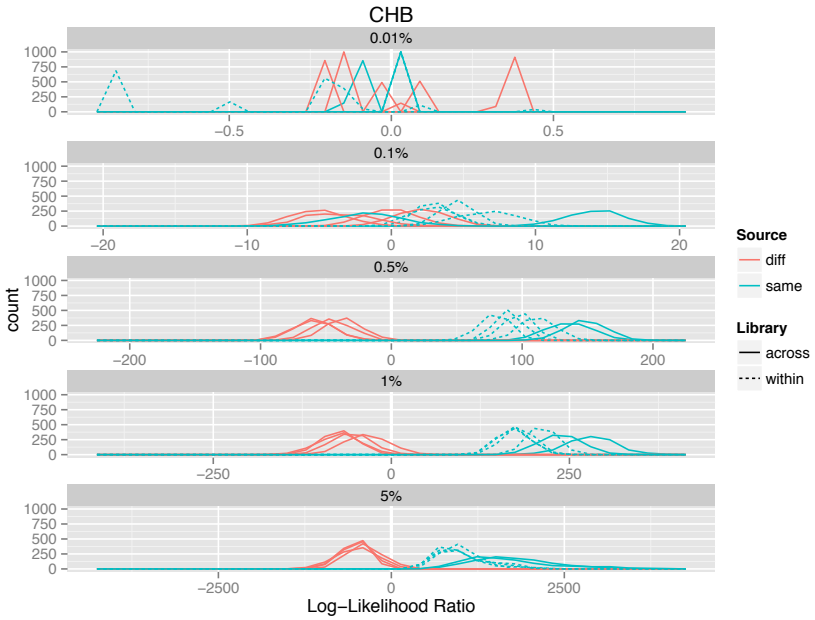

Figure A.7: Results from low-coverage sequencing comparison using Han Chinese in Beijing, China (CHB) population as the reference panel.

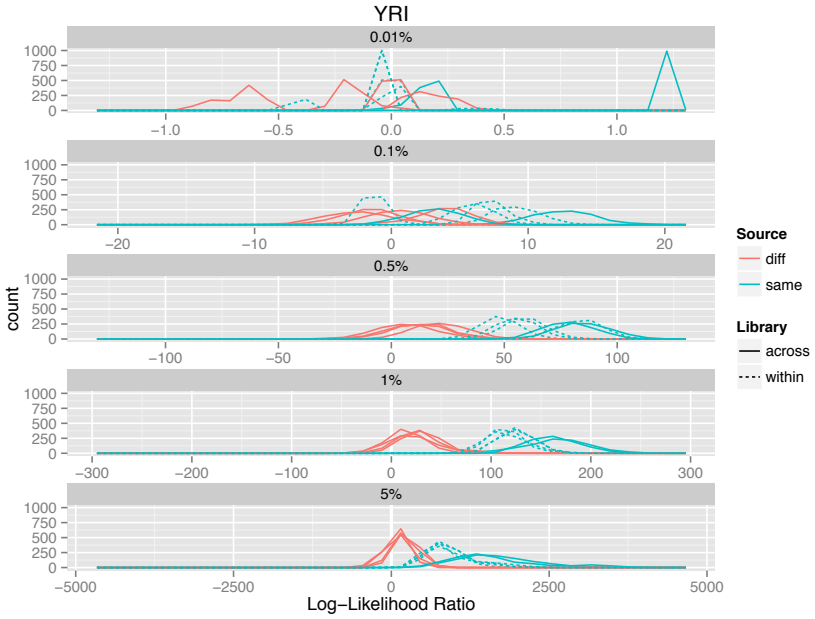

Figure A.8: Results from low-coverage sequencing comparison using Yoruba in Ibadan, Nigeria (YRI) population as the reference panel.

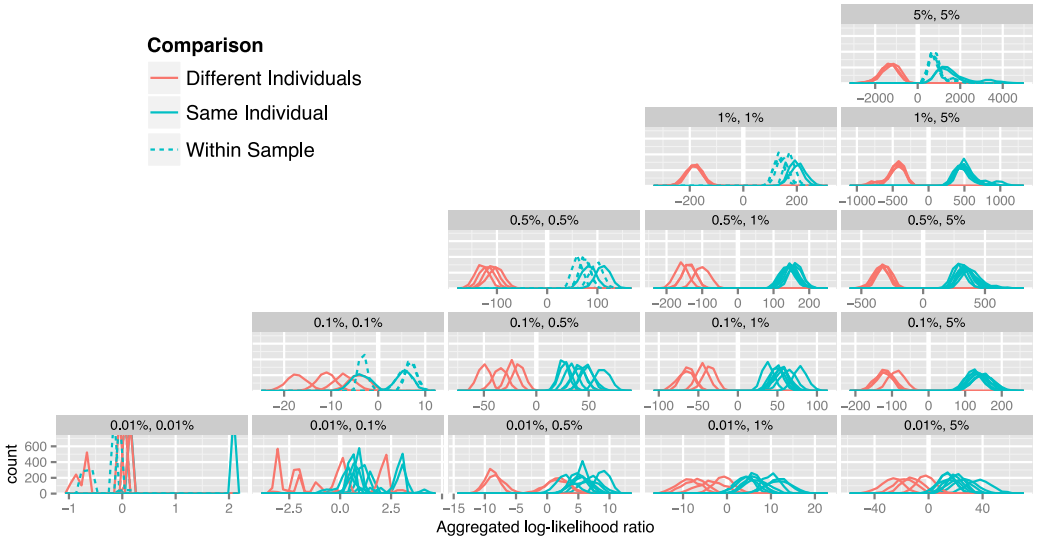

Figure A.9: Comparisons of samples with different levels of coverage using CEU reference panel.

to have a base quality of 30. Positions where more than one read mapped were excluded.

