GENETIC RELATEDNESS ESTIMATION USING ANCIENT GENOMIC DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

AYSHIN GHALICHI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

SEPTEMBER 2017

Approval of the thesis:

GENETIC RELATEDNESS ESTIMATION USING ANCIENT GENOMIC DATA

submitted by **AYSHIN GHALICHI** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Orhan Adalı
Head of Department, **Biology** _____

Assoc. Prof. Dr. Mehmet Somel
Supervisor, **Biology Dept., METU** _____

Examining Committee Members:

Prof. Dr. Mayda Gürsel
Biology Dept., METU _____

Assoc. Prof. Dr. Mehmet Somel
Biology Dept., METU _____

Assoc. Prof. Dr. Sreeparna Banerjee
Biology Dept., METU _____

Assist. Prof. Dr. Nihal Terzi Çizmecioğlu
Biology Dept., METU _____

Assist. Prof. Dr. Banu Şebnem Önder
Biology Dept., Hacettepe University _____

Date: 15.09.2017

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Ayshin Ghalichi

Signature :

ABSTRACT

GENETIC RELATEDNESS ESTIMATION USING ANCIENT GENOMIC DATA

Ghalichi, Ayshin
M.S., Department of Biology
Supervisor : Assoc. Prof. Dr. Mehmet Somel

September 2017, 89 pages

One distinct feature of the early Neolithic settlements in the Near East was their burial customs. Both in the Levant and in Anatolia, people dug graves inside their houses, and multiple individuals were buried in these intramural graves; a custom that reached its climax in Çatalhöyük. Archaeological evidence suggests that individuals buried in a house were socially related, which has motivated anthropologists to estimate biological relatedness among individuals who share the same grave. Such information, which could be obtained from ancient DNA data, could shed light on the social structure of these ancient communities, and be valuable for archaeological studies. The challenge of working with ancient DNA is that it is highly degraded and usually in minute amounts, which results in limited DNA data availability. Importantly, in ancient DNA datasets usually only one allele can be detected per individual. There exist a number of methods to estimate genetic relatedness designed for modern high coverage genomic data, but their performance on ancient DNA data has not been tested. Here we

apply two of these methods, KING and PLINK, on low coverage whole genome data from real family pedigrees, as well as ancient DNA data from simulated pedigrees. We further propose a new approach to calculate relatedness between ancient individuals, which would require minimal coverage and SNP numbers to accurately estimate relatedness. We show that our approach can more efficiently estimate the relatedness coefficients compared to the KING and PLINK software. Our approach is expected to promote the application of ancient DNA to address new archaeological questions.

Keywords: Relatedness, Kinship coefficient, Coancestry, Ancient DNA, Identical by descent (IBD)

ÖZ

ANTİK GENOM DİZİLEME VERİSİ İLE GENETİK AKRABALIK İLİŞKİSİNİN BELİRLENMESİ

Ghalichi, Ayshin
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi : Doç. Dr. Mehmet Somel

Eylül 2017 , 89 sayfa

Yakın Doğu'da yer alan Neolitik yerleşim yerlerinin en belirgin özelliklerinden biri ölü gömme gelenekleridir. Hem Levant'ta hem de Anadolu'da, ölülerin evlerin içine gömüldüğü bilinmektedir. Ev içinde bulunan mezarlara birden fazla ölünün gömülmesi şeklinde görülen bu gelenek, Çatalhöyük'te çok yaygındır. Arkeolojik bulgular, aynı evin içine gömülen bireylerin sosyal olarak birbirleri ile ilişkili olabileceklerini göstermektedir. Bu doğrultuda, aynı mezarda bulunan bireyler arasındaki biyolojik akrabalık derecesinin bilinmesi de antropologlar için ilgi çekici sorulardan biridir. Antik DNA'dan elde edilmesi mümkün olan bu bilgi, arkeolojik araştırmalar için de çok önemli olan antik yerleşim yerlerindeki toplulukların sosyal yapıları ile ilgili önemli bilgiler sağlayabilir. Antik DNA ile çalışmanın en önemli zorluğu ise yüksek oranda parçalanmış olması ve çok az miktarda elde edilebiliyor olması sebebiyle sınırlı DNA dizi bilgisi sağlamasıdır. Antik DNA ile üretilen dizileme verilerinde, dizilenen pozisyonlar için her bireyde

yalnızca bir alelin bilgisi elde edilebilmektedir. Günümüzde, modern genom dizileme verisi ile biyolojik akrabalık derecesini belirlemeye olanak sağlayan metotlar olmasına rağmen, bu metotların antik DNA verisi ile etkin şekilde kullanılıp kullanılamayacağı henüz test edilmemiştir. Bu çalışmada, modern genomlar için dizayn edilmiş bu yöntemlerden ikisini, KING ve PLINK'i, gerçek ve simüle edilmiş aile ağaçlarında, düşük kapsamda (derinlikte) üretilmiş tüm genom dizileme verisi kullanarak test ettik. Ayrıca, minimum sayıda tekil nukleotid polimorfizmi (TNP) ve antik DNA dizileme derinliğine sahip veri ile doğru akrabalık katsayısı hesaplamak için farklı bir yöntem geliştirdik. Yönteminizin KING ve PLINK'ten daha verimli sonuç verdiğini belirledik. Bulgularımız, antik DNA dizileme verilerinin yeni arkeolojik soruların cevaplanmasındaki kullanımını kolaylaştıracaktır.

Anahtar Kelimeler: Akrabalık, Akrabalık katsayısı, Ortak atadan gelme, Antik DNA, Türeme yoluyla özdeş

To my beloved family

تقدیم به خانواده عزیزم

ACKNOWLEDGMENTS

First and foremost, I would like to express the special appreciation to my supervisor *Mehmet Somel*, for the opportunity of doing research with his team. I have greatly benefited from his immense knowledge, valuable guidance, generous support and continuous enthusiasm.

I want to express my deepest gratitude to *Gülşah Dal Kılınç* for her guidance and friendship. Without her encouragement, useful feedback and insightful comments I would never have been able to finish my thesis.

My sincere thanks to *Melike Dönertaş* and *Dilek Koptekin* for providing much needed assistance and for tirelessly answering all my questions. I also want to thank *Hazal Moğultay* and *Poorya Parvizi* for helping me with the creation of (LaTeX) template and any formating problems I had.

I must thank the many members of Compevo (Somel) lab for being the Sam to my Frodo, without their friendship and endless support, I could not have completed this challenging journey.

Also I should thank *Howard Shore*, for his amazing score for LOTRs. I would not have finished my thesis without this wonderful and epic score which kept me motivated. I would like to acknowledge *J.R.R Tolkien*, *Bill Finger* and *Bob Kane* for the fascinating characters and stories they created. Their work inspired me to be more courageous, work harder and never give up no matter how hard the situation gets.

I thank Scientific and Technological Research Council of Turkey (TÜBİTAK) for financially supporting me through 3501 project scholarship with "114Z927" code and "Niğde Tepecik-Çiftlik ve Konya Çatalhöyük kazılarına ait insan kemiklerinden antik DNA eldesi ve genom dizilemesi yoluyla Orta Anadolu Neolitik Çağ insan popülasyonlarının karakterizasyonu" title.

Lastly, I would like to thank my family: my dear father *Nader*, for his infinite belief and trust in me, my amazing aunt *Narges*, for her constant love and support through the most difficult times of my life and my awesome brother *Elyar*, for always being there for me. I love you so much.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Genetic relatedness estimation

One of the first fields to use information about family history and their lineages was genealogy. The information gathered about relatedness, or more precisely the degree by which people are related to each other was used to regulate laws about marriage and inheritance (Bishop, 2008; Weir, Anderson, & Hepler, 2006). Throughout the years, the genetic relatedness concept was used in fields such as agriculture, forensics, human genetics, conservation programs, and most recently, archaeology (Monroy Kuhn, Jakobsson, & Günther, 2017; Weir et al., 2006).

Although relatedness analysis is a primary concept in medical genetics, defining it has not been an easy task (Speed & Balding, 2014). In general, members of a family or a population are said to be related because they share a common ancestor (Weir et al., 2006). These related individuals share segments of their DNA that are identical by descent (IBD), meaning that they were inherited from a common ancestor (**Figure 1.1**). Actually each individual is a mix of many of these segments that come from different ancestors that shape human genome into a mosaic (Monroy Kuhn et al., 2017). However IBD is a quantity that cannot be measured directly from data, what is done is using the property of being "identical by state" to make an inference about IBD (Theunert, Racimo, & Slatkin, 2017).

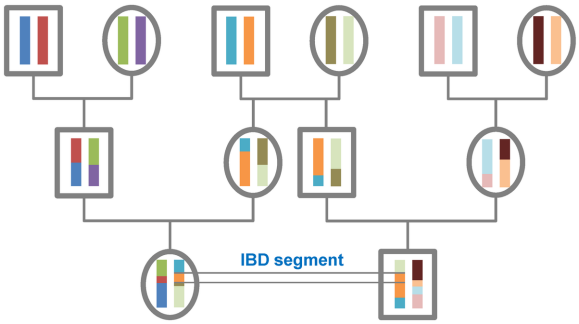

Figure 1.1: IBD segments. Adapted from the Wikimedia Commons file "Pedigree, recombination and resulting IBD segments, schematic representation.png"

There are two types of relatedness measures: **i)** pedigree-based, **ii)** marker based.

1.1.1 Pedigree-based relatedness

Traditionally, pedigrees have been used in the formation and explanation of joint frequencies and covariance of genetic markers between relatives (Cockerham, 1971; Harris, 1964; Wright, 1921). Pedigrees and their lineage paths, the shortest number of (parent-child) steps that would take to link two individuals together in a pedigree, had been used to calculate theoretical pairwise relatedness and inbreeding coefficient (Wang, 2016; Weir et al., 2006).
As **Figure 1.2** shows, the half-siblings B and C are connected together via their most recent common ancestor, which is A. The kinship coefficient (coancestry coefficient) Θ between two individuals (here between B and C), probability of a randomly chosen allele from individual B and individual C being IBD, meaning coming from their most recent common ancestor individual A. One way that the pairwise kinship coefficient can be calculated is by,

$$\Theta(B,C) = \sum_{A} \frac{1+f_A}{2^{g_A+1}} \tag{1.1}$$

Where f_A represents the inbreeding coefficient for individual A and g_A is the

number of lineage path that connects B and C through individual A. This is summed over all possible most recent common ancestors between B and C (e.g. there would be two such individuals for cousins).

As Harris (1964) described there are 15 possible patterns of IBD, if a single locus with two alleles is considered for each individual B and C. With the assumptions of unrelated ancestors (f_B=f_C=0) and unordered alleles, the patterns could be reduced to three (IBD = 0, 1 or 2). When a single locus is compared between two individuals, they can share two alleles that came from their common ancestor(s) (IBD=2), or only one allele (IBD=1) or none (IBD=0).

Considering these, the formula can be revised to,

$$\Theta = \frac{E[IBD]}{4} = \frac{\phi}{4} + \frac{\Delta}{2} \tag{1.2}$$

Where $\phi = P[IBD = 1]$ and $\Delta = P[IBD = 2]$. This is because, if there is one IBD allele between B and C, the chance of randomly choosing that allele in both B and C is $\frac{1}{2}$, and zero for the case of two IBD alleles between B and C.

Although using pedigrees are crucial for calculations of expected kinship degrees, they don't always represent real-life situations completely. Most of the time, it is impossible to reconstruct the whole pedigree. In addition, the assumption that the founders of the pedigree are unrelated (f=0) is almost never valid for real-life models because there is always some level of above-random relationship among members of a population. Besides, addition of extra ancestors to the original pedigree would increase the coancestry value to the point that it converges to one and make the whole pedigree redundant in practice (Speed & Balding, 2014).

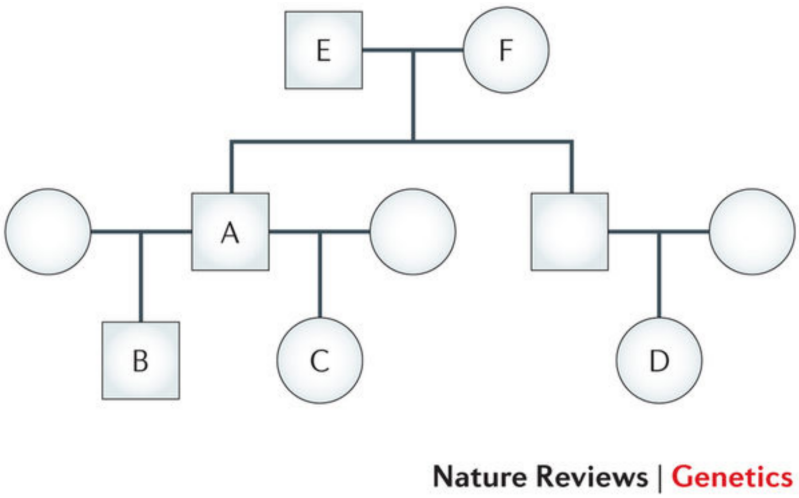

Figure 1.2: Using pedigree to define and calculate relatedness. Adapted from (Speed & Balding, 2014).

1.1.2 Marker-based relatedness

In addition to the problems listed above (**Section 1.1.1**), obtaining pedigrees in real-life is very difficult, time consuming, laborious and for most cases impossible. For this reason, many researchers use molecular markers.

One of the first forms of relatedness studies was paternity tests that became accessible in the 1920s. Formerly, these studies used blood-group antigens (ABO, Rh and MNS) to determine whether there was a parental relationship between two individuals. However, these markers only represent very small number of alleles and their dominance characteristics over other alleles made the relatedness estimations inconclusive. Although a breakthrough occurred with the discovery of minisatellites by Jeffreys et al. in 1985, technical problems such as selection of gel systems, insufficient lab control and ambiguity in interpretation of matches between electrophoretic bands, made the implementation of this method in legal context debatable (Ramel, 1997; Weir et al., 2006).

Today, using microsatellites and single nucleotide polymorphisms (SNPs) as markers for analysis of relatedness generates more detailed and accurate results than using the aforementioned markers, because microsatellites are well-studied, and advances in next generation sequencing (NGS) technologies has made high-density SNP data readily available.

1.1.2.1 Microsatellites

Microsatellites are a class of VNTR (variable number of tandem repeats) that consists of repetitive DNA segments of 2-5 nucleotides, which are common in multicellular organisms. The number of times these segments are repeated along the genome is highly variable within a population. Since 1990s, microsatellites replaced minisatellites as molecular markers used in fields of forensic science and paternity testing.

Using microsatellites as molecular markers is advantageous because they are multiallelic, locus-specific and codominant, meaning alleles do not mask each other. Microsatellites are very abundant (around 32,000 microsatellite markers were identified for all members of CEPH families) and show noticeable difference among and within populations, which is why they have been used in many linkage studies and demographic analyses (Chistiakov, Hellemans, & Volckaert, 2006; Ramel, 1997; Weir et al., 2006). Meanwhile, some features of microsatellites, such as their high mutation rate and instability makes their use in some genetic analysis problematic. Because of their high mutation rate, a reliable conclusion cannot be made about IBS (identical by state) alleles being IBD (identical by descent) or not. As a consequence of these limitations, many association studies and population genetic analysis uses SNPs as their genetic marker that are much more common (Weir et al., 2006).

1.1.2.2 Biallelic SNPs

Through the recent developments in Next Generation Sequencing (NGS) techniques, it is possible to sequence several samples cheaply in a very short time. The reduction in time and cost enables sequencing of thousands of individuals for studies of evolutionary biology, clinical genetics, forensics and metagenomics (Behjati & Tarpey, 2013). Even if in comparisons individual SNPs are much less informative than individual microsatellites, the profound abundance, greater genetic stability (in mammals), straightforward terminology and convenient data analysis tools available for SNPs make them a more efficient and reliable data source than microsatellites (Fernández et al., 2013; Pemberton, 2008).

SNP data has been successfully adapted to many individual identification and parentage analysis in animal breeds and conservation programs (Heaton et al., 2002; Rohrer, Freking, & Nonneman, 2007; Tokarska et al., 2009), as well as in forensic science (Phillips et al., 2007) and inferring kinship degrees in humans (Lipatov, Sanjeev, Patro, & Veeramah, 2015; Manichaikul et al., 2010; Purcell et al., 2007; Speed & Balding, 2014; Weir et al., 2006).

Several equations, methods and software have been developed for relatedness estimations in human genetics, plant and animal breeding which take into account different sets of assumptions for each case. One way of estimating pairwise relatedness is to use genetic relatedness matrices (GRM) or as Speed and Balding (2014) suggests, genetic similarity matrices (GSM) to avoid confusion with pedigree-based relatedness methods (Speed & Balding, 2014). Using GSM methods makes the relatedness estimation computations more efficient (Dodds et al., 2015). The softwares developed by Patterson et al. 2006 (PCA) and Pickrell and Pritchard 2012 (TreeMix) use a similar approach.

In order to generate the similarity matrix, the biallelic SNPs between two individuals are coded as 0, 1 and 2 according to the state of minor allele in a population. If the individual's alleles are homozygous for minor allele they are coded as 2, heterozygotes are coded as 1 and homozygous for major allele are coded as 0 (Speed & Balding, 2014).