The comparison describe in the main text was repeated with the substitution of the CEU reference panel with alternatives (FIN, GBR, IBS, TSI, CHB, YRI). We found that all European populations produced similar results.

Comparisons were also made across coverage levels for the original CEU reference panel (Figure A.9). This demonstrates that comparisons with a sample with extremely low coverage can benefit higher coverage in the other sample.

A.3 Ancient DNA

Illumina sequencing data from DNA extracted from 12 samples from 11 Bronze Age Eurasian humans [4] were downloaded from the European Nucleotide Archive (project

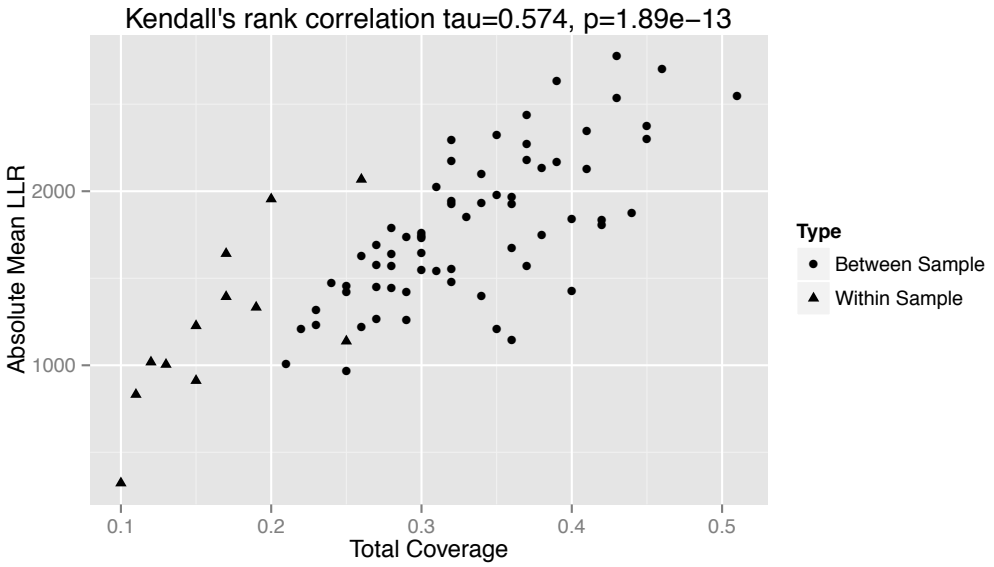

Figure A.10: In ancient DNA samples, total sequence coverage in comparison is correlated with the absolute value of the mean aggregated log-likelihood ratio.

accession number PRJEB9021). We downloaded mapped reads in BAM format for samples RISE109, RISE154, RISE240, RISE247, RISE480, RISE483, RISE507, RISE508, RISE510, RISE546, RISE554, and RISE586.

Again, single base observations were made from mapped reads using output from `samtools mpileup` [64]. Reads were required to have a map quality greater than 30 and bases were required to have a base quality of 30. Positions where more than one read mapped were excluded. In addition, only mutations that represented transversion mutations were included in the analysis.

Appendix B

Additional materials for "A fast method for mitochondrial haplotype analysis of sequence data from complex DNA mixtures"