Speed and Balding (2014) describe four different SNP-based measures to estimate kinship coefficient Θ which are K_{as}, $k_{as}{}'$, K_{c0} and K_{c-1}. These are all based on counting the number of IBS SNPs but differ in how they weight each SNP. Generally the K_{c-1} is used in the field of human genetics while the K_{c0} is mostly favored in plant and animal breeding. The former assumes that the phenotypic variance is explained equally by all SNPs and therefore as the minor allele frequency (MAF) decreases, their effect size would increase, whereas the latter assumes all SNPs have similar effect size (Speed & Balding, 2014). Nevertheless it is better to use the most informative set of SNPs in regards to the traits in question which requires a well contemplated SNP filtering process.

1.2 Ancient DNA

Although archeogenomics enables study of archaic individuals and events, strict criteria is required for obtaining reliable results. Fragmented DNA, contamination and other post-mortem DNA decay are some of the technical drawbacks of working with ancient DNA (aDNA). However, many advances in the fields of NGS and accompanying bioinformatics allow using this technique to study historic human migration and biological structure (Hofreiter, Serre, Poinar, Kuch, & Pääbo, 2001; Monroy Kuhn et al., 2017; Shapiro & Hofreiter, 2014).

After the death of an organism, the DNA degradation process begins with activity of enzymes such as endogenous nucleases. Other environmental factors contributing to DNA degradation are temperature, humidity, degree of microbial attack, salt and pH concentration (Allentoft et al., 2012; Hofreiter et al., 2001). Allentoft et al. (2012) estimated the half life of DNA to be 521 years for a 242 bp mtDNA sequence. DNA degradation and high content of microbial DNA results in extracting low percentage of authentic (endogenous) ancient DNA. Even with advances in experimental aDNA techniques, in most cases only 0.1%-0.01% of the extracted DNA belongs to the targeted ancient DNA (**Figure 1.3**) (Shapiro & Hofreiter, 2014).

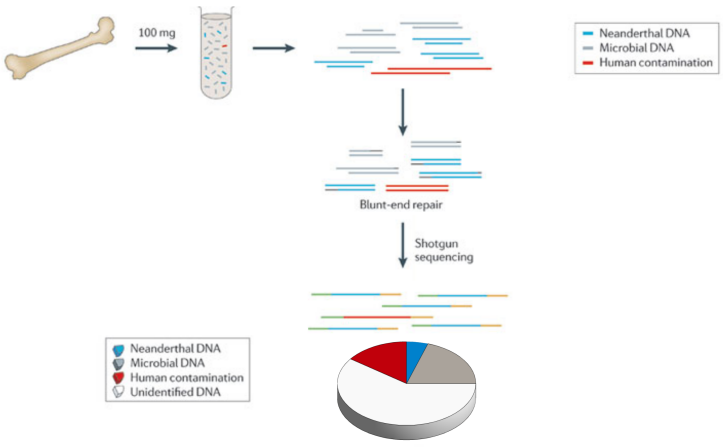

Figure 1.3: Challenges of sequencing ancient DNA. Modified from (Stoneking & Krause, 2011).

One distinct feature of aDNA is the pattern of cytosine to thymine substitution at the 5' end of fragmented sequence reads (**Figure 1.4**). This pattern is accompanied by a complementary substitution from G to A in the 3' end of the sequence that accumulates overtime (Skoglund et al., 2014).

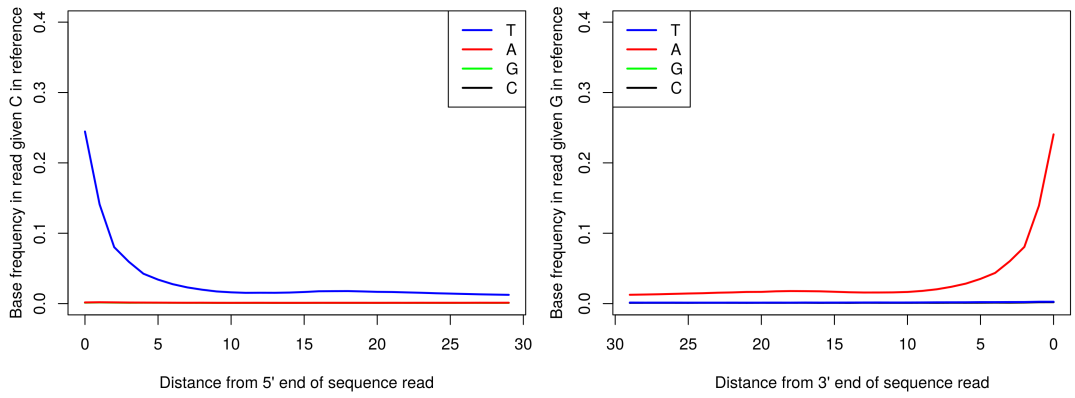

Figure 1.4: Patterns of post-mortem decay in ancient DNA (Kılınç et al., 2016).Plots of the positions' specific substitutions from the 5' (left) and the 3' end (right). The blue line shows the C to T substitutions while the red line shows the G to A substitutions.

Over the years many special experimental techniques and bioinformatic tools were generated to overcome these drawbacks in aDNA studies. These developments included the whole-genome in-solution capture (WISC) or SNP capture methods to increase the yield of endogenous aDNA extraction (Carpenter et al., 2013; Haak et al., 2015), methods for identification and removal of DNA contamination in ancient samples (Green et al., 2010; Skoglund et al., 2014) and computational tools for population genetic analysis of aDNA data (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013; Green et al., 2010; Pickrell & Pritchard, 2012).

1.3 Ancient DNA studies of the Neolithic period

The Neolithic Era is one of the most important turning points in human history

that started for the first time in the Fertile Crescent (including parts of Taurus-Zagros, Levant and Central Anatolia) around 12,000-11,000 years ago (**Figure 1.5**) (Belfer-Cohen & Goring-Morris, 2011; Byrd, 2005). The transition from mobile foraging to sedentary farming during this revolutionary period had dramatic consequences for human health, workloads, population growth rate, labor division and overall social structure (Bar-Yosef, 2001; Byrd, 2005; Larsen, 1995).

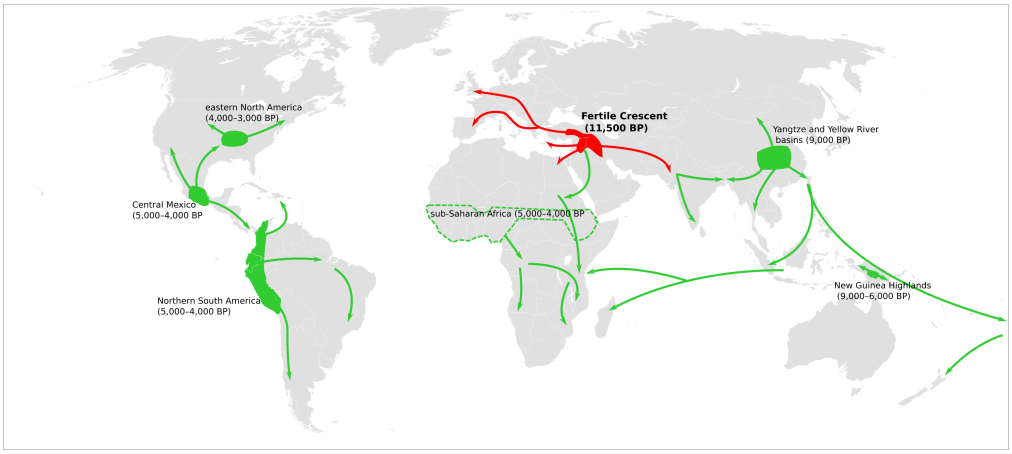

Figure 1.5: Centers of origin and expansion of agriculture across the world. Modified from (Diamond & Bellwood, 2003)

Another peculiar change was burial custom of these ancient cultures. The new sedentary communities had begun to bury their dead in organized fashion (Byrd, 2005). Even a more specialized version of these customs was practiced in the Levant and Central Anatolia, and particularly intensely in Çatalhöyük, where the dead were buried together beneath the floor of settlement (Hodder, 2007). The stable isotope analysis revealed the correlation between differentiated social activities (such as diet) and burial practices (Pearson, Grove, Özbek, & Hongo, 2013). Archaeological and anthropological studies indicate that these individuals buried in the same house were socially related. However, whether or not these individuals were also biologically related has remained a largely open question. There have been some studies attempting to answer this by using indirect methodologies such as comparison of dental morphology between individuals of a settlement (Pilloud & Larsen, 2011). Anthropological studies suggest a structured social organization due to the difference in dental phenotype and diet of

these ancient individuals. Nevertheless, these limited studies are not sufficient to derive a definite conclusion about the biological kinship and social structure of ancient populations.
Most ancient DNA studies focus on the Neolithization process and its spread through Eurasia. Other ancient studies focus on human migration and how each culture influence other populations. Not many genetic studies have yet focused on in-depth analysis of social organization in each settlement in itself. Combining archaeological studies with genetic analysis of social relations of these Neolithic (and other ancient) cultures could help us better understand the effects of cultural shifts to human populations and their demographic dynamics.

1.4 Research objectives

The aim of this study is to propose a new approach to infer kinship relations among ancient individuals. Ancient samples are degraded and have low coverage, hence the published software cannot accurately or efficiently estimate kinship degrees. Knowing kinship degrees between ancient individuals could help answer the long-lasting questions about the social organization of ancient human cultures. Archeological and anthropological studies can identify ancient human individuals who were socially related. However, whether or not these individuals were biologically related remains a mystery.
To examine the performance of our approach in different conditions, I used an ancient dataset for construction of a simple four-generation pedigree and also a modern family dataset because of its realistic error structure.
The objectives are:

- To test a different approach in estimation of kinship coefficients among ancient individuals which is not currently used in relatedness estimation software,
- To accurately estimate first and second degree (grandparent-grandchild, half-siblings) relatives that construct a core family,

- To characterize the lowest possible SNP number required for precise estimation of relatedness,
- To identify the potential error factors that decrease the accuracy and precision of our kinship coefficient estimations.

CHAPTER 2

MATERIALS AND METHODS

2.1 Study samples

Three different genomic datasets were used in this study. These datasets include **i)** Whole genome sequence data of CEPH Family 1463 (Eberle et al., 2017), **ii)** Genotype data for 594,924 autosomal SNPs of a total of 2,730 individuals from 203 modern populations from the Human Origins genotype dataset (Lazaridis et al., 2016; Patterson et al., 2012) and, **iii)** Genotype data for 1,240,000 autosomal SNPs captured in 230 ancient individuals (Mathieson et al., 2015).

2.1.1 Whole genome sequence data from CEPH Family 1463

Whole genome sequence data of the CEPH Family 1463 (Eberle et al., 2017) consisting of 17 individuals' full genome sequences were used in this study (**Figure 2.1**). Centre d'Etude du Polymorphisme Humain (Human polymorphism study center) CEPH, is a Paris based international genetic research center that has sequenced 61 reference families thus far. The family collection includes samples from France, Utah (North & Central European descent), Venezuela, and the Amish populations. One of these family sets is the CEPH pedigree 1463 that is sequenced to 50X depth.

Genome sequences of the five individuals from CEPH Family 1463 (NA12877, NA12883, NA12885, NA12889 and NA12890) that are mapped to the human reference genome (version hg19) and available as BAM (Binary sequence Alignment Map) files were downloaded from the European Nucleotide Archive (`https://`

www.ebi.ac.uk/ena ENA accession: PRJEB3381), which is also currently available in the NCBI dbGAP (database of Genotypes and Phenotypes; https://www.ncbi.nlm.nih.gov/gap; Study Accession: phs001224.v1.p1). These selected individuals represent a pair of unrelated individuals, five pairs of first-degree relatives and four pairs of second-degree relatives (**Figure 2.1**) (**Table 2.1**).

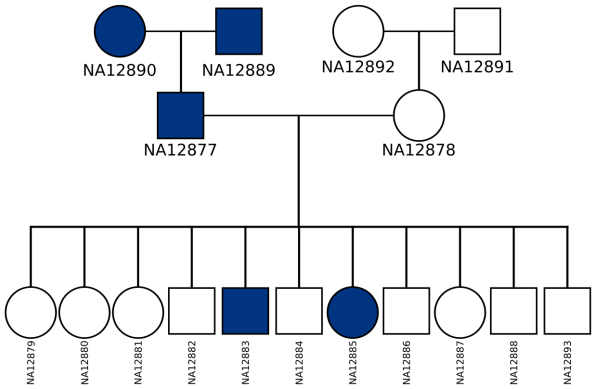

Figure 2.1: Complete pedigree of CEPH family 1463. The samples with blue color (five) are the individuals used for this study.

Table 2.1: Summary of CEPH family 1463 data

Sample	Coverage	Relation to Proband	Gender
NA12877	50	father	male
NA12883	50	son	male
NA12885	50	daughter	female
NA12889	50	Paternal grandfather	male
NA12890	50	Paternal grandmother	female

2.1.2 Genotype data of modern-day individuals from Human Origins dataset

The latest version of Human Origins (HO) SNP Array dataset, which contains 594,924 autosomal SNPs' genotype calls for 203 different populations (**Figure**

2.2) with 2,730 present-day individuals, was used in this study. I only used 50 West Eurasian populations for the PCA analysis in **Section 2.2.3**. This dataset was generated by the David Reich group (Lazaridis et al., 2016; Patterson et al., 2012) with the objective of facilitating human population history analysis. I downloaded the data from `http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets_files/NearEastPublic.tar.gz` in EIGENSTRAT format and converted them to PED file format via EIGENSOFT.

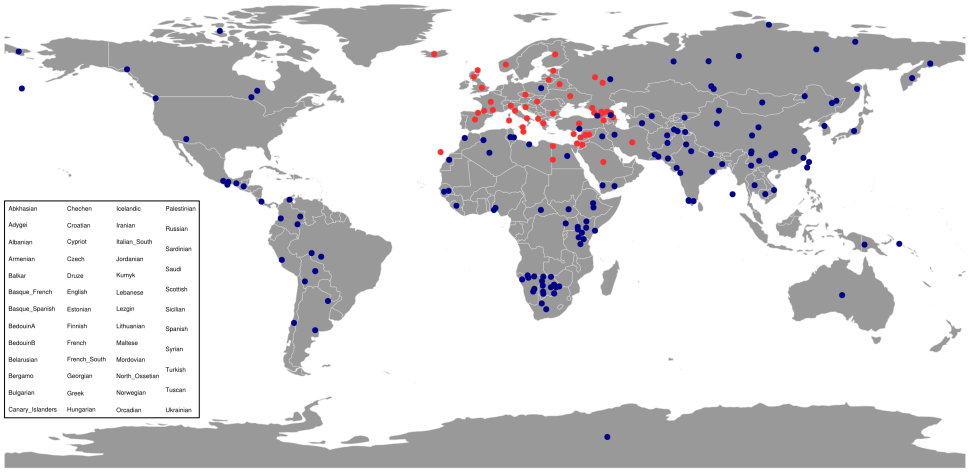

Figure 2.2: Geographic location of populations in HO dataset. This map shows the location of all the 203 populations available in Human Origins (HO) dataset. The red color represents the 50 West Eurasian population used in PCA analysis.

2.1.3 Genotype data of ancient individuals

The genome-wide data of 163 West Eurasian ancient individuals published by Mathieson et al. (2015) were used for studying relatedness in ancient populations in this thesis. The reason for selecting these samples is their high SNP coverage that is crucial for our analysis. The samples represent individuals from six material cultures and 19 distinct populations (**Table 2.2**, Mathieson et al. 2015), spanning a time period of 6500 to 300 BCE. The

full dataset has 1,237,207 SNPs. I downloaded the EIGENSTRAT format of the data from `http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets_files/MathiesonEtAl_genotypes_April2016.tar.gz` and converted the files to PED format using EIGENSOFT.

Table 2.2: Brief description of 163 ancient individuals used in this study.Subset of Extended Data Table 1 from (Mathieson et al., 2015). N, represents the total number of samples for each populations. Out: the outliers determined by PCA by the authors. Rel: the individuals estimated to be related and removed from the analysis by the authors.

Population	**N**	**Out**	**Rel**	**Date range**
Anatolia_Neolithic	26	1	1	8.4-8.3 kya
Bell_Beaker_LN	10	0	1	4.5-4.5 kya
Central_LNBA	26	0	2	4.9-4.6 kya
Central_MN	6	0	0	5.9-5.8 kya
EHG	3	0	0	7.7-7.6 kya
Hungary_BA	2	0	0	4.2-4.1 kya
Hungary_EN	10	0	0	7.7-7.7 kya
Iberia_Chalcolithic	14	1	2	4.8-4.2 kya
Iberia_EN	5	0	1	7.3-7.2 kya
Iberia_MN	4	0	0	5.9-5.6 kya
LBK_EN	14	1	0	7.5-7.1 kya
Motala_HG	6	0	0	7.9-7.5 kya
Poltavka	5	1	0	4.9-4.7 kya
Potapovka	3	0	0	4.2-4.1 kya
Samara_Eneolithic	3	0	0	7.2-6.0 kya
Scythian	1	0	0	2.4-2.2 kya
Srubnaya	14	1	1	3.9-3.6 kya
WHG	2	0	0	8.2-8.0 kya
Yamnaya_Samara	9	0	0	5.4-4.9 kya

In their paper Mathieson et al. (2015) had excluded five population genetic outliers (I0056, I0354, I0432, I0581 and I0725) based on their divergence patterns identified by principal components analysis (PCA) (**Table 2.2**). There were also excluded from further analysis in my work.