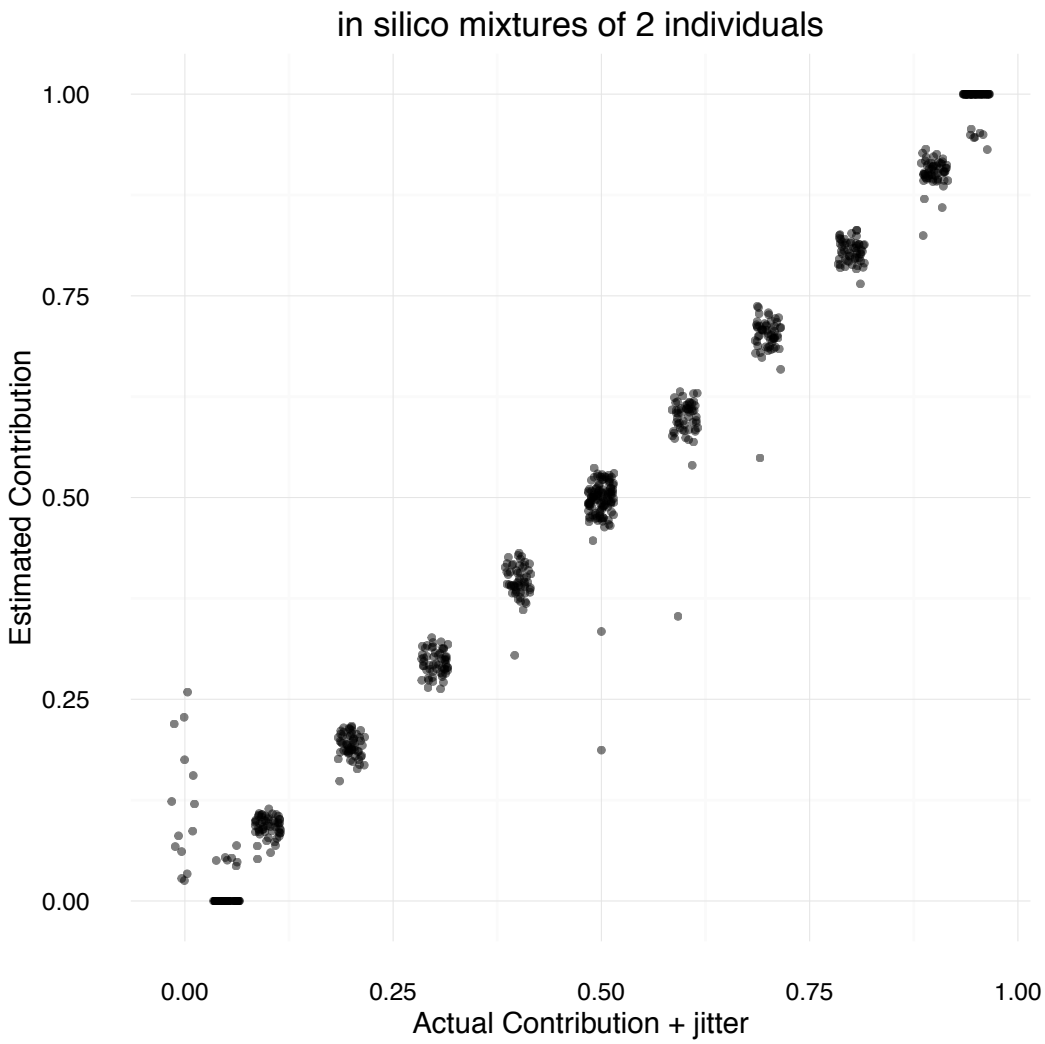

Figure B.1: Contribution estimates for *in silico* mixtures of 2 individuals (same data set as Figure 3.2) refined with a second round of expectation maximization performed on only the haplogroups that pass all filtering steps.

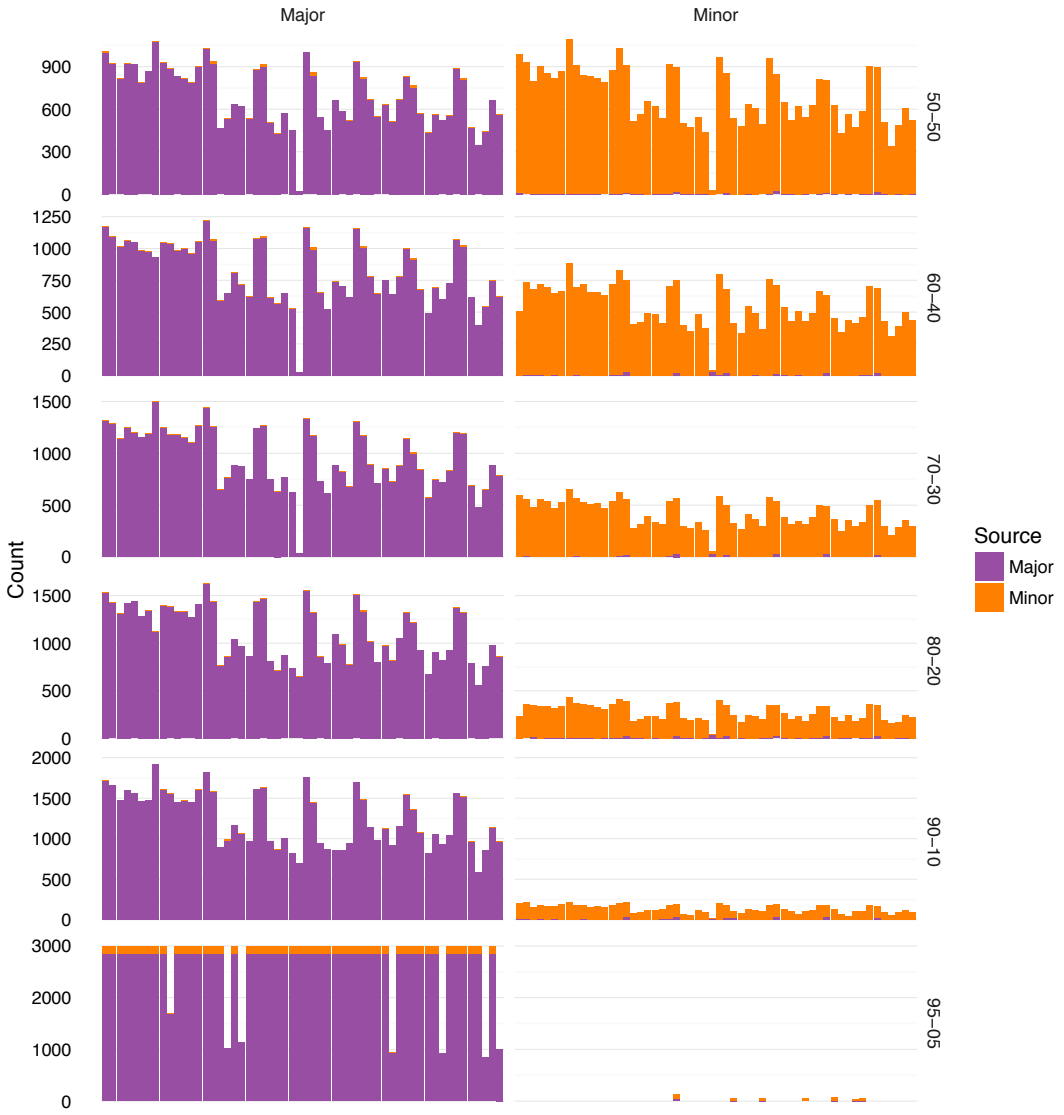

Figure B.2: Source contributor for each fragment after partitioning by detected haplogroups. Counts for fragments originating from the major contributor are shown in purple and counts for fragments from the minor contributor are shown in orange (3,000 total fragments in each mixture). The vast majority of fragments are assigned to the correct contributor. Misassigned fragments in the 95-5 mixtures are due to the failure to detect the minor haplogroup at this coverage level.