From this ancient polymorphism dataset, eight individuals (I0054, I0100, I0108, I0112, I0172, I0408, I0412 and I1549) from West Eurasian Neolithic period populations (**Table 2.3**) were selected to generate an in-silico pedigree. The relat-

edness estimation process (described in **Sections 2.2.3**, **2.2.4** and **2.4** for SNP filtering, reference population selection and relatedness estimation respectively) was performed on the whole ancient dataset. These eight individuals selected represent unrelated samples (see **Appendix A**) from a similar time period. Here, I only report the calculations and results based on analysis of these eight individuals.

Table 2.3: Archaeological background of eight ancient individuals used in relatedness estimation analysis (Mathieson et al., 2015)

Sample ID	**Archaeological Culture**	**Date**	**Location**	**Country**
I0054	LBK_EN	5,122 *BCE	Unterwiederstedt	Germany
I0100	Anatolia_NE	6,350 BCE	Barcın	Turkey
I0108	Bell_Beaker_LN	2,437 *BCE	Rothenschirmbach	Germany
I0112	Bell_Beaker_LN	2,300 *BCE	Quedlinburg XII	Germany
I0172	Central_MN	3,223 *BCE	Esperstedt	Germany
I0408	Iberia_MN	3,750 BCE	La Mina	Spain
I0412	Iberia_EN	5,194 *BCE	Els Trocs	Spain
I1549	Bell_Beaker_LN	2,275 BCE	Benzingerode -Heimburg	Germany

2.2 Data processing

2.2.1 Sequence data processing

2.2.1.1 Down-sampling the sequence data of modern genomes

The modern sequence data (**Section 2.1.1**) was down-sampled at coverage and SNP levels to mimic the features of degraded ancient DNA samples.

The depth of coverage or shortly the coverage of a genome sequencing dataset is the average number of reads aligned to an individual base from the reference genome. The coverage level of a dataset is a determinant of accurate variant discovery at specific base positions. At higher sequence coverages, the degree of confidence increases because each base is covered by more sequenced reads. High coverage data could account for some of the inevitable sequencing errors

of Next Generation Sequencing (NGS) methods (Liu et al., 2012; Nielsen, Paul, Albrechtsen, & Song, 2011). Ancient DNA samples generally have low sequence coverages that would generate low confidence SNP calls.

To reduce the coverage of modern BAM files, I used the samtools software's "view" algorithm that subsamples the file by choosing a fraction of read pairs randomly. I down-sampled the BAM files from 50X coverage to 10X, 2X, 1X and 0.1X coverages. The fractions used for down-sampling of each sample is summarized in **Table 2.4**.

Table 2.4: Fraction of read pairs for down-sampling process

Sample	**Fraction for 0.1X**	**Fraction for 1X**	**Fraction for 2X**	**Fraction for 10X**
NA12877	0.0021	0.021	0.042	0.21
NA12883	0.0022	0.022	0.044	0.22
NA12885	0.0023	0.023	0.046	0.23
NA12889	0.0019	0.019	0.038	0.19
NA12890	0.0024	0.024	0.048	0.24

I also checked the quality of the original and downsampled BAM files using FastQC software (Andrews, 2010). The next part was sub-sampling of SNPs in modern sequence data, again to mimic low quantity ancient DNA datasets. For all coverages except 0.1X, I randomly decreased the SNP numbers to 100K, 50K, 20K, 10K, 5K and 1K. For 0.1X coverage the SNP number was reduced to 40K, 30K, 20K, 10K, 5K and 1K due to total available SNP number.

2.2.1.2 Checking accuracy of down-sampling

The success of down-sampling was confirmed by calculating the coverage of sub-sampled BAM files. Using samtools, first I computed the number of mapped reads with the "flagstat" algorithm, and then multiplied it with read lengths obtained with the "view" parameter. Dividing this number by the genome size gives the coverage for the file (**Figure 2.3**). This step verifies the accuracy of down-sampling.

$$Coverage\ of\ BAM\ file = \frac{number\ of\ mapped\ reads\ X\ read\ length}{genome\ size\ (bp)}$$

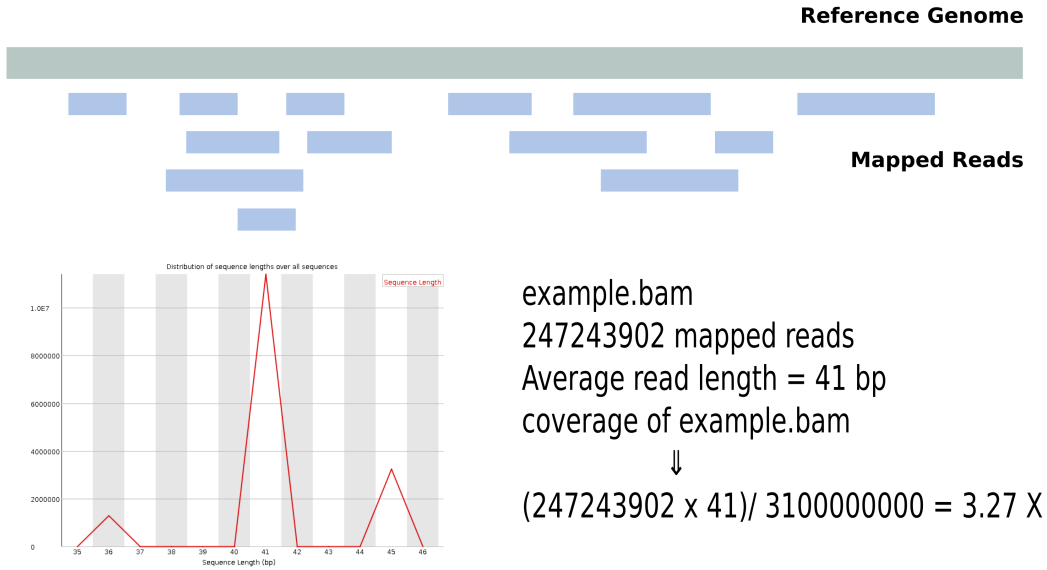

Figure 2.3: Sequence coverage estimation. An example illustration to describe steps of sequence coverage calculations.

2.2.2 SNP discovery

Variant calling is one of the steps in NGS (next generation sequencing) data analysis that reveals nucleotide differences (variants) between an individual and a reference genome.

There are different software packages for variant calling of sequenced genomes. Samtools (Li et al., 2009) "mpileup" is one of these algorithms that calls variants from reads aligned to the reference genome. Although GATK is a more sophisticated SNP calling algorithm, many ancient DNA studies use the pipeline of samtools for variant calling. According to Hwang et al. (2015) samtools is a better pipeline for SNP calling while GATK is more efficient in calling indels. However, most aDNA data have low sequencing coverage and indel calling is not possible, that is why we used samtools (Li et al., 2009) version 1.1 instead of GATK.

The variant calling process with the samtools software's "mpileup" command requires a BAM file, a reference genome (in accordance with the reference genome used in the mapping process) and an optional SNP list in BED format as input data. The BED file functions as a list of positions to be called. The Human Origins dataset (**Section 2.1.2**) was used as a list of positions to be called. We do not call de novo SNPs because aDNA samples have short read length (very fragmented) and low coverage. To solve this problem in modern samples: i) high amount of DNA is sequenced, ii) paired-end and mate-pair sequencing techniques are used on longer DNA fragments. However, these methods cannot be incorporated into aDNA sequencing. De novo assembly of low coverage, fragmented aDNA would result in erroneous base calling. Most of these errors could be detected by using reference based approaches (Seitz & Nieselt, 2017).

The "mpileup" parameters -q and -Q represent the minimum mapping quality and minimum base quality used for SNP calling respectively. Performing mapping and base quality filters during variant calling helps to rule out false positive SNP calls. The next step of variant calling is to convert the generated BCF file to VCF file format while removing indels using bcftools (Li et al., 2009) version 1.4.1. Presence of indels near mismatches could lead to false variant calling (Li & Homer, 2010).

The VCF file is a text file format that stores called variants along with the reference genome. This is a more efficient way of representing large number of genotype data compared to other file formats like General Feature Format (GFF) that stores all genetic data. In order to simplify and make files compatible with other software used in this analysis, the VCF files were converted to PLINK PED/MAP format using vcftools (Danecek et al., 2011) version v0.1.12b with parameter ("-cV indels"). This pipeline is used for SNP discovery of both ancient and modern genome sequences.

2.2.2.1 SNP discovery from ancient genome sequences

For the ancient SNP discovery, 163 ancient samples (see **Section 2.1.3**) from Mathieson et. al. (2015) were selected for the analysis. The human reference genome (version hs37d5), the Human Origins dataset SNPs (**Section 2.1.2**) and minimum mapping (-q) and base (-Q) quality 30 were used in SNP calling process. This quality threshold is commonly used for ancient DNA samples (Kılınç et al., 2016). After removing indels, the files were converted to PLINK PED/MAP format. All the heterozygous sites were haplodized by random selection of one of the alleles to avoid post-mortem nucleotide changes (Skoglund et al., 2012).

2.2.2.2 SNP discovery from modern genome sequences

The genotype calling of modern sequence data (five sets from **Section 2.2.1.1**) were performed with the pipeline in **Section 2.2.2**. In agreement with the mapping step, human reference genome version hg19 was used. The HO dataset (**Section 2.1.2**) was utilized to specify regions of genome for variant calling. For consistency with ancient samples, the minimum mapping and base quality was set to 30. The indels were removed and the files were converted to PLINK PED/MAP format.

2.2.3 Principal Component Analysis (PCA)

Principle Component Analysis (PCA) is a well-established mathematical technique that is used in analysis of high-throughput biological data. This method is used as a data quality check in many biological studies to visualize variation among many samples. By using PCA the biological signal of interest is compared against the signals coming from experimental conditions and bias. By using orthogonal transformation, PCA reduces the dimensions and noise but retains the original variation of the data. It is a way of compressing data into uncorrelated, orthogonal dimensions (components) with the highest data variation (Ringnér, 2008; Yao, Coquery, & Lê Cao, 2012).

In aDNA studies PCA is used to describe the genetic structure of ancient samples compared to modern-day populations. I performed PCA to investigate the genetic relationship of study samples: i) modern (CEPH 1463 **Section 2.1.1**), ii) ancient (**Section 2.1.3**) to modern-day individuals of the Human Origins Dataset. The populations that are closest (genetically) to our study samples (i & ii) were selected as reference populations. These two reference population sets were later used for minor allele frequency (**Section 2.2.4.4**) and linkage disequilibrium (**Section 2.2.4.3**) calculations. The smartpca program of EIGENSOFT (Patterson, Price, & Reich, 2006) software was used for PCA calculations of modern and ancient datasets. For both case, 50 modern West Eurasian populations were selected from HO dataset.

For CEPH 1463 PCA, the five samples of the family were merged with 50 West Eurasian populations from HO dataset. The eigenvector (principal components) calculations were done among the merged 51 populations. The first two components were used for plotting.

A total of eight ancient individuals (**Section 2.1.3**) and 50 West Eurasian populations from HO dataset were used for ancient PCA calculations. First the transition sites were removed from the ancient samples (**Section 2.2.4.1**). Also, all the multiallelic variants were excluded from the HO dataset. Then using PLINK's "–merge" parameter the two datasets were merged together. All the heterozygous sites were haploidized by selecting one of the alleles randomly to generate a completely homozygous dataset using a custom Python code. This

step is performed for consistency, because low coverage aDNA creates a non-diploid dataset. The eigenvector calculation was performed on modern populations. Then the ancient individuals were projected onto first two components using the "lsqproject: YES" option of smartpca.

2.2.4 Filtering the SNPs

2.2.4.1 Filtering the transitions

As mentioned before, one property of aDNA is cytosine to uracil/thymine transitions that accumulate gradually overtime. To avoid post-mortem degradation effects, the analyses are limited to transversion sites in the HO dataset (Skoglund et al., 2014). All the positions that had transitions ($C \leftrightarrow T$ or $G \leftrightarrow A$) in the HO dataset were excluded from the next steps of analysis using custom Bash code (see **Appendix B** for workflow of SNP filtering steps).

2.2.4.2 Filtering the positions based on missingness

The next step after removal of transitions is to look for genotyping (missingness) rates among ancient samples. Due to the technical reasons (e.g. biochemistry of the in-solution SNP capture procedure) it is possible that some SNPs systematically tend not to be called and are reported as missing. In our group's future work using in-solution SNP capture, we will prefer not to use these SNPs. I therefore decided to remove any such SNPs from my analysis here as well. I note that missing SNPs are not included in my related calculation. I calculated missingness frequencies (rates) for each SNP (locus) of ancient samples using PLINK's "–missing" parameter. This calculates missingness frequencies for each SNP. I removed SNPs that had missingness frequency higher than 0.5, meaning that these SNPs were not genotyped in more than half of the samples.

2.2.4.3 Filtering the SNPs based on Linkage Disequilibrium (LD)

The other step of filtering is LD-based SNP pruning. The situation where alleles from distinct loci have a non-random association is called linkage disequilibrium (LD) (Slatkin, 2008). It is better for our calculations to use SNPs that are independent from each other. This way we could avoid over- or underestimation of relatedness estimates our measurements. PLINK's calculations depend on pairwise genotypic correlation.

The linkage disequilibrium coefficient D, is one of the measures to estimate LD which is defined as

$$D_{AB} = p_{AB} - p_A p_B \tag{2.1}$$

where, p_A is the frequency of allele A at one locus. In another locus, p_B is the frequency of allele B. In the same manner, p_{AB} epresents the AB haplotype frequency, that is the situation where A and B occur together in the same gamete. However, D measurements are hard to interpret because their range are dependent on the frequency of alleles. This makes the comparison of pairs of markers based on D difficult (Ardlie, Kruglyak, & Seielstad, 2002; Devlin & Risch, 1995). Another way of estimating LD is Pearson's coefficient of correlation (r); the square of r is often used to avoid the introduced arbitrary sign. The r^2 value is measured as

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{2.2}$$

The range of r^2 extends from 0 to 1. Zero indicates a situation where two loci are in complete linkage equilibrium where 1 means that they are in complete linkage disequilibrium (Ardlie et al., 2002; Devlin & Risch, 1995; Lewontin, 1964; Slatkin, 2008).

The LD-based SNP pruning was performed using modern reference populations selected from PCA analysis (10 West Eurasian population for ancient dataset as described in **Section 3.1.2** and six Northern European populations for modern dataset described in **Section 3.2.2**). Loci in the modern reference population dataset with LD of 0.4 and higher were removed using PLINK's "–indep-pairwise" parameter. I set the program to take a window size of 200 SNPs, shift it 25 SNPs each time and calculate pairwise SNP LD. If the pairwise LD was higher than 0.4, one of the SNP pairs was removed recursively from both

the reference (**Section 2.2.3**) and test (**Section 2.1.3**) datasets (Purcell et al., 2007).

2.2.4.4 Filtering the SNPs based on Minor Allele Frequency (MAF)

Allele frequency per locus is determined as the proportion of a particular allele type to total number of measured alleles at that locus in a population. The measured allelic frequency can be polarized to minor/major, reference/alternative or ancestral/derived states. At that precise locus, the least frequent allele in the population is designated as the minor allele (Gillespie, 2004).
Using the PLINK "–freq" parameter, I calculated MAF (minor allele frequency) for each position by determining minor and major alleles in filtered reference populations. In order to use SNPs with accurate genotype calling, the positions with MAF lower than 0.1 were removed from analysis. Because of the features of aDNA data (Skoglund et al., 2014), I used a higher cutoff for MAF values than the usual 0.01.

2.3 Generation of *in-silico* pedigrees

In order to test the performance of our method, I simulated a four generation pedigree that included relatives up to third-degree relatives. For this I used the genomes of eight real ancient individuals (**Section 2.1.3**) that belong to the Anatolian, Central European and Iberian population of the Neolithic period. I used three criteria for selecting these individuals, which I call "founders": i) the founders should belong to the same time period, ii) the founders should have the highest number of SNPs possible, iii) the founders should be unrelated.
A pedigree was simulated using these founders, and using the remaining SNPs after the pruning process in **Section 2.2.3**. First, I generated four trios from the founders. The formation of each trio is summarized in the **Figure 2.4**.
Each pedigree trio is composed of two founder individuals and a simulated offspring. From each founder genome, I randomly selected one of the biallelic SNPs at each locus. By combining these two haploid sets, I created a diploid offspring

that is the random mixture of its parents' genomes.

The first set of simulated offspring (Child1, Child2, Child3 and Child4) was created in this fashion from four founder pairs. This set of simulated offspring genomes was used recursively to create a third-degree family pedigree in the same manner as shown in the **Figure 2.5**.

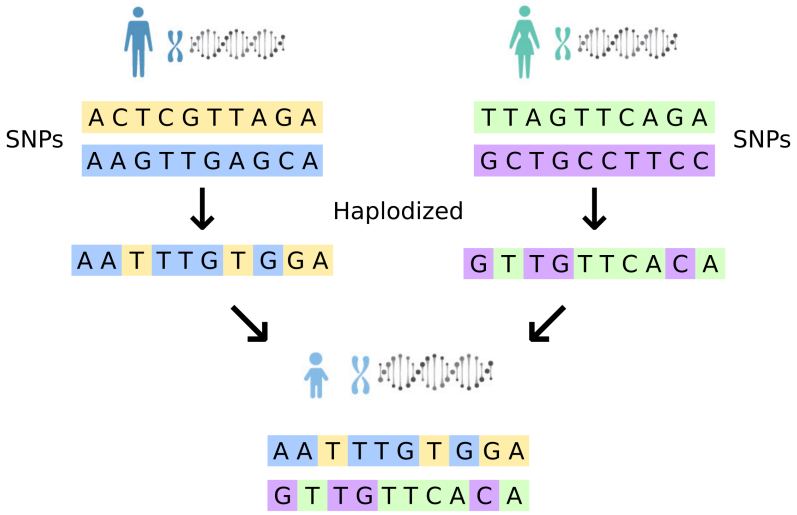

Figure 2.4: Schematic representation of Child simulation, modified from `http://www.natera.com/science-informatics`.

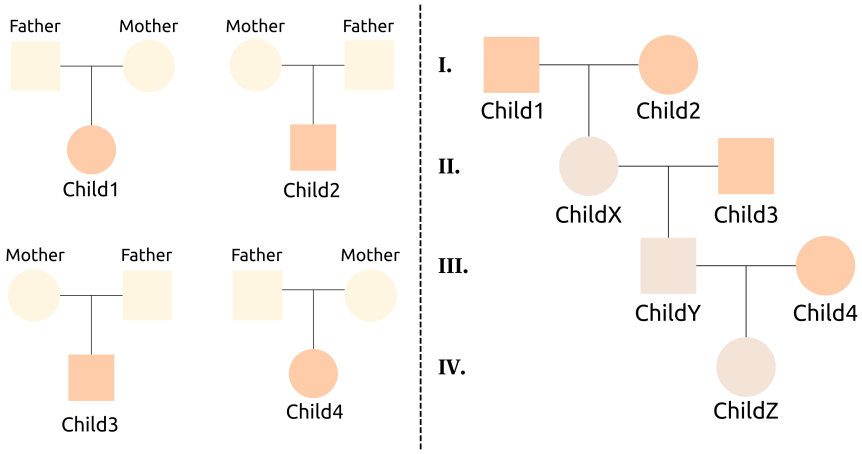

Figure 2.5: Example of pedigree simulation

2.4 Estimation of relatedness

2.4.1 Evaluation of published methods

2.4.1.1 PLINK

PLINK is a free, open-source platform for whole-genome association analysis (Purcell et al., 2007). The tool set performs a wide range of genotype/phenotype data analyses. One feature of the program is to calculate pairwise identity by descent (IBD). For this purpose, the program uses shared rare variants between two individuals. The algorithm searches through common SNPs to identify the regions of extended sharing. These regions are assumed to be inherited via IBD, and identifying these regions enables us to estimate IBD. PLINK assumes homogeneous population structure (random mating) for IBD estimation among apparently unrelated pairs. This feature of the program is mainly designed for quality-control purposes like identifying contamination, duplication, sample mix up, pedigree errors and undiscovered relatedness (Mathieson & McVean, 2014; Purcell et al., 2007).
To estimate pairwise relatedness as a measure of relatedness, I used PLINK's (Purcell et al., 2007) "–genome" parameter that calculates the relatedness coefficient r from the IBS information with formula

$$PI_{HAT} = P(IBD = 2) + 0.5 \times P(IBD = 1) \tag{2.3}$$

where P stands for probability. The IBD probabilities are calculated based on the observed IBS proportions.

2.4.1.2 KING

KING (Manichaikul et al., 2010) is one of the software generated to calculate relatedness among individual pairs using genome-wide polymorphism data. According to Manichaikul et al. (2010), KING's algorithm could calculate kinship up to third-degree among thousands of related pairs of a high-throughput data, even in the presence of unknown population substructure (Manichaikul et al.,

2010). I conducted pair-wise relationship inference analysis using KING's "–kinship" parameter that calculates Θ, the kinship coefficient.

2.5 Method development for estimating the relatedness

Two main approaches to relatedness analysis, are pedigree-based (**Section 1.1.1**) and SNP-based measurements (**Section 1.1.2.2**) (Speed & Balding, 2014). Here our aim is to estimate relatedness from ancient DNA data therefore we use a SNP-based approach of relatedness.

Our kinship coefficient estimations are based on Genetic Similarity Matrices (GSMs) as described by Speed and Balding (2014). The calculations are based on the unbiased estimation of allelic correlation coefficient with the equation,

$$K_{c-1}(B,C) = \frac{1}{m} X_B X_C^T \; where \; X_{Bj} = \frac{S_{Bj} - 2p_j}{\sqrt{2p_j(1-p_j))}} \tag{2.4}$$

K_{c-1} is one of the pairwise measures of genome similarity, that estimates 2Θ (allelic correlation coefficient) averaged over SNPs. $\frac{1}{m}XX^T$ is the genetic similarity matrix where the genotype scores are centered and standardized. S_{Bj} is the genotype of B at the j^{th} dialleleic SNP, with the genotypes coding system of 0,1,2. The total SNP number is indicated by m and the population MAF at the j^{th} SNP is denoted by p_j.

Subtracting the population frequency of minor alleles (MAF) from genotype scores accounts for the level of background relatedness in a population. Individuals of a finite population have some degree of genetic similarity due to ancestral relatedness (sharing a common ancestor in the past) (Weir et al., 2006). In this way, we could differentiate the familial relatedness from ancestral relatedness. This rescaling puts more weight on rare shared alleles. Frequency of an allele could be used to estimate its age. Rare alleles are typically newly arisen variants compared to common variants. Sharing rare alleles could be the indicator of a recent common ancestor among individuals (Mathieson & McVean, 2014).

The relatedness calculations were performed in the R programming environment with the SNPs remaining from the filtering process (**Section 2.2.4**). The genotypes of test individuals were transformed to the 0,1,2 coding system after

comparison to that of reference population set. If the test allele genotype is the same as the minor allele, it is coded as 2. If it shares the same genotype with the major allele, it is coded as 0. When there is a missing data or a different genotype than both minor and major alleles, the allele information is transformed to NA and removed from analysis. The pairwise relatedness was calculated with the equation (2.4). **Table 2.5** describes the expected kinship coefficient (Θ) values for each relationship degree, and their corresponding 95% credible interval (CI).

Table 2.5: Theoretical criteria of kinship coefficient (Θ) estimation (Speed & Balding, 2014)

Relationship	Θ	**95% CI of** Θ
first degree (parent-offspring) (full-siblings)	0.25	(0.204, 0.296)
second degree (grandparent-grandchild)	0.125	(0.092, 0.158)
third degree (first cousins)	0.0625	(0.038, 0.089)
unrelated	0	(0.000, 0.000)

2.6 Statistical analysis

One of the crucial factors in determining validity of a test is knowing its accuracy. In order to evaluate the performance of our method in estimation of kinship degrees, five statistical measures were calculated. The true positive (TP) rate measures the amount of related individuals that were correctly inferred as related (in the correct kinship degree) by our approach. The proportion of unrelated pairs which were estimated as related individuals constitute the false positive (FP) rate. In contrast, the false negative (FN) rate represents the proportion of related individuals that were mistakenly identified as unrelated pairs with our approach. The category of incorrectly related (IR) pairs were composed of related pairs assigned to the wrong kinship degree. And finally the undecided

(UN) proportion represents estimates that were in between the expected range of relatedness degrees.

These statistical measures were calculated separately for the first, second and third degree individual pairs in the in-silico pedigree simulated from ancient samples and first and second degree relatives of real-life modern data of 50X and 10X coverages.

For the individuals in the datasets, the SNP numbers were randomly reduced to 25K, 20K, 15K, 10K, 5K and 1K to observe how the efficiency of our approach vary as the SNP number diminishes. Each set of SNPs were sampled 1000 times, then the kinship estimation method was performed on all of them. Every statistical measure was in percentages and calculated by dividing the number of true positives to total number of truly related pairs for TP, dividing the number of false positives to the total number of pairs that were unrelated for FP, dividing the false negatives to all the pairs that were truly related for FN, dividing the number of pairs that were categorized to an incorrect kinship degree to total number of pairs in accurate kinship degree for IR and lastly by dividing the undecided pairs to the total number of true related pairs.

CHAPTER 3

RESULTS

3.1 Simulated ancient data

3.1.1 Sequencing data summary and statistics for ancient data

Our method's performance was tested on a four generation in-silico pedigree simulated from real ancient data (**Section 2.1.3**). The eight founder individual (ancient samples I0054, I0100, I0112, I0172, I0408, I0412 and I1549) were selected because they had the highest number of SNPs among unrelated individuals of European Neolithic period.

The pipeline in **Section 2.2.2** and **2.2.2.1** were used for SNP discovery of ancient samples. An ancient sample's DNA usually is in low amount and highly fragmented. It is better to use reference-based approaches (conditioning on confidently identified SNPs) to avoid incorrect base calling. The SNPs that overlapped with the Human Origins dataset were used for the analysis. This is a SNP array that represents genetic variation in human populations from all around the world and was designed by the David Reich group (Lazaridis et al., 2016; Patterson et al., 2012) for studying human population genetics and history. The number of SNPs from the chosen ancient genomes overlapping with the HO dataset is shown in **Table 3.1**.

Table 3.1: Summary statistics of eight ancient individuals

Sample ID	N. SNPs overlapping with HO dataset	Genome coverage
I0054	561,535	0.82
I0100	545,367	0.56
I0108	399,902	0.12
I0112	533,394	0.23
I0172	580,157	0.98
I0408	537,636	0.59
I0412	553,669	1.13
I1549	415,223	0.12

3.1.2 Reference population selection for ancient individuals

Principle Component Analysis (PCA) of the eight ancient study samples (**Section 2.1.3**) was performed for determination of the reference populations set. The allele frequencies in reference populations are important for discovery of background relatedness in study samples. Considering a finite population and a single locus, any two individuals share some degree of relatedness because at some point in time they had a common ancestor. Therefore, any observed relatedness among individuals should be measured against this background relatedness in the population (Weir et al., 2006). Accurate detection of reference populations is crucial for our relatedness calculations.

The selection of reference populations were based on their genetic affinities to study samples. I chose 50 West Eurasian populations from the HO dataset to test this relationship. The selected ancient samples belong to Anatolian, Central European and Iberian populations of the Neolithic period. As described in previous studies (Haak et al., 2015; Kılınç et al., 2016; Lazaridis et al., 2016; Mathieson et al., 2015) Neolithic samples from Europe usually cluster together with modern-day West and South European populations. For this reason, the other populations of HO dataset (African, American and East Asian) was not included in PCA.

One of the most prominent post-mortem damages of ancient samples, is the cytosine deamination at the end of DNA fragments (Briggs et al., 2007). This

translates into high amount of C to T substitution in ancient DNA especially at molecule ends (up to 30-40%) and could lead to inaccurate base calling (Seitz & Nieselt, 2017; Skoglund et al., 2014). Hence all the transition sites were removed from ancient samples data, leaving only transversion SNPs (**Section 2.2.4.1**). A total of 108,535 SNPs remained. **Table 3.2** illustrates the detailed result of transition filtering.

Table 3.2: Remaining SNPs after transition filtering for ancient samples. N. missing indicates the number of SNPs that are not identified for that individual (among all transversion SNPs).

Sample ID	**N. of SNPs**	**N.missing**
I0054	102,986	5,549
I0100	99,819	8,716
I0108	73,005	35,530
I0112	97,685	10,850
I0172	106,498	2,037
I0408	98,095	10,440
I0412	101,186	7,349
I1549	75,235	33,300

In addition, I haploidized the whole merged data by randomly selecting one of the alleles at each heterozygous position. Due to low coverage of ancient samples, haploidized version of the data is used for analysis (Green et al., 2010; Skoglund et al., 2012). The eigenvectors were calculated from 50 modern populations, ancient samples were later projected onto the first two calculated principal components (PCs) using the "lsqproject: YES" option of smartpca (Patterson et al., 2006). This option uses least squares equations instead of orthogonal projection which is appropriate for situations where some of the samples have a high proportion of missing data. Using modern data (few missing genotype) to calculate PCs and then projecting ancient data (many missing genotype) onto it would minimize the error and increase power (Patterson et al., 2012). The PCA result is shown in **Figure 3.1**.

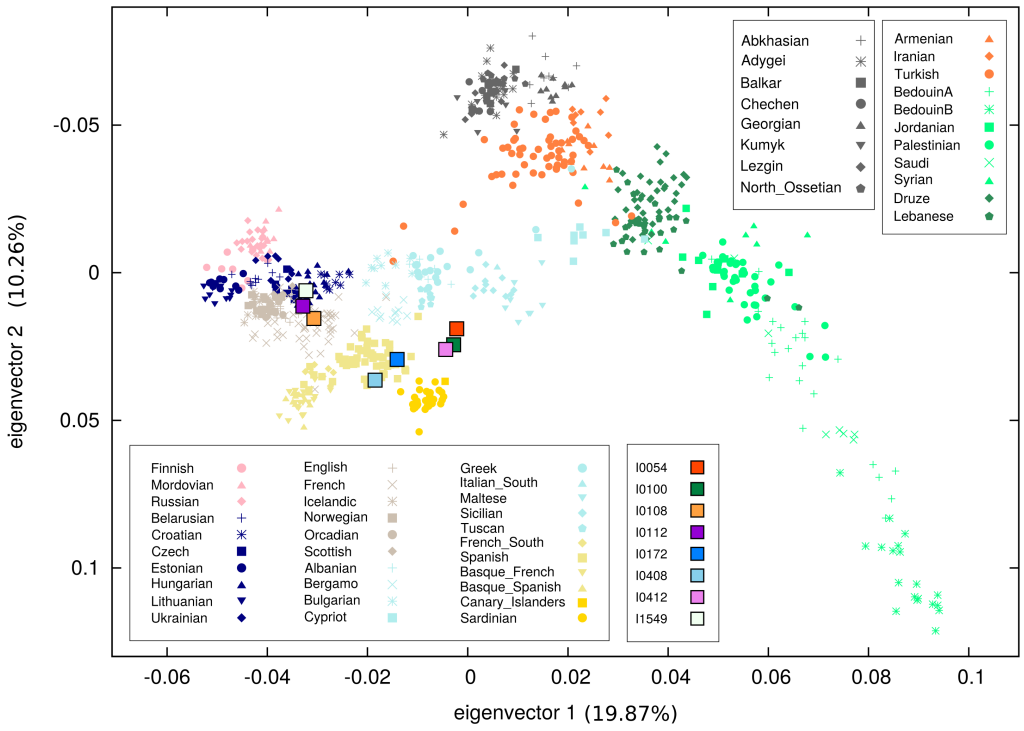

Figure 3.1: Genetic structure and population affinities of ancient samples. The principle component analysis (PCA) of 50 modern West Eurasian population on which eight ancient samples were projected.

The positioning of modern populations in the PCA is similar to their geographic distribution. As the genetic affinity between population increases, they form tighter clusters in the PCA. Consistent with previously reports (Lazaridis et al., 2014; Omrak et al., 2016), the Neolithic individuals are mostly (five out of eight) grouped with populations from Southern Europe. Considering the results of previous studies (Günther & Jakobsson, 2016; Kılınç et al., 2016) and this PCA, I chose Albanian, Bergamo, Greek, Italian_South, Maltese, Sicilian, Tuscan, Spanish, Canary Islanders and Sardinians as reference populations for the relatedness calculations. The reference population set for ancient data includes 153 individuals from 10 South European populations.

3.1.3 SNP filtering for ancient data

The next step after selection of reference populations, was to calculate missing genotype frequency in ancient samples. Ancient samples have a high proportion of missing data due to their low coverage and some level of SNP filtering is required to decrease their effect on our calculations. I removed all the SNPs that are systematically missing in more than half of the individuals (>50%). A total of 60,663 SNPs pass this filter. This step would still allow for some level of missing data in the samples. If all SNPs with any missing genotypes were removed completely, the few remaining SNPs would not be sufficient for relatedness estimations. The result of filtering missing SNPs is presented in **Table 3.3**.

Table 3.3: Remaining SNPs after missingness filtering for ancient samples

Sample ID	**N. of SNPs**	**N.missing**
I0054	60,476	187
I0100	60,306	357
I0108	48,310	12,353
I0112	59,586	1,077
I0172	60,638	25
I0408	60,489	174
I0412	60,542	121
I1549	53,891	6,772

The same set of SNPs were also removed from the reference population set. The SNP filtering process is executed in parallel between ancient data and reference population set.