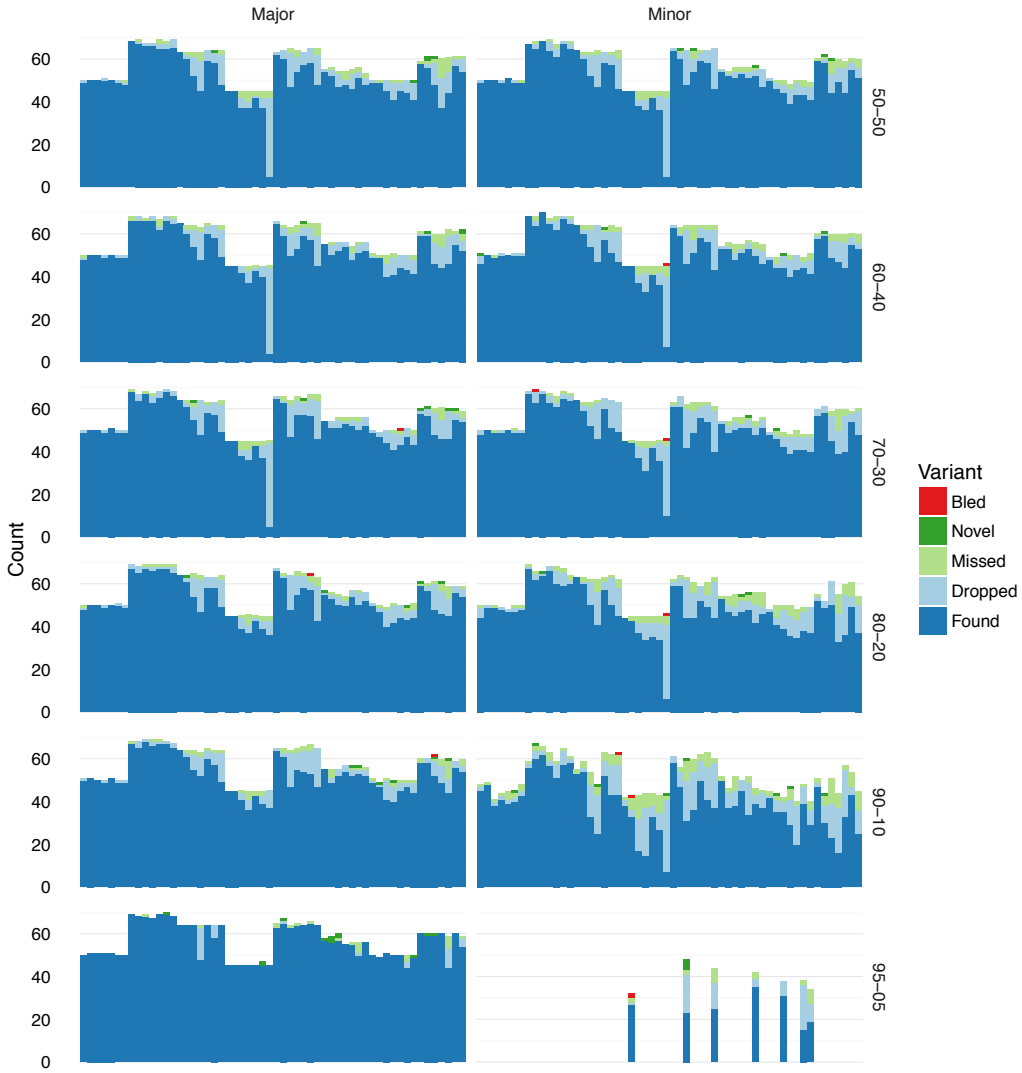

Figure B.3: Summary of variants detected from single-source fragment sets and mixtures for 2 individuals. Variants were called from individual contributions and fragment sets partitioned by haplogroup. Counts are shown for variants detected in both (“Found”), variants not detected in the mixemt results due to lack of coverage (“Dropped“), variants that were covered but not detected (“Missed“), variants that were detected in the mixemt results but not in the single-individual set (“Novel”) and variants that are not found in the single-individual set but do occur in the other contributor (“Bled”). We found that most variants are recovered correctly, most missing variants are due to lack of coverage, and that there are only a few occurrences of novel variants or variants that bleed-over from the other contributor.