The following two steps of SNP filtering (LD and MAF filtering) was performed on reference population set and not the ancient samples. This is because LD and MAF calculations are usually performed at the population level and we have few samples recovered from each ancient location. Their number is not enough for population level analysis, so for now, we have to use the genetically closest modern populations instead. We chose the 10 South European populations from the PCA as a reference for allele frequencies of ancient samples.

Because of replicated signals, variants in high LD would generate noisy related-

ness estimations. A common solution is pruning of SNPs to reduce correlation effect (Speed, Hemani, Johnson, & Balding, 2012). I used the PLINK (Purcell et al., 2007) program for LD pruning, which takes 200 bp size windows, calculates pairwise r^2 between SNPs and removes one of them if the r^2 value is larger than 0.4 (described in detail **Section 2.2.4.3**). The parameters used in LD calculations (200-window size, 25-size of sliding window and 0.4-r^2 threshold) were selected according to Kilinc et al. (2016). After pruning a total of 37,577 SNPs remains (details in **Table 3.4**).

Table 3.4: Remaining SNPs after LD filtering for ancient samples. N. missing indicates the number of SNPs that are not identified for that individual.

Sample ID	**N. of SNPs**	**N.missing**
I0054	37,459	118
I0100	37,339	238
I0108	29,871	7,706
I0112	36,867	710
I0172	37,559	18
I0408	37,458	119
I0412	37,498	79
I1549	33,456	4,121

The final step of SNP filtering based on MAF was performed using MAF estimations from the reference population set. Knowing the minor allele and its frequency in the population, is crucial for our approach in relatedness estimations. The HapMap (Frazer et al., 2007) and 1000 Genomes (Auton et al., 2015) projects demonstrate that accurate identification of rare variants (with 1% MAF) requires sequencing of thousands of individuals (Nielsen et al., 2011). However, there are few ancient samples available and these have low coverage data and damaged DNA. Therefore, I used a higher MAF threshold than the usual 1% to obtain more accurate results. All the SNPs that had a frequency lower than 0.1 were excluded from the reference population set. A total of 25,990 SNPs pass the filtering. The same group of SNPs were removed from ancient data as well. The detailed result of MAF SNP filtering is summarized in **Table 3.5**.

Table 3.5: Remaining SNPs after MAF filtering for ancient samples. N. missing indicates the number of SNPs that are not identified for that individual.

Sample ID	N. of SNPs	N. missing
I0054	25,905	85
I0100	25,824	166
I0108	20,644	5,346
I0112	25,489	501
I0172	25,977	13
I0408	25,905	85
I0412	25,938	52
I1549	23,110	2,880

3.1.4 Pedigree simulation and relatedness estimation

Following the SNP filtering, a total of 25,990 SNPs remained. These same group of SNPs was selected in both ancient data and reference population set. The 25,990 SNPs in the ancient individuals were used in simulation of a four generation pedigree according to the pipeline in **Section 2.3**. Randomly selected haploid alleles from a pair of founders were combined to generate each diploid Child. **Figure 3.2** shows the four trios constructed from the selected eight ancient individuals. The four simulated Children were used to create the four generation pedigree displayed in **Figure 3.2**.

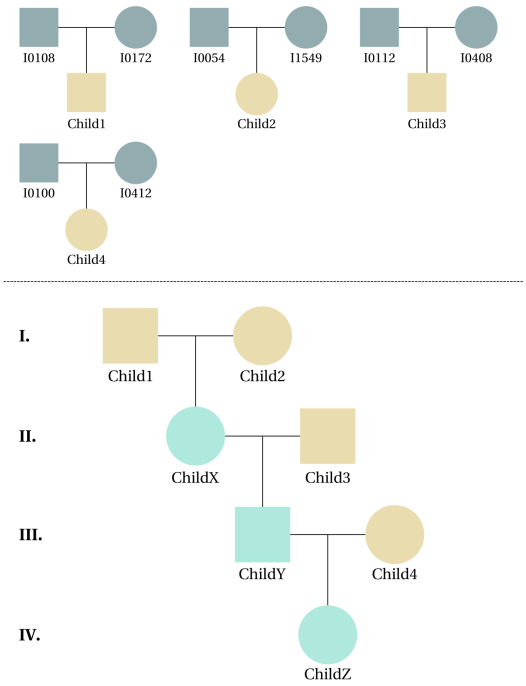

Figure 3.2: Topology of simulated pedigree from eight ancient individuals

After the pedigree was created, a haploid file of each sample was generated for relatedness calculations. For every sample, one of the alleles in each position is randomly selected. The genotype data of simulated individuals were merged with that of the reference population set (10 West Eurasian populations **Section 3.1.2**). For every individual, the genotype of each allele was compared with the genotype of minor allele in the population. If they had the same genotype with the minor allele, they were coded as 2. Whenever, they had the major allele, they were coded as 0. By this way, we would give more weight to shared rare alleles between individuals. When there was a missing allele or any other genotype than the minor or major alleles, the allele was removed from analysis. The pairwise relatedness between simulated individuals of the pedigree in **Figure 3.2** was calculated with the equation (2.4) in the **Section 2.5**. In addition, relatedness between simulated children was calculated with the software KING and PLINK. In the simulated pedigree, there are six pairs of first-degree relatives, four pairs of second-degree relatives, two pairs of third-degree relatives and nine pairs of unrelated individuals (**Table 3.6**).

Table 3.6: Relatedness degree of simulated pairs

Sim. pairs	**Rel. degree**	**Sim. pairs**	**Rel. degree**
Child1-ChildX	First-degree	Child2-ChildZ	Third-degree
Child2-ChildX	First-degree	Child1-Child2	Unrelated
Child3-ChildY	First-degree	Child1-Child3	Unrelated
ChildX-ChildY	First-degree	Child1-Child4	Unrelated
ChildY-ChildZ	First-degree	Child2-Child3	Unrelated
Child4-ChildZ	First-degree	Child2-Child4	Unrelated
Child1-ChildY	Second-degree	Child3-Child4	Unrelated
Child2-ChildY	Second-degree	Child4-ChildX	Unrelated
ChildX-ChildZ	Second-degree	Child4-ChildY	Unrelated
Child3-ChildZ	Second-degree	Child3-ChildX	Unrelated
Child1-ChildZ	Third-degree		

I used three methods (described in **Section 2.4.1.1**, **2.4.1.2** and **2.5**) for relatedness estimation among related and unrelated pairs in the Neolithic pedigree, as well as the 8 unrelated Neolithic individuals (see **Appendix A**):

- GSM-based allelic correlation coefficient (GACO), or Kc-1,
- PLINK,
- KING.

As the plot in **Figure 3.3** show, we were able to correctly infer the relatedness degree of all the simulated pairs with our approach. The gray shaded areas are the theoretical 95% credible intervals (CI) for each degree of relatedness as calculated by Speed and Balding (2015) (**Table 2.5**). The error bars at each point represents the 95% CI determined by bootstrapping. For each pairwise comparison, 5,000 SNPs were randomly sampled and the kinship coefficients were calculated. This process was repeated for 1,000 times. The estimates which were outside of the 95% two-sided confidence bounds were excluded.

If the reference population allele frequencies were different from those of the original ancient Neolithic population, we might find a general bias toward overestimating relatedness in this pedigree. We do observe a modest overestimation of relatedness in five of the unrelated individuals, but the accuracy in general appears high.

The KING software could not produce any informative results about the relationship degree of simulated individuals. Although the PLINK program performed the calculations, the obtained result was zero degree of relatedness for every pair, which is not accurate.

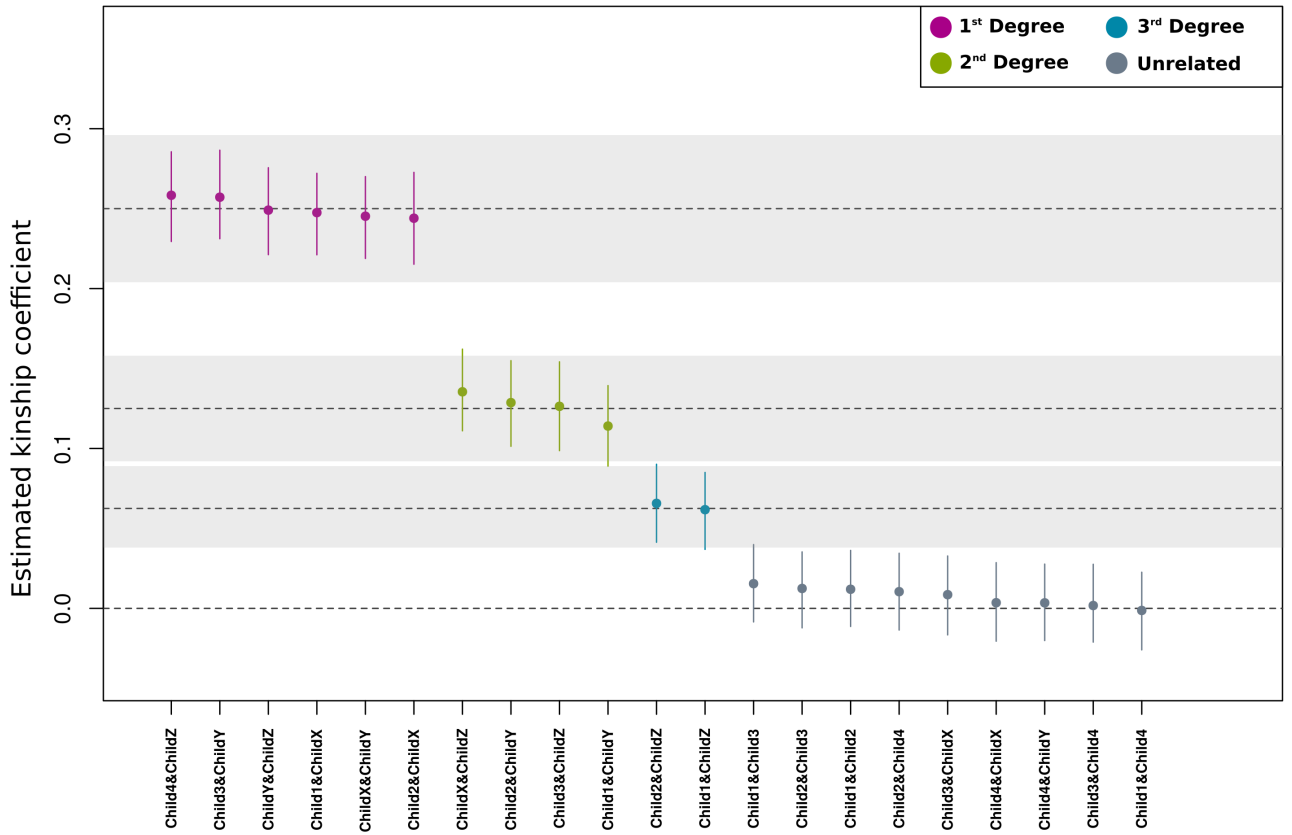

Figure 3.3: Kinship coefficient (Θ) estimation of simulated pedigree pairs. Each dot represents a pairwise kinship comparison (Table 3.6). Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).

3.1.5 Statistical analysis of kinship estimation method performance using the ancient simulated dataset

The seven members of the in-silico pedigree simulated from the ancient samples were used to test the performance of our approach. These seven members generate 21 pairwise comparisons that include six pairs of first-degree, four pairs of second-degree, two pairs of third-degree and nine unrelated pairs. We had six different SNP sets (25K, 20K, 15K, 10K, 5K and 1K), each of them were randomly subsampled 1000 times with replacement to yield a total of 126,000 pairwise comparisons (see **Appendix C**).

The **Figure 3.4** illustrates in detail the result of these statistical test for first, second and third-degree comparisons. Overall our approach could accurately infer kinship degrees in the simulated pedigree. As the SNP number decreases the precision of the results decline as well. Another factor affecting the estimation of kinship coefficient, is the degree by which individuals are related. Just as the degree of kinship increases, the level of shared genomic regions inherited from a common ancestor decreases and starts to overlap with each other (**Table 2.5**). For the six 1st degree pairs, all the SNP sets except the 1000 SNP one, showed perfect results (**Figure 3.4**). The FP, FN and IR rates were zero for the first-degree pairs. Only in the 1000 SNP set, some estimates were above the expected range of the first-degree pairs (> 0.296) and some estimates were in between the expected range of first-degree and second-degree relatives, yielding an UN rate of 15.4%. For the second and third-degree pairs, the high number SNP sets produced largely true result. However, for the 1000 SNP set, among the second-degree pairs, the IR rate increased to 8.8% and UN rate to 22.3% while in third-degree pairs, the FP rate was 16.4%, IR 26.4% and UN 14.7%. Therefore, only a 1000 SNPs (of MAF$>$0.1) may not be enough to make decisive estimations about kinship degrees. In summary, our approach could successfully estimate up to third degree kinship degrees with a minimum of 5000 SNPs.

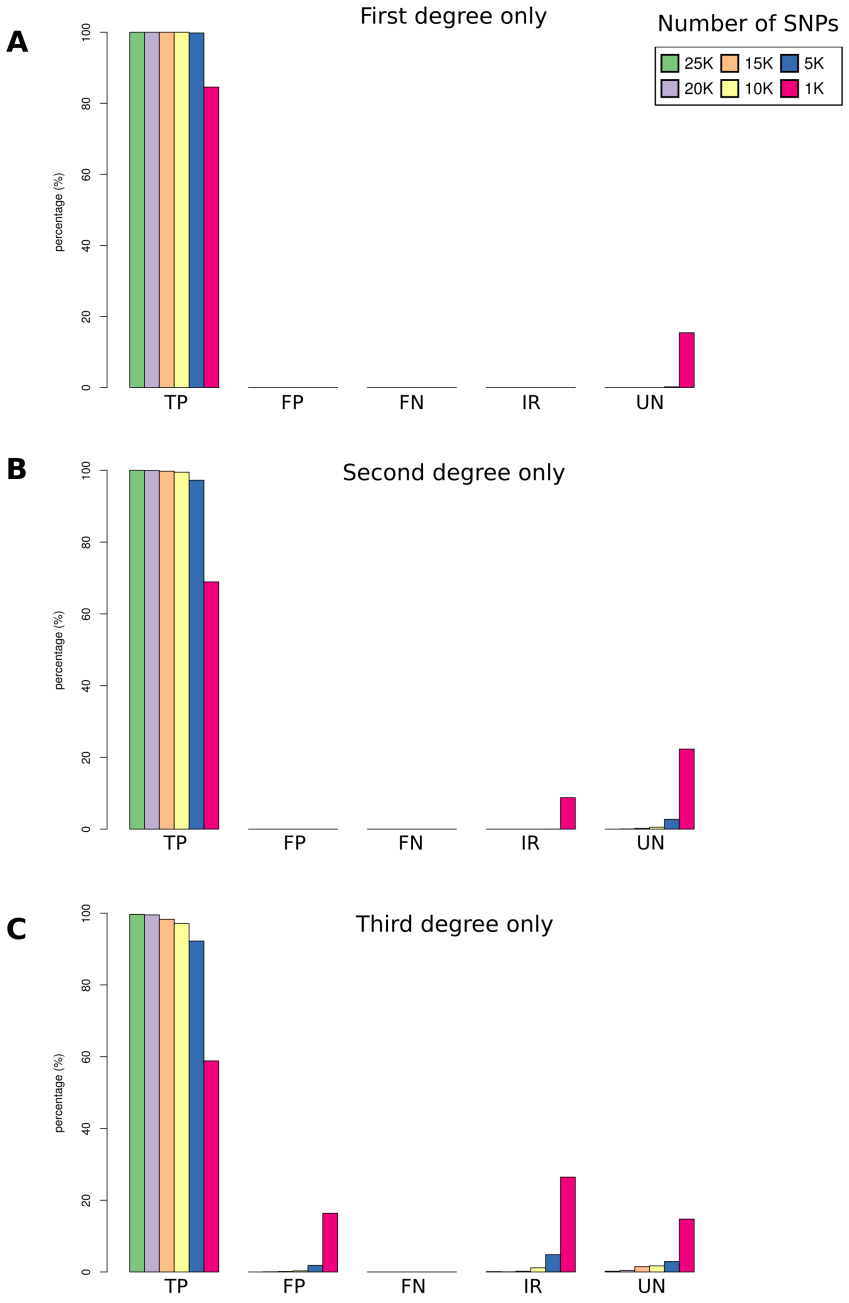

Figure 3.4: Error rates for ancient simulated data. All three plots demonstrate error rates for different SNP numbers. The x-axis shows the five different statistical measures, true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided". (A) For first-degree relatives only, (B) second-degree relatives only and (C) third-degree relatives only.

3.2 Real-life modern data with known relationship

3.2.1 Sequencing data summary and statistics for modern data

To test the performance of our approach on real-world data, I used five individuals from CEPH 1463 family of North/West European descent from Utah, USA (**Section 2.2.1**) who were sequenced to 50X coverage. These modern samples were processed (down-sampling and transition removal) to resemble properties of ancient DNA. Moreover, KING and PLINK software were used to estimate kinship degrees of these processed real-life data. The coverage of modern sequence data (**Section 2.1.1**) was gradually reduced (from 50X to 10X, 2X, 1X and 0.1X) to mimic the low coverage features of aDNA. The performance of down-sampling was tested by calculating coverage of each file using the formula in **Section 2.2.1.2**. This step verifies the accuracy of down-sampling (details in **Table 3.7**).

Table 3.7: Coverage calculation and conformation of down-sampling process

Sample ID	**Number of mapped reads**				**Average read length**	**Observed coverage**			
	0.1X	**1X**	**2X**	**10X**		**0.1X**	**1X**	**2X**	**10X**
NA12877	3,278,213	32,746,765	65,491,328	327,429,800	101	0.11	1.07	2.13	10.67
NA12883	3,138,538	31,359,111	62,709,238	313,534,139	101	0.10	1.02	2.04	10.22
NA12885	3,239,668	32,399,287	64,794,236	324,029,115	101	0.11	1.06	2.11	10.56
NA12889	3,153,582	31,517,048	63,038,309	315,215,827	101	0.10	1.03	2.05	10.27
NA12890	3,264,022	32,646,729	65,302,268	326,488,405	101	0.11	1.06	2.13	10.64

The SNP discovery of all modern samples at five different coverages were performed according to the pipelines in textbfSection 2.2.2 and **Section 2.2.2.2**. The total number of SNPs for each sample that overlapped with the HO dataset (**Section 2.1.2**) is summarized in **Table 3.8**.

Table 3.8: Number of SNPs of CEPH family 1463 used in this study

	N. of SNPs Overlapping with HO Array				
Sample ID	**0.1X**	**1X**	**2X**	**10X**	**50X**
NA12877	55,059	367,847	506,126	592,023	592,122
NA12883	55,112	367,119	505,425	592,027	592,118
NA12885	55,858	370,363	507,664	592,012	592,120
NA12889	53,591	361,254	500,820	592,006	592,121
NA12890	56,138	372,842	510,150	592,041	592,117

3.2.2 Reference population selection for modern data

I performed PCA to determine the reference populations set of CEPH family 1463 (with the pipeline in **Section 2.2.3**). The 50 West Eurasian populations (**Section 2.1.2**) from the HO array were merged with the five individuals of CEPH family. The eigenvector calculations were performed on the whole merged (51 populations) dataset. **Figure 3.5** displays the PCA result from the first two components.

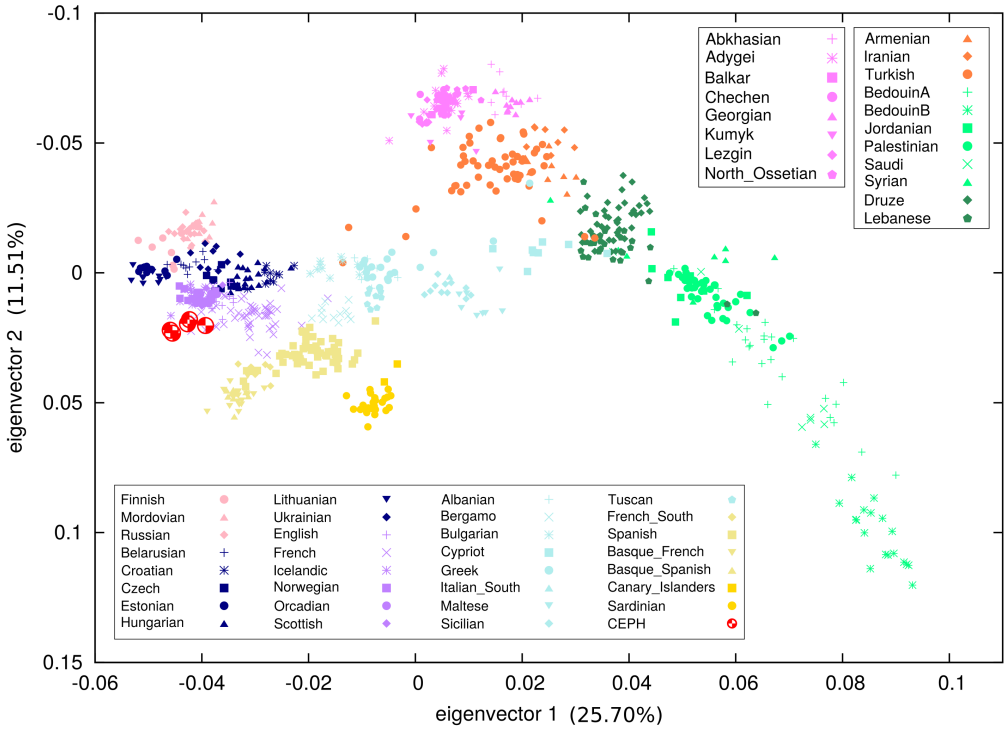

Figure 3.5: Genetic structure and population affinities of CEPH family 1463. PCA of 50 contemporary West Eurasian populations with the five members of CEPH family.

According to the PCA result, the individuals from CEPH family 1463 mostly group with populations from Northern Europe. CEPH 1463 is a family from Utah with European ancestry and belonging to the Mormon community (Zhang et al., 2004), so this is expected. However, why CEPH 1463 form a slightly distinct group at the edge of PCA is not so obvious and could be related to the consanguineous marriage practices and consequent genetic homogeneity of Mormons (Jorde, 2001). I chose the closest populations (English, French, Icelandic, Norwegian, Orcadian and Scottish) to be the reference populations set for the CEPH 1463 family.

3.2.3 SNP filtering for modern data

Because of post-mortem decay of aDNA, the analysis of ancient samples is usually restricted to transversion sites only (Skoglund et al., 2014). The modern

data (down-sampled to resemble the features of ancient DNA) was used as a model to investigate the performance of software estimating kinship coefficient. Therefore, after SNP calling, all the transition sites were removed from modern data as well (**Section 2.2.4.1**). However, I did not remove the transition sites from the data for 0.1X coverage, because the reduction in SNP numbers made the calculations in the subsequent steps impossible. **Table 3.9** shows the number of remaining SNPs for each coverage after filtering.

Table 3.9: Remaining SNPs after transition filtering for modern data

Sample ID	**1X**	**2X**	**10X**	**50X**
NA12877	68,137	93,836	110,078	110,097
NA12883	68,228	93,999	110,083	110,098
NA12885	68,888	94,397	110,083	110,100
NA12889	67,214	93,148	110,078	110,100
NA12890	69,103	94,788	110,079	110,097

For consistency with the ancient pipeline, all the SNP filtering steps in **Section 2.2.4** was also applied to all coverages of the modern data. Genotype frequency was calculated, then all the SNPs that were missing in more than half of the samples (>50%) were excluded from analysis. The result of missingness filtering is reported in **Table 3.10**.

Table 3.10: Remaining SNPs after missingness filtering for modern data

Sample	**0.1X**	**1X**	**2X**	**10X**	**50X**
NA12877	55,059	57,505	92,744	110,077	110,094
NA12883	55,112	57,706	92,898	110,083	110,094
NA12885	55,858	58,026	93,240	110,083	110,094
NA12889	53,591	56,836	92,064	110,076	110,094
NA12890	56,138	58,164	93,607	110,078	110,094

The next step was LD SNP pruning according to the pipeline described in **Section 2.2.4.3**. If the calculated pairwise LD in the reference population set was greater than 0.4, one of the SNPs were removed from both of the reference and study data (details in **Table 3.11**).

The last step was filtering SNPs according to their MAF values. The allele fre-

quencies were calculated in the reference population set and one of the alleles at each position was determined as minor allele. The SNPs with minor allele frequency (MAF) of 0.1 and lower were excluded from the analysis. The remaining SNPs is shown in **Table 3.12**.

Table 3.11: Remaining SNPs after LD filtering for modern data

Sample	**0.1X**	**1X**	**2X**	**10X**	**50X**
NA12877	22,104	31,957	47,934	56,615	56,625
NA12883	22,111	32,088	48,010	56,618	56,625
NA12885	22,595	32,236	48,121	56,618	56,625
NA12889	21,549	31,539	47,488	56,613	56,625
NA12890	22,726	32,442	48,420	56,616	56,625

Table 3.12: Remaining SNPs after MAF filtering for modern data

Sample	**0.1X**	**1X**	**2X**	**10X**	**50X**
NA12877	15,291	22,961	34,353	40,678	40,685
NA12883	15,380	23,149	34,482	40,678	40,685
NA12885	15,552	23,157	34,479	40,678	40,685
NA12889	14,971	22,729	34,057	40,678	40,685
NA12890	15,718	23,338	34,697	40,678	40,685

3.2.4 Relatedness estimation for modern pedigree data

The relatedness calculations were carried out with the SNPs that remained after the filtering process (**Section 2.5**). One of the alleles were selected randomly for each individual to simulate ancient DNA. After comparison with the minor allele in the reference populations set, the alleles were coded as 2 if they had the same genotype as minor allele or 0 when they were carrying the major allele. The SNPs that were missing or carrying other genotypes were excluded from the data. The pairwise kinship coefficient was calculated with the equation (2.4) in **Section 2.5** in R environment.

I also used PLINK and KING software for relatedness calculations in another set of processed modern data. In addition to down-sampling of coverages, the SNP numbers were reduced randomly to observe the effect of low coverage and

SNP number on accuracy of relatedness estimations. After transition removal, the SNP numbers were randomly reduced to 100K, 50K, 25K, 10K, 5K and 1K for all the coverages except for 0.1X. After the SNP calling and transition removal, a total of 45,440 SNPs remained for modern samples of 0.1X coverage. Therefore, its SNP number was randomly decreased to 40K, 30K, 20K, 10K, 5K and 1K.

The **Figures 3.6–3.10** show all the plots for kinship estimation of CEPH family 1463 by three methods: our approach, PLINK and KING. The PLINK software could not infer relatedness accurately in any of the five coverages. Our approach and KING were able to estimate kinship degrees for 50X and 10X coverages within the expected range. The 50X coverage data was the original version of CEPH family 1463 data. As the **Figure 3.6** shows both our approach and KING software could correctly estimate the kinship relations between first-degree, second-degree and unrelated members of the CEPH family. In our approach, first-degree pairs are correctly inferred and their 95% CI is within the expected range of (0.204-0.296) for 1st degree relatives.

There is a slight overestimation in the measured kinship coefficient of NA12883-NA12889 (second-degree) pair with our approach. The other three 2nd degree pairs' estimated kinship coefficients are in the expected range although some of the 95% CIs are noisy: overestimated for the NA12885-NA12889 pair and underestimated for the NA12883-NA12890 pair. The NA12889-NA12890 pair was correctly assigned as unrelated.

KING software could also infer kinship degrees of 50X coverage data correctly. However, as the SNP numbers is reduced randomly for each pair (lighter colors), the estimated Θ's get distorted for 2nd degree and unrelated pairs (**Figure 3.6**). Even with the original 50X data, PLINK overestimated kinship degrees to such an extent that unrelated individuals were inferred as first degree relatives.

The results for 10X coverage data (**Figure 3.7**) were similar to that of 50X coverage, our approach and KING could determine kinship degrees properly while PLINK's results were highly overestimated.

The PLINK program estimations became more inconsistent as the coverage of the data declined. Interestingly, for 2X (**Figure 3.8**) and 1X (**Figure 3.9**) coverages KING underestimated all the kinship degrees, while our approach and

PLINK overestimated the results. KING Kinship values estimated by KING at 2X coverage were negative. Across all methods, the estimations for 2X, 1X and 0.1X (**Figure 3.10**) coverages were not accurate, which was surprising. In a typical 2X coverage genomic data, there has to be enough information for accurate relatedness calculations. In the ancient samples, I was able to measure kinship degrees by using individuals that had much lower coverages.
For both our approach and KING, the accuracy of first-degree estimations was higher than that of second-degree relatives.

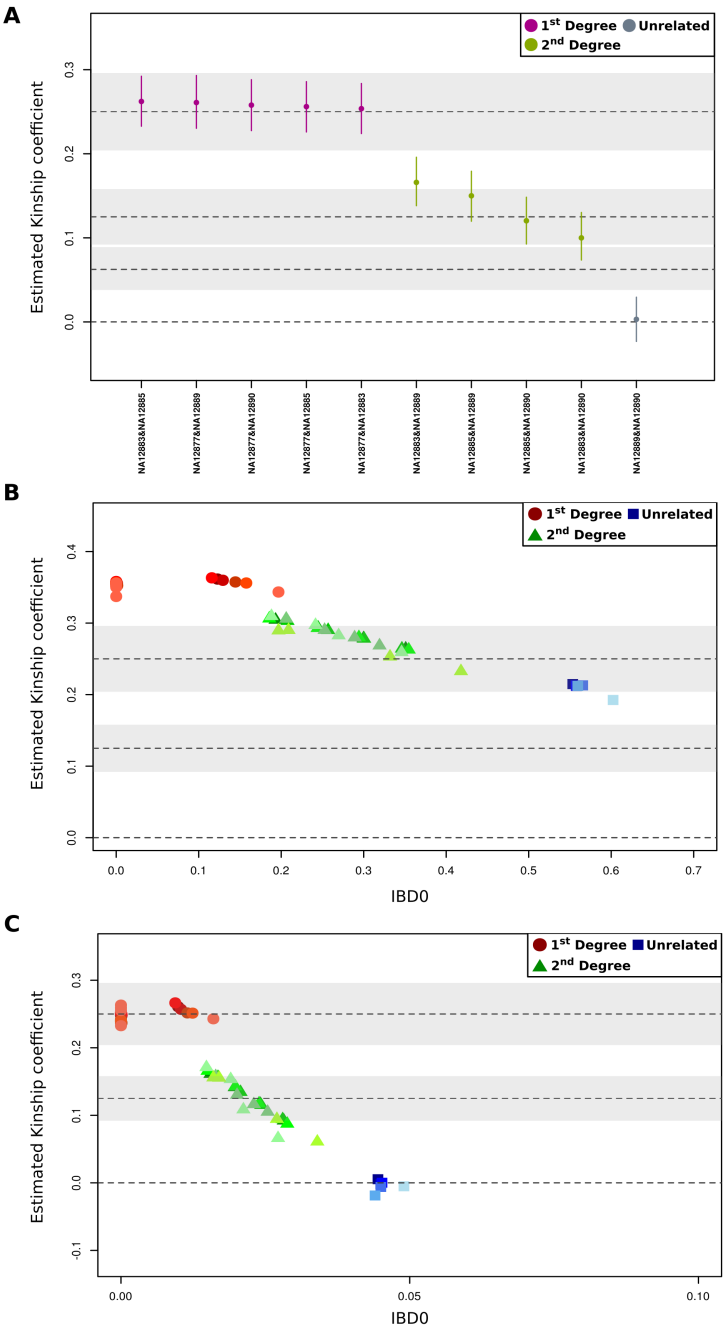

Figure 3.6: Kinship coefficient (Θ) estimation for 50X coverage data. Kinship estimation using three different methods: (A) our approach, (B) PLINk and (C) KING. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).The x-axis for (A) is the 10 possible pairwise comparisons for the five member of CEPH family, for (B) and (C) is the probability of IBD=0. (B) and (C) the red circles shows the 1st degree relatives, green triangles show 2nd degree and the blue rectangles represents the unrelated individuals. As the SNP number decreases the color intensity of the dots decrease as well.

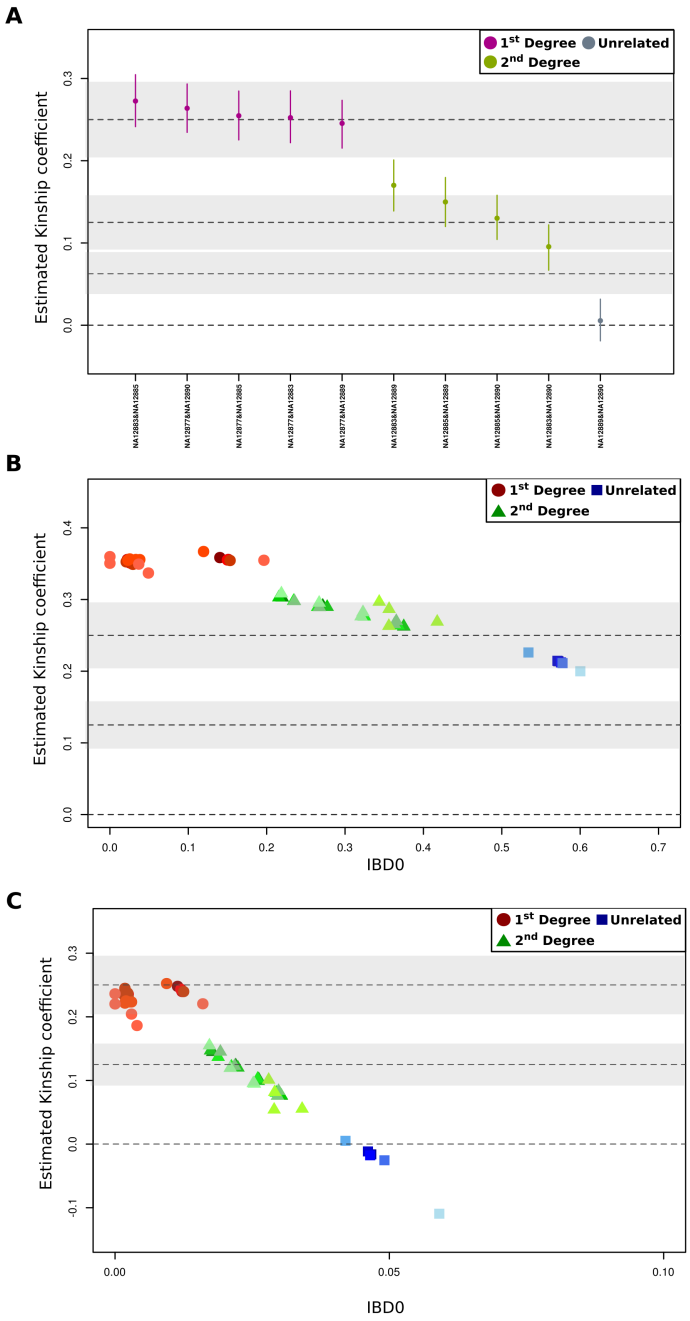

Figure 3.7: Kinship coefficient (Θ) estimation for 10X coverage data. Kinship estimation using three different methods: (A) our approach, (B) PLINk and (C) KING. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).The x-axis for (A) is the 10 possible pairwise comparisons for the five member of CEPH family, for (B) and (C) is the probability of IBD=0. (B) and (C) the red circles shows the 1st degree relatives, green triangles show 2nd degree and the blue rectangles represents the unrelated individuals. As the SNP number decreases the color intensity of the dots decrease as well.