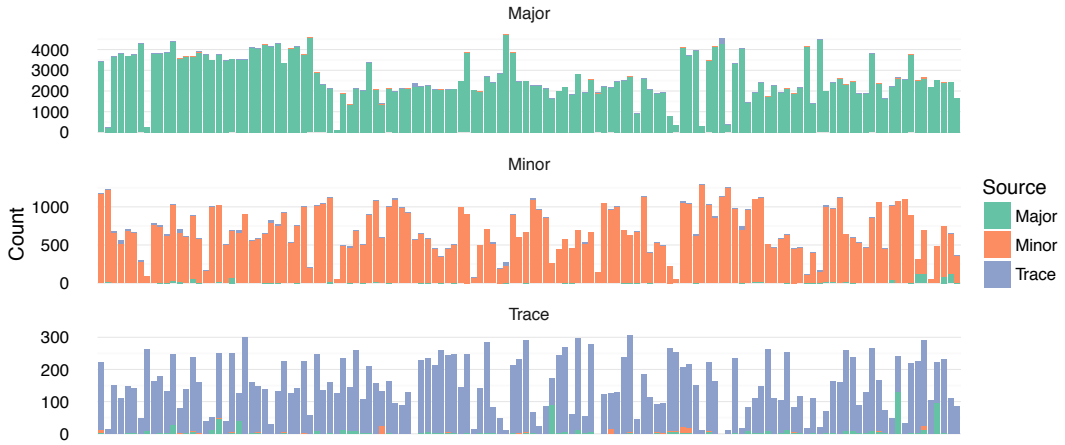

Figure B.4: Source contributor for each fragment after partitioning by detected haplogroups in *in silico* mixtures of 3 individuals. Counts for fragments originating from the major contributor are shown in green, counts for fragments from the minor contributor are shown in orange, and counts for fragments from the trace individual are in blue (10,000 total fragments in each mixture). The fragments are generally assigned to the correct contributor. The reconstructed set for the smallest contributor ("Trace") is affected the most by misassigned fragments.

Bibliography

[1] A. Achilli, U. A. Perego, H. Lancioni, A. Olivieri, F. Gandini, B. Hooshiar Kashani, V. Battaglia, V. Grugni, N. Angerhofer, M. P. Rogers, R. J. Herrera, S. R. Woodward, D. Labuda, D. G. Smith, J. S. Cybulski, O. Semino, R. S. Malhi, and A. Torroni. Reconciling migration models to the Americas with the variation of North American native mitogenomes. *Proceedings of the National Academy of Sciences*, 110(35):14308–14313, 2013.

[2] C. J. Adler, W. Haak, D. Donlon, and A. Cooper. Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science*, 38(5):956–964, 2011.

[3] M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. P. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1748):4724–4733, Dec 2012.

[4] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. D. Casa, P. Dąbrowski, P. R. Duffy, A. V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolá, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtchyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, Ł. Pospieszny, T. D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V. I. Soenov, V. Szeverényi, G. Tóth, S. V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S. Brunak, R. Nielsen, K. Kristiansen, and E. Willerslev. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172, 2015.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct 1990.

[6] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23(2):147, Oct 1999.

[7] M. C. Avila-Arcos, E. Cappellini, J. A. Romero-Navarro, N. Wales, J. V. Moreno-Mayar, M. Rasmussen, S. L. Fordyce, R. Montiel, J.-P. Vielle-Calzada, E. Willerslev, and M. T. P. Gilbert. Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Scientific Reports*, 1:74, 2011.

[8] D. J. Balding. Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proceedings of the National Academy of Sciences*, 110(30):12241–12246, 2013.

[9] H. J. Bandelt, M. Van Oven, and A. Salas. Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *International Journal of Legal Medicine*, 126(6):901–916, 2012.

[10] A. Barbaro, P. Cormaci, and A. Barbaro. DNA analysis from mixed biological materials. *Forensic Science International*, 146, Suppl:S123–S125, Dec 2004.

[11] D. M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.-L. Loogväli, N. M. Silva, T. Kivisild, A. Torroni, and R. Villems. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics*, 90(4):675–684, Apr 2012.

[12] E. C. Berglund, A. Kiialainen, and A.-C. Syvänen. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative Genetics*, 2(1):23, 2011.

[13] M. Billinger and J. W. Ives. Inferring demographic structure with moccasin size data from the promontory caves, Utah. *American Journal of Physical Anthropology*, 156(1):76–89, 2015.

[14] D. Bornman, M. Hester, J. Schuetter, M. Kasoji, A. Minard-Smith, C. Barden, S. Nelson, G. Godbold, C. Baker, B. Yang, J. Walther, I. Tornes, P. Yan, B. Rodriguez, R. Bundschuh, M. Dickens, B. Young, and S. Faith. Short-read, high-throughput sequencing technology for STR genotyping. *BioTechniques - Rapid Dispatches*, pages 1–6, Apr 2012.

[15] C. Børsting and N. Morling. Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics*, 18:78–89, 2015.

[16] A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maricic, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, R. Schmitz, V. B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, and S. Pääbo. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*, 325(5938):318–321, Jul 2009.

[17] J. S. Buckleton, J. M. Curran, and P. Gill. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics*, 1(1):20–28, 2007.

[18] B. Budowle, B. Budowle, T. R. Moretti, T. R. Moretti, S. J. Niezgoda, S. J. Niezgoda, B. L. Brown, and B. L. Brown. CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools. *The Second European Symposium on Human Identification*, pages 73–88, 1998.

[19] B. Budowle and A. Van Daal. Forensically relevant SNP classes. *BioTechniques*, 44(5):603–610, 2008.

[20] J. M. Butler. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Elsevier Academic Press, Burlington, MA, second edition, 2005.

[21] C. Calabrese, D. Simone, M. A. Diroma, M. Santorsola, C. Gutta, G. Gasparre, E. Picardi, G. Pesole, and M. Attimonelli. MToolBox: A highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, 30(21):3115–3117, 2014.

[22] M. L. Carpenter, J. D. Buenrostro, C. Valdiosera, H. Schroeder, M. E. Allentoft, M. Sikora, M. Rasmussen, S. Gravel, S. Guillén, G. Nekhrizov, K. Leshtakov, D. Dimitrova, N. Theodossiev, D. Pettener, D. Luiselli, K. Sandoval, A. Moreno-Estrada, Y. Li, J. Wang, M. T. P. Gilbert, E. Willerslev, W. J. Greenleaf, and C. D. Bustamante. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American Journal of Human Genetics*, 93(5):852–864, Nov 2013.

[23] T. M. Clayton, J. P. Whitaker, R. Sparkes, and P. Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1):55–70, 1998.

[24] Y. Cui, L. Song, D. Wei, Y. Pang, N. Wang, C. Ning, C. Li, B. Feng, W. Tang, H. Li, Y. Ren, C. Zhang, Y. Huang, Y. Hu, and H. Zhou. Identification of kinship and occupant status in Mongolian noble burials of the Yuan Dynasty through a multidisciplinary approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130378, Dec 2014.

[25] J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, and M. Meyer. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, 110(39):15758–15763, Sep 2013.

[26] J. Dabney, M. Meyer, and S. Pääbo. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, 5(7):a012567, Jul 2013.

[27] M. Date Chong, C. D. Calloway, S. B. Klein, C. Orrego, and M. R. Buoncristiani. Optimization of a duplex amplification and sequencing strategy for the HVI/HVII regions of human mitochondrial DNA for forensic casework. *Forensic Science International*, 154(2-3):137–148, 2005.

[28] C. B. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8):897–899, 2008.

[29] European Molecular Biology Laboratory, European Bioinformatics Institute. European Nucleotide Archive. `https://www.ebi.ac.uk/ena/data/view/PRJEB9021`, 2015 (Accessed July 18, 2015).

[30] I. W. Evett, C. Buffery, G. Willott, and D. Stoney. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *Journal of the Forensic Science Society*, 31(1):41–47, Jan 1991.

[31] L. Fan and Y. G. Yao. MitoTool: A web server for the analysis and retrieval of human mitochondrial DNA sequence variations. *Mitochondrion*, 11(2):351–356, 2011.

[32] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, I. Yakub, B. W. Birren, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li,

Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, Oct 2007.

[33] A. Fungtammasan, G. Ananda, S. E. Hile, M. S.-W. Su, C. Sun, R. Harris, P. Medvedev, K. Eckert, and K. D. Makova. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research*, 25(5):736–749, May 2015.

[34] C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kvári, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T. F. G. Higham, M. Hofreiter, D. G. Bradley, and R. Pinhasi. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5:5257, Oct 2014.

[35] M.-T. Gansauge and M. Meyer. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols*, 8(4):737–48, 2013.

[36] M. T. P. Gilbert, T. Kivisild, B. Grønnow, P. K. Andersen, E. Metspalu, M. Reidla, E. Tamm, E. Axelsson, A. Götherström, P. F. Campos, M. Rasmussen, M. Metspalu, T. F. G. Higham, J.-L. Schwenninger, R. Nathan, C.-J. De Hoog, A. Koch, L. N. Møller, C. Andreasen, M. Meldgaard, R. Villems, C. Bendixen, and E. Willerslev. Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Science*, 320(5884):1787–1789, Jun 2008.

[37] P. Gill. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine*, 114(4-5):204–210, 2001.

[38] P. Gill, C. H. Brenner, J. S. Buckleton, A. Carracedo, M. Krawczak, W. R. Mayr, N. Morling, M. Prinz, P. M. Schneider, and B. S. Weir. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2-3):90–101, 2006.

[39] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91(1):41–53, 1998.

[40] A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, 27(2):182–189, 2009.

[41] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund,

A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, May 2010.

[42] R. E. Green, A.-S. Malaspinas, J. Krause, A. W. Briggs, P. L. F. Johnson, C. Uhler, M. Meyer, J. M. Good, T. Maricic, U. Stenzel, K. Prüfer, M. Siebauer, H. A. Burbano, M. Ronan, J. M. Rothberg, M. Egholm, P. Rudan, D. Brajković, Z. Kućan, I. Gusić, M. Wikström, L. Laakkonen, J. Kelso, M. Slatkin, and S. Pääbo. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–26, Aug 2008.

[43] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695, Oct 2009.

[44] M. Gymrek, D. Golan, S. Rosset, and Y. Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22(6):1154–1162, Jun 2012.

[45] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, Mar 2015.

[46] M. C. Hansson and B. P. Foley. Ancient DNA fragments inside Classical Greek amphoras reveal cargo of 2400-year-old shipwreck. *Journal of Archaeological Science*, 35(5):1169–1176, 2008.

[47] M. Hofreiter, D. Serre, H. N. Poinar, M. Kuch, and S. Pääbo. Ancient DNA. *Nature Reviews Genetics*, 2(5):353–359, May 2001.

[48] M. M. Holland, M. R. McQuillan, and K. A. O'Hanlon. Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. *Croatian Medical Journal*, 52(3):299–313, 2011.

[49] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, 2002.

[50] Illumina Inc. Platinum Genomes. `http://www.illumina.com/platinumgenomes/`, 2014 (Accessed Aug 29, 2014).

[51] J. W. Ives, D. G. Froese, J. C. Janetski, F. Brock, and C. B. Ramsey. A High Resolution Chronology for Steward's Promontory Culture Collections, Promontory Point, Utah. *American Antiquity*, 79(4):616–637, 2014.

[52] M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, Feb 2008.

[53] M. A. Jobling and P. Gill. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10):739–751, 2004.

[54] H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, and L. Orlando. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29(13):1682–1684, 2013.

[55] M. Kayser and P. de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192, 2011.