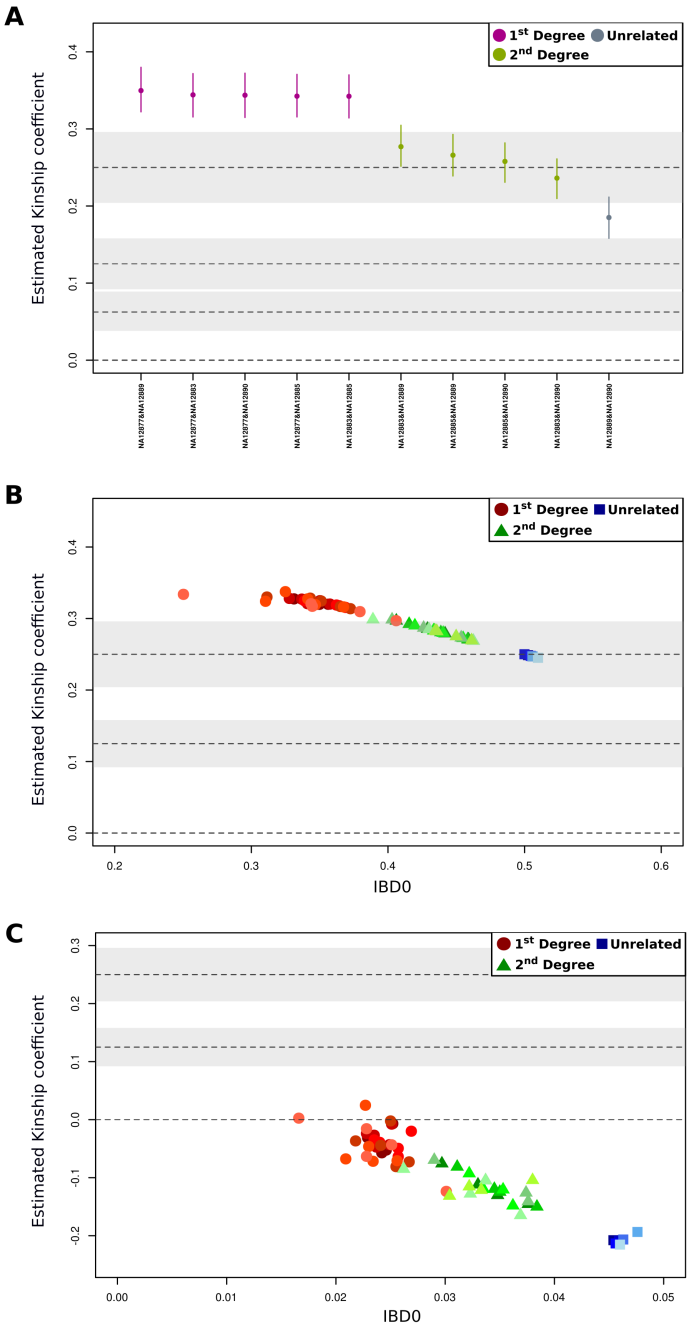

Figure 3.8: Kinship coefficient (Θ) estimation for 2X coverage data. Kinship estimation using three different methods: (A) our approach, (B) PLINk and (C) KING. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).The x-axis for (A) is the 10 possible pairwise comparisons for the five member of CEPH family, for (B) and (C) is the probability of IBD=0. (B) and (C) the red circles shows the 1st degree relatives, green triangles show 2nd degree and the blue rectangles represents the unrelated individuals. As the SNP number decreases the color intensity of the dots decrease as well.

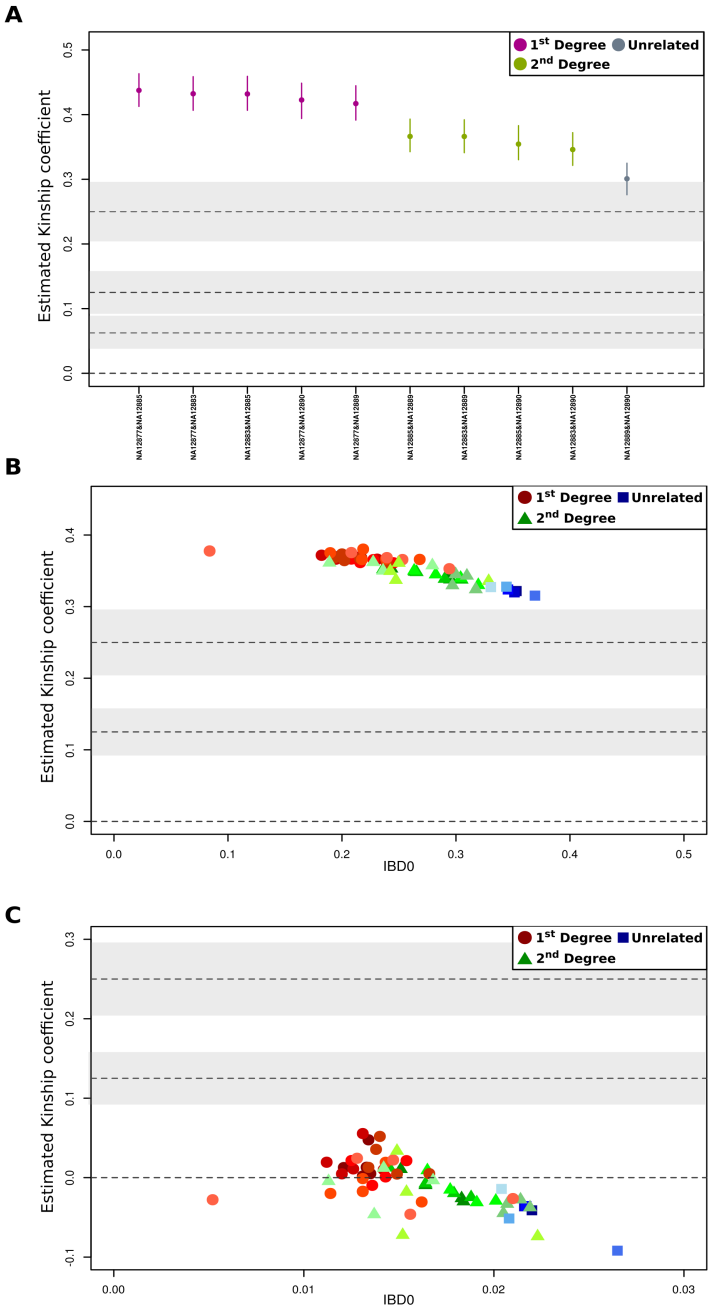

Figure 3.9: Kinship coefficient (Θ) estimation for 1X coverage data. Kinship estimation using three different methods: (A) our approach, (B) PLINk and (C) KING. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).The x-axis for (A) is the 10 possible pairwise comparisons for the five member of CEPH family, for (B) and (C) is the probability of IBD=0. (B) and (C) the red circles shows the 1st degree relatives, green triangles show 2nd degree and the blue rectangles represents the unrelated individuals. As the SNP number decreases the color intensity of the dots decrease as well.

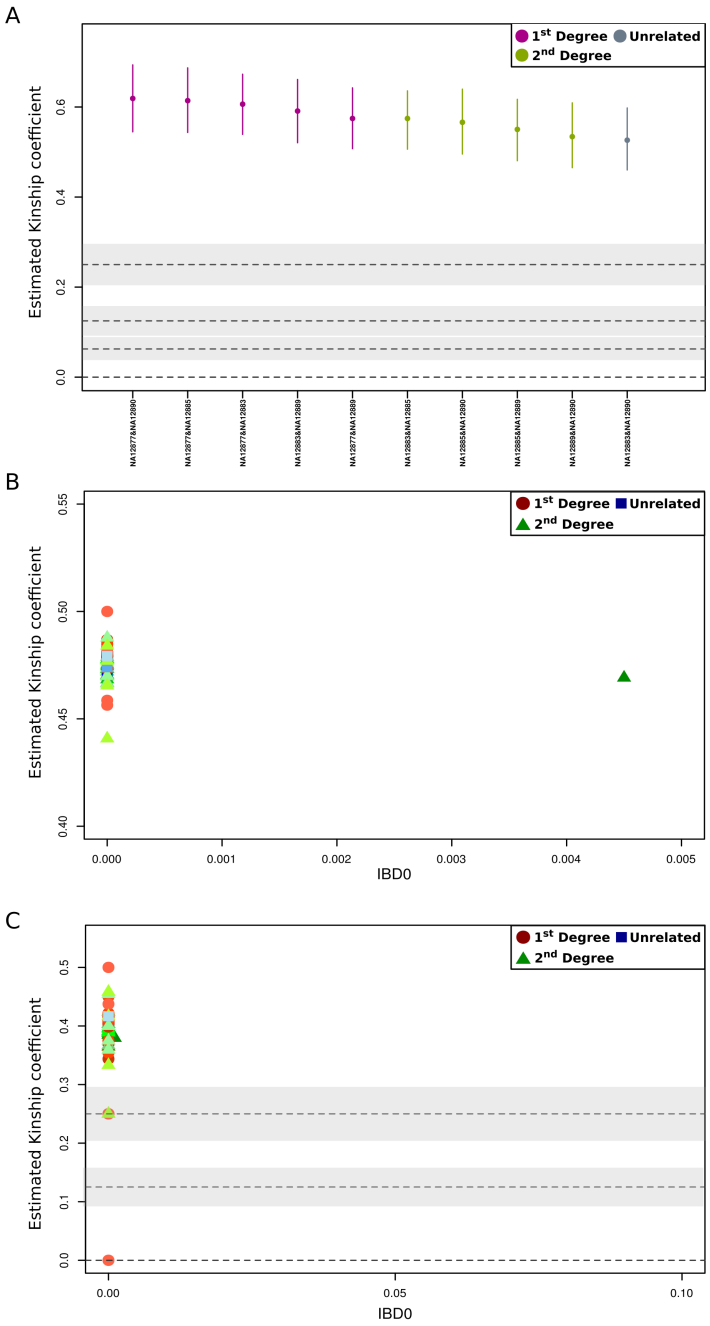

Figure 3.10: Kinship coefficient (Θ) estimation for 0.1X coverage data. Kinship estimation using three different methods: (A) our approach, (B) PLINk and (C) KING. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5).The x-axis for (A) is the 10 possible pairwise comparisons for the five member of CEPH family, for (B) and (C) is the probability of IBD=0. (B) and (C) the red circles shows the 1st degree relatives, green triangles show 2nd degree and the blue rectangles represents the unrelated individuals. As the SNP number decreases the color intensity of the dots decrease as well.

3.2.5 Statistical analysis of kinship estimation method performance using modern-day genomic data with a known pedigree

We tested the performance of our approach based on kinship estimations on a real-life family genomic data with a known pedigree. The family consisted of five first-degree related pairs, four second-degree pairs and one pair of unrelated individuals. For the statistical measures, I only used the 50X (original) and 10X (down-sampled) data (see **Appendix D, E**). The rest of the down-sampled data (2X, 1X and 0.1X) showed high overestimation in inferred kinship degrees and therefore they were not used in this analysis.

For each coverage a total of 60,000 pairwise comparisons were examined. The results are summarized in **Figure 3.11** and **3.12** for 50X coverage and 10X coverage data, respectively. Both 50X and 10X coverage data show similar outcomes. We can accurately infer kinship for first-degree pairs, while for the second-degree the rates of IR and UN increases to around 17% and 25% respectively. It became obvious that 1000 SNPs (with MAF > 0.1) are too few to make a reliable estimation about the kinship relations of individuals in this setting. In both datasets the rate of undecided (UN) pairs increase substantially for second-degree pairs across all six different SNP number sets. Therefore, we should develop a strategy to assign some of these pairs, that are mostly in the area between the expected range of first-degree and second-degree ($0.158 < \Theta < 0.204$) to their correct kinship degree.

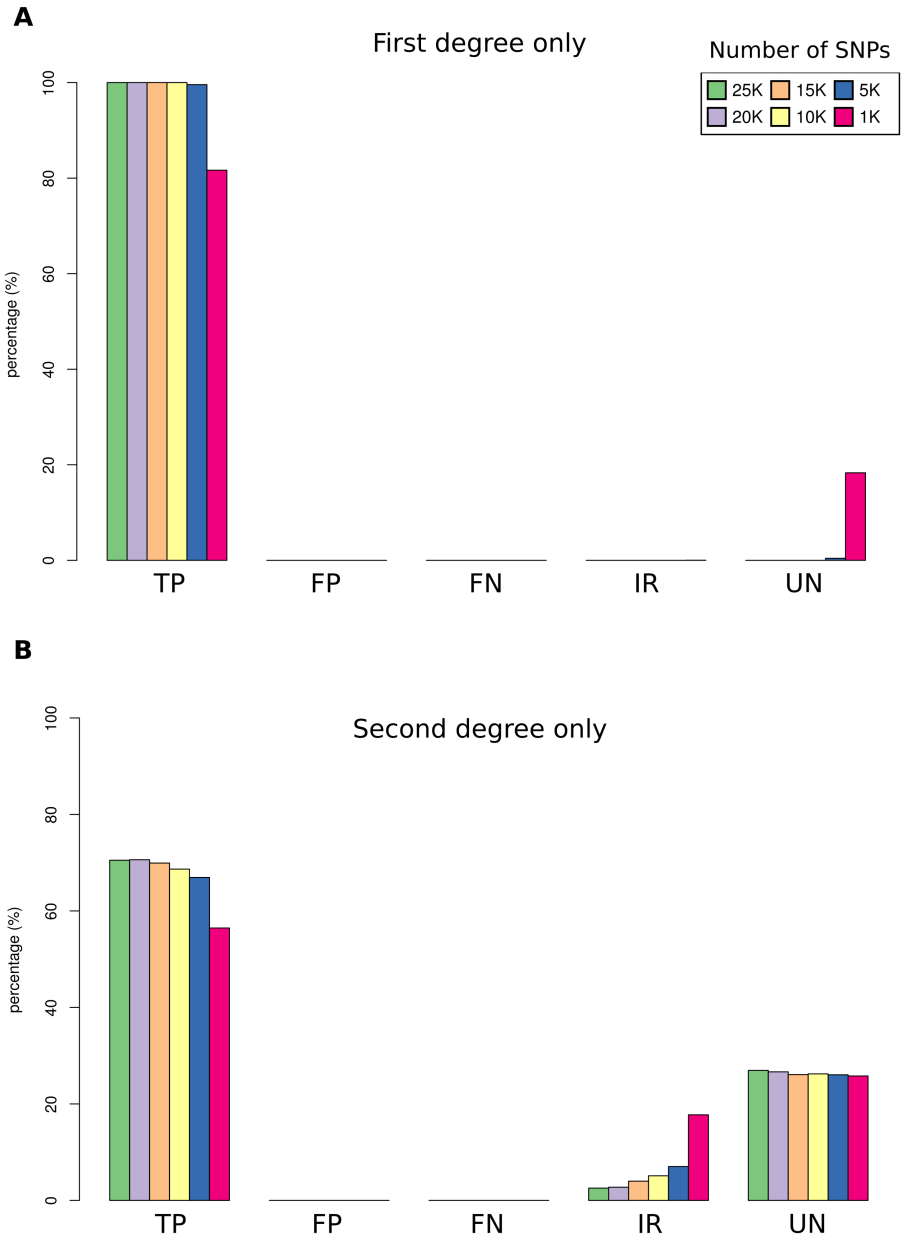

Figure 3.11: Error rates for 50X modren data. Plots demonstrate error rates for different SNP numbers. The x-axis shows the five different statistical measures, true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided". (A) For first-degree relatives only and (B) second-degree relatives only.