[56] C. Keyser, C. Bouakaze, E. Crubézy, V. G. Nikolaev, D. Montagnon, T. Reis, and B. Ludes. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Human Genetics*, 126(3):395–410, Sep 2009.

[57] H. Kim, H. A. Erlich, and C. D. Calloway. Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions. *Croatian Medical Journal*, 56(3):208–217, Jun 2015.

[58] H. Y. K. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder. Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, 30(6):562–562, 2012.

[59] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik,

F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, and J. Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, Sep 2014.

[60] S. A. LeBlanc, L. S. C. Kreisman, B. M. Kemp, F. E. Smiley, S. W. Carlyle, A. N. Dhody, and T. Benjamin. Quids and aprons: Ancient DNA from artifacts from the American southwest. *Journal of Field Archaeology*, 32(2):161–175, 2007.

[61] B. Letts and B. Shapiro. Case study: ancient DNA recovered from pleistocene-age remains of a Florida armadillo. In B. Shapiro and M. Hofreiter, editors, *Ancient DNA: Methods and Protocols, Methods in Molecular Biology*, volume 840, chapter 12, pages 87–92. Clifton, NJ, 2012.

[62] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint*, 1303.3997, Mar 2013. `http://arxiv.org/abs/1303.3997`.

[63] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–95, Mar 2010.

[64] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[65] E. R. Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387–402, Jan 2008.

[66] T. Melton, G. Dimick, B. Higgins, L. Lindstrom, and K. Nelson. Forensic mitochondrial DNA analysis of 691 casework hairs. *Journal of Forensic Sciences*, 50(1):73–80, 2005.

[67] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226, Oct 2012.

[68] M. Montesino, A. Salas, M. Crespillo, C. Albarrán, A. Alonso, V. Álvarez-Iglesias, J. A. Cano, M. Carvalho, D. Corach, C. Cruz, A. Di Lonardo, R. Espinheira, M. J. Farfán, S. Filippini, J. García-Hirschfeld, A. Hernández, G. Lima, C. M. López-Cubría, M. López-Soto, S. Pagano, M. Paredes, M. F. Pinheiro, A. M. Rodríguez-Monge, A. Sala,

S. Sóñora, D. R. Sumita, M. C. Vide, M. R. Whittle, A. Zurita, and L. Prieto. Analysis of body fluid mixtures by mtDNA sequencing: An inter-laboratory study of the GEP-ISFG working group. *Forensic Science International*, 168(1):42–56, 2007.

[69] J. J. Mulero, C. W. Chang, R. E. Lagacé, D. Y. Wang, J. L. Bas, T. P. McMahon, and L. K. Hennessy. Development and validation of the AmpFSTR® MiniFiler PCR amplification kit: A miniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA. *Journal of Forensic Sciences*, 53(4):838–852, 2008.

[70] K. B. Mullis and F. A. Faloona. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, 155(C):335–350, 1987.

[71] National Center for Biotechnology Information, National Library of Medicine, Bethesda (MD). Sequence Read Archive. `http://www.ncbi.nlm.nih.gov/sra/`, 2013 (accessed Aug 29, 2014).

[72] National Center for Biotechnology Information, National Library of Medicine, Bethesda (MD). Database of Single Nucleotide Polymorphisms (dbSNP) (dbSNP Build ID: 144). `http://www.ncbi.nlm.nih.gov/SNP`, 2015 (accessed Jul 27, 2015).

[73] D. Navarro-Gomez, J. Leipzig, L. Shen, M. Lott, A. P. M. Stassen, D. C. Wallace, J. L. Wiggs, M. J. Falk, M. Van Oven, and X. Gai. Phy-Mer: A novel alignment-free and reference-independent mitochondrial haplogroup classifier. *Bioinformatics*, 31(8):1310–1312, 2015.

[74] S. Pääbo. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, 86(6):1939–1943, Mar 1989.

[75] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S. M. Gomes, L. Souto, L. Fendt, R. Delport, R. Langit, S. Wootton, R. Lagacé, and J. Irwin. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Science International: Genetics*, 7(6):632–639, Sep 2013.

[76] B. A. I. Payne, I. J. Wilson, P. Yu-Wai-Man, J. Coxhead, D. Deehan, R. Horvath, R. W. Taylor, D. C. Samuels, M. Santibanez-Koref, and P. F. Chinnery. Universal heteroplasmy of human mitochondrial DNA. *Human Molecular Genetics*, 22(2):384–390, 2013.

[77] G. H. Perry, J. C. Marioni, P. Melsted, and Y. Gilad. Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, 19(24):5332–5344, Dec 2010.

[78] J. K. Pickrell and D. Reich. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics*, 30(9):377–389, Aug 2014.

[79] H. N. Poinar, C. Schwarz, J. Qi, B. Shapiro, R. D. E. Macphee, B. Buigues, A. Tikhonov, D. H. Huson, L. P. Tomsho, A. Auch, M. Rampp, W. Miller, and S. C. Schuster.

Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311(5759):392–394, Jan 2006.

[80] M. Prinz, A. Ishii, A. Coleman, H. J. Baum, and R. C. Shaler. Validation and casework application of a Y chromosome specific STR multiplex. *Forensic Science International*, 120(3):177–188, 2001.

[81] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–9, Jan 2014.

[82] G. Renaud, V. Slon, A. T. Duggan, and J. Kelso. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, 16(1):224, 2015.

[83] C. Renfrew and P. G. Bahn. *Archaeology: Theories, Methods, and Practice*. Thames and Hudson, 1996.

[84] R. Reynolds, K. Walker, J. Varlaro, M. Allen, E. Clark, M. Alavaren, and H. Erlich. Detection of sequence variation in the HVII region of the human mitochondrial genome in 689 individuals using immobilized sequence-specific oligonucleotide probes. *Journal of Forensic Sciences*, 45(6):1210–1231, Nov 2000.

[85] K. K. Richter. *DNA Preservation Under Extreme Conditions*. Doctoral dissertation, The Pennslyvania State University, 2014.

[86] A. W. Röck, A. Dür, M. van Oven, and W. Parson. Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Science International: Genetics*, 7(6):601–609, 2013.

[87] J. M. Rodriguez, S. Bercovici, L. Huang, R. Frostig, and S. Batzoglou. Parente2: a fast and accurate method for detecting identity by descent. *Genome Research*, 25(2):280–289, Feb 2015.

[88] L. Roewer. DNA fingerprinting in forensics: past, present, future. *Investigative Genetics*, 4(1):22, 2013.

[89] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. A. Smit, M. L. Speir, A. S. Zweig, D. Haussler,

R. M. Kuhn, and W. J. Kent. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, 43(D1):D670–D681, 2014.

[90] R. Saiki, S. Scharf, F. Faloona, K. Mullis, G. Horn, H. Erlich, and N. Arnheim. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–1354, Dec 1985.

[91] R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

[92] T. Schultes, S. Hummel, and B. Herrmann. Ancient DNA-typing Approaches for the Determination of Kinship in a Disturbed Collective Burial Site. *Anthropologischer Anzeiger*, 58(1):37–44, Mar 2000.

[93] S. B. Seo, J. L. King, D. H. Warshauer, C. P. Davis, J. Ge, and B. Budowle. Single nucleotide polymorphism typing with massively parallel sequencing for human identification. *International Journal of Legal Medicine*, 127(6):1079–1086, 2013.

[94] B. Shapiro and M. Hofreiter. A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science*, 343(6169):1236573, Jan 2014.

[95] P. Skoglund, H. Malmstrom, M. Raghavan, J. Stora, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Gotherstrom, and M. Jakobsson. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science*, 336(6080):466–469, Apr 2012.

[96] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234, Feb 2014.

[97] M. Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–85, Jun 2008.

[98] M. Slatkin and F. Racimo. Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, 113(23):6380–7, Jun 2016.

[99] C. I. Smith, A. T. Chamberlain, M. S. Riley, C. Stringer, and M. J. Collins. The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution*, 45(3):203–217, 2003.

[100] J. E. L. Templeton, P. M. Brotherton, B. Llamas, J. Soubrier, W. Haak, A. Cooper, and J. J. Austin. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative Genetics*, 4(1):26, Dec 2013.

[101] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Oct 2012.

[102] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015.

[103] P. F. Thomsen and E. Willerslev. Environmental DNA An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, 183:4–18, Mar 2015.

[104] M. D. Timken, K. L. Swango, C. Orrego, and M. R. Buoncristiani. A duplex real-time qPCR assay for the quantification of human nuclear and mitochondrial DNA in forensic samples: implications for quantifying DNA in degraded samples. *Journal of Forensic Sciences*, 50(5):1044–1060, Sep 2005.

[105] A. Torroni, A. Achilli, V. Macaulay, M. Richards, and H. J. Bandelt. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*, 22(6):339–345, 2006.

[106] P. A. Underhill and T. Kivisild. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*, 41:539–564, 2007.

[107] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30, 2011.

[108] R. A. van Oorschot, K. N. Ballantyne, and R. J. Mitchell. Forensic trace DNA: a review. *Investigative Genetics*, 1(1):14, 2010.

[109] M. van Oven and M. Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2):386–394, 2009.

[110] D. Vianello, F. Sevini, G. Castellani, L. Lomartire, M. Capri, and C. Franceschi. HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. *Human Mutation*, 34(9):1189–1194, 2013.

[111] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. African populations and the evolution of human mitochondrial DNA. *Science*, 253(5027):1503–1507, 1991.

[112] S. H. Vohr, C. F. Buen Abad Najar, B. Shapiro, and R. E. Green. A method for positive forensic identification of samples from extremely low-coverage sequence data. *BMC Genomics*, 16(1):1034, Dec 2015.

[113] J. A. Walker, R. K. Garber, D. J. Hedges, G. E. Kilroy, J. Xing, and M. A. Batzer. Resolution of mixed human DNA samples using mitochondrial DNA sequence variants. *Analytical Biochemistry*, 325(1):171–173, 2004.

[114] D. H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, J. L. King, and B. Budowle. STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Science International: Genetics*, 7(4):409–417, 2013.

[115] B. S. Weir, C. M. Triggs, L. Starling, L. I. Stowell, K. A. Walsh, and J. Buckleton. Interpreting DNA mixtures. *Journal of Forensic Sciences*, 42(2):213–22, Mar 1997.

[116] H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, and S. Schönherr. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research*, 44(W1):W58–W63, Jul 2016.

[117] M. R. Wilson, J. A. DiZinno, D. Polanskey, J. Replogle, and B. Budowle. Validation of mitochondrial DNA sequencing for forensic casework analysis. *International Journal of Legal Medicine*, 108(2):68–74, 1995.

[118] E. Zietkiewicz, M. Witt, P. Daca, J. Zebracka-Gala, M. Goniewicz, B. Jarzab, and M. Witt. Current genetic methodologies in the identification of disaster victims and in forensic analysis. *Journal of Applied Genetics*, 53(1):41–60, 2012.