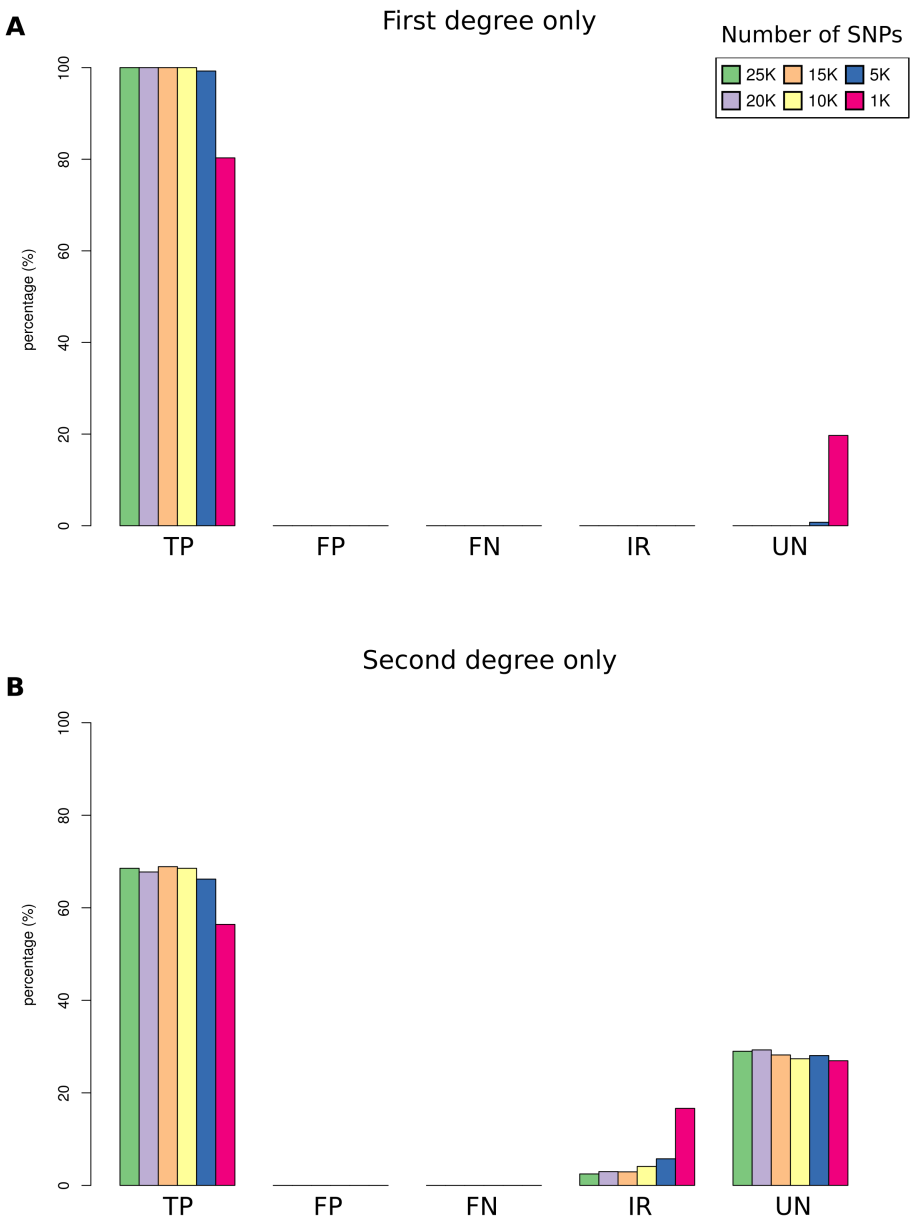

Figure 3.12: Error rates for 10X modern data. Plots demonstrate error rates for different SNP numbers. The x-axis shows the five different statistical measures, true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided". (A) For first-degree relatives only and (B) second-degree relatives only.

CHAPTER 4

DISCUSSION

Relatedness and the degree by which people are related to each other is a fundamental concept for various disciplines such as forensics, medical genetics and conservation programs. Thus, numerous distinct methods and software had been developed to answer these questions. Generally these methods were developed to infer relatedness in high-throughput sequencing data with the assumption of homogenous population structure. One example of these methods, is the ERSA 2.0 developed by Li et al. (2014) that could detect kinships up to 11th degree using perfect (diploid data that without errors) whole-genome sequence data. Others like Manichaikul et al. (2010) and Thornton et al. (2012) have developed methods for relationship estimations in structured and admixed populations. Most recently, the focus is on the kinship coefficient estimators that use low coverage (as low as 2X) data (Korneliussen & Moltke, 2015; Lipatov et al., 2015).

However, low coverage is not the only problem facing the ancient DNA studies. Ancient DNA is usually in trace amounts and degraded, for this reason special experimental and computational techniques are required for its analysis (Skoglund et al., 2014).

Estimation of kinship degrees among ancient individuals could shed light on culture and social structure of prehistoric human societies. For a long time, archaeologists and anthropologists have attempted to answer this question using indirect methods. I propose a new approach to estimate kinship patterns in low coverage ancient data.

For this study, I used one of the kinship coefficient estimation methods suggested

by Speed and Balding (2014), based on genetic similarity matrix calculation that is centered and scaled for population minor allele frequency, and tested it on ancient samples. To achieve this, I used a series of methods and SNP filtering processes to select the lowest possible informative set of SNPs to be used in kinship coefficient estimations.

I applied our approach to both a simulated pedigree created using ancient genome data (Mathieson et al., 2015) and a modern genome dataset with known relationship (Eberle et al., 2017), to test its performance. In the first case, eight ancient individuals were used to simulate a four generation in-silico pedigree. Simulations are beneficial for investigation of distinct real-life processes in a controlled manner, however they cannot completely represent the real-life situations. Therefore, I used real ancient samples and only simulated the relationships between them to preserve the authenticity of aDNA features. The modern data with known relationships was used to test the performance of our approach on empirical data that would have more realistic error structure. In addition, I down-sampled the original data (50X coverage) to 10X, 2X, 1X and 0.1X to make them mirror the low coverage nature of aDNA samples.

The overall performance of our approach, for both the ancient in-silico pedigree and modern data was better compared to results obtained from published software such as KING and PLINK. I could accurately infer third-degree relatives in simulated ancient pedigree with as little as 5,000 overlapping SNPs. The accuracy of the estimation dropped to 60% with the 1,000 SNP set, hence, its results should be used with caution. For this SNP set, the IR rate increased to 26.4%, FP to 16.4% and UN to 14.7%. However the biggest problem was the rise in FP rate, that made its usage unfeasible. In many settings, the cost of classifying unrelated individuals as related may be much higher than assigning related individuals incorrectly to the third degree relatives class. For this reason, it is preferable to use higher number of SNPs to avoid false positives.

Considering the 50X and 10X coverages of modern data, our approach performed better than KING and PLINK in kinship relation estimations. Even for the second degree relatives, the FP or FN rate were zero. The decline in accuracy of estimated second-degree pairs was due to the increase in undecided (UN) rate. When the coverages of the samples were down-sampled to 2X and below, all

three methods produced inaccurate results. Our approach and the PLINK software overestimated the kinship coefficient values for all the pairwise comparisons while KING underestimated them.

The most critical step of our approach is the selection of a reference population set, which were modern populations that operated as surrogates for ancient samples to estimate their population level features. The reference population set was used in SNP filtering steps and in MAF calculations. The equation (2.4) implemented for kinship coefficient estimation uses MAF to resolve the level of background relatedness in the population. Moreover, it would give extra weight for the rare shared alleles, as they are more informative. The choice of reference population is critical for the estimation of kinship coefficient with our approach. Selection of genetically distant populations compared to the test individuals as the reference should result in overestimation of kinship degrees.

However, I do not think the observed overestimation in pairwise kinship coefficients was due to incorrect reference population selection, because the same set of populations were also used for calculation of 50X and 10X data and in this case we did not observe any bias. I used samtools for down-sampling process that randomly extracts a fraction of the reads from original data. One possible explanation to this observed pattern could be that samtools down-sampling process was not random. I down-sampled the modern data to mimic limited ancient data. However, the artificially down-sampled data could be different than the low coverage ancient data possibly due to a bug in my code or the software.

4.1 Limitations and possible improvements

1. In the analysis of this study, I used haploidized ancient data. Because ancient DNA data is usually at low coverage, it is common to randomly haploidize all the heterozygous/diploid SNP calls, which I repeated here. For allele frequency-based analysis (such as PCA and f-statistics), the modern data is haploidized as well to make the data comparable. Many population genetics analyses that require complete genotypes or haplotypes cannot be used because of this loss in data content. For our method,

this means that we would never have comparison pairs with IBD=1. There are two possible approaches to overcome this problem. First, is to modify these methods in a way to accommodate the uncertainties. Second, imputation could be used to improve the content of the data and increase overlapping SNP numbers. Using genotype likelihoods (Korneliussen, Albrechtsen, & Nielsen, 2014) could be yet another alternative.

2. I used PCA analysis and modern samples to determine the reference populations which had the closest genetic affinity to the ancient samples. These modern populations were used as a reference to predict the population level allele frequencies of ancient samples. However the composition of human populations changes overtime because of migration, drift and selection. For example, the ancient hunter-gatherers' genetic variation is different than any modern population so that they form a separate cluster in PCA analysis. Modern populations' genetic variation might not always accurately represent ancient samples' genetic composition. With the increase in aDNA studies and publications, it would be possible to use ancient populations as the reference which improve the accuracy of the calculations.

3. In this study, I used a simple four-generation simulated pedigree for kinship estimation calculations. However, the real-life scenarios are much more complicated than my simulated pedigree. It would be better to construct different, complex pedigrees with inbreeding to better assess the performance of our approach in real-life situations.

4. The next step after using different simulated pedigrees for kinship calculations, would be to test our approach's performance on the inferred related ancient samples as reported by Mathieson et al. (2015).

5. Despite the efforts to improve our approach's kinship estimations, there is a limit to this progress due to the distinctive properties of aDNA. For some cases, it might be difficult to accurately identify the incorrectly assigned relate pairs. Another factor contributing to this problem is that the proportion of overlapping expected kinship coefficient values increase for more distant relatives. As Monroy Kuhn et al. (2017) suggests, it

might be useful to combine genetic data with radiocarbon dating of ancient samples and their uniparental markers such as mitochondrial and Y-chromosome haplogroups to eliminate some of the possible inaccurate relatedness patterns.

6. Likewise, the rate of undecided pairs increased in the second and third-degree relatives which decreases the sensitivity of our estimations. It would be essential to establish a plan to accurately incorporate these pairs into their correct relatedness degree.

7. The observed overestimation of kinship values for down-sampled modern data (2X, 1X and 0.1X coverages) should be investigated to identify the underlying causes.

8. Also, the effect of MAF filtering should be examined in detail using different cut offs.

CHAPTER 5

CONCLUSION

Traditional methods for estimating kinship patterns are ineffective against degraded, low quality characteristics of aDNA. Identifying familial relations among ancient individuals are essential to understand the complex social organization of prehistoric civilizations.

To achieve this, I based the methodology of our approach on a SNP-based measure of relatedness suggested by Speed and Balding (2014). I performed additional SNP filtering steps to adjust it for features of aDNA. I tested the performance of our approach on a four-generation in-silico pedigree simulated from ancient samples and a modern three-generation family pedigree with varying genomic coverages. The primary findings of this study are:

- The overall performance of our approach is better than that of published software KING and PLINK.
- By using 5,000 SNPs our approach could successfully detect the first-degree, second-degree and third-degree relatives with accuracy of 99.8%, 97.2% and 92.2% respectively.
- For the empirical modern data, the accuracy of estimation is 99.2% for first-degree pairs and around 70% for second-degree pairs.

Overall, the initial results demonstrated that our approach could successfully infer the second-degree relatives in both the simulated and real data. The second-degree relatives which are grandparent-grandchild and half-siblings, are

sufficient to characterize members of a core family who shared the same burial site. This knowledge could help us better understand the social dynamics of ancient populations.

REFERENCES

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., ... Bunce, M. (2012, dec). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4724–4733. Retrieved from `http://rspb.royalsocietypublishing.org/content/279/1748/4724http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2012.1745` doi: 10.1098/rspb.2012.1745

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved from `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002, apr). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *3*(4), 299–309. Retrieved from `http://www.nature.com/doifinder/10.1038/nrg777` doi: 10.1038/nrg777

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., ... Abecasis, G. R. (2015, sep). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. Retrieved from `http://www.nature.com/doifinder/10.1038/nature15393` doi: 10.1038/nature15393

Bar-Yosef, O. (2001). From sedentary foragers to village hier- archies: the emergence of social institutions. *Proceedings of the British Academy*, *101*, 1–38.

Behjati, S., & Tarpey, P. S. (2013, dec). What is next generation sequencing? *Archives of disease in childhood - Education & practice edition*, *98*(6), 236–238. Retrieved from `http://ep.bmj.com/lookup/doi/10.1136/archdischild-2013-304340` doi: 10.1136/archdischild-2013-304340

Belfer-Cohen, A., & Goring-Morris, A. N. (2011, oct). Becoming Farmers: The Inside Story. *Current anthropology*, *52*(S4),

S209–S220. Retrieved from http://www.jstor.org/stable/10.1086/658861http://www.journals.uchicago.edu/doi/10.1086/658861http://www.jstor.org/stable/10.1086/658861?ref=no-x-route:7fb4e350474ebacfbd0172f47c9d236c{%}5Cnpapers://d2952c50-9509-4ba2-9a03-22fbc04267d4/Paper/p33326 doi: 10.1086/658861

Bishop, R. (2008, jun). In the Grand Scheme of Things: An Exploration of the Meaning of Genealogical Research. *The Journal of Popular Culture*, *41*(3), 393–412. Retrieved from http://doi.wiley.com/10.1111/j.1540-5931.2008.00527.x doi: 10.1111/j.1540-5931.2008.00527.x

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prufer, K., ... Paabo, S. (2007, sep). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, *104*(37), 14616–14621. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1976210{&}tool=pmcentrez{&}rendertype=abstracthttp://www.pnas.org/cgi/doi/10.1073/pnas.0704665104 doi: 10.1073/pnas.0704665104

Byrd, B. F. (2005, sep). Reassessing the Emergence of Village Life in the Near East. *Journal of Archaeological Research*, *13*(3), 231–290. Retrieved from http://link.springer.com/10.1007/s10814-005-3107-2 doi: 10.1007/s10814-005-3107-2

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., ... Bustamante, C. D. (2013, nov). Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics*, *93*(5), 852–864. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S000292971300459X doi: 10.1016/j.ajhg.2013.10.002

Chistiakov, D. A., Hellemans, B., & Volckaert, F. A. (2006, may). Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. *Aquaculture*, *255*(1-4), 1–29. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0044848605007313 doi: 10.1016/j.aquaculture.2005.11.031

Cockerham, C. C. (1971, oct). Higher order probability functions of identity of allelles by descent. *Genetics*, *69*(2), 235–46. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5135830http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1212700 doi: Article

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011, aug). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330 doi: 10.1093/bioinformatics/btr330

Devlin, B., & Risch, N. (1995, sep). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, *29*(2), 311–322. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0888754385790039 doi: 10.1006/geno.1995.9003

Diamond, J., & Bellwood, P. (2003, apr). Farmers and their languages: the first expansions. *Science (New York, N.Y.)*, *300*(5619), 597–603. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1078208http://www.ncbi.nlm.nih.gov/pubmed/12714734 doi: 10.1126/science.1078208

Dodds, K. G., McEwan, J. C., Brauning, R., Anderson, R. M., van Stijn, T. C., Kristjánsson, T., & Clarke, S. M. (2015, dec). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, *16*(1), 1047. Retrieved from http://biorxiv.org/content/early/2015/08/24/025379.abstracthttp://www.biomedcentral.com/1471-2164/16/1047 doi: 10.1186/s12864-015-2252-3

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., ... Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, *27*(1), 157–164. doi: 10.1101/gr.210500.116

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013, oct). Robust Demographic Inference from Genomic and SNP Data.

PLoS Genetics, *9*(10), e1003905. Retrieved from http://dx.plos.org/10.1371/journal.pgen.1003905 doi: 10.1371/journal.pgen.1003905

Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H., Ripoli, M. V., ... Giovambattista, G. (2013). Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genetics and Molecular Biology*, *36*(2), 185–191. Retrieved from http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1415-47572013000200008{&}lng=en{&}tlng=en doi: 10.1590/S1415-47572013000200008

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., ... Stewart, J. (2007, oct). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. Retrieved from http://www.nature.com/doifinder/10.1038/nature06258 doi: 10.1038/nature06258

Gillespie, J. H. (2004). *Population genetics : a concise guide*. Baltimore, Md.: Johns Hopkins University Press.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... Paabo, S. (2010, may). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. Retrieved from http://www.sciencemag.org/content/328/5979/710.abstracthttp://www.sciencemag.org/cgi/doi/10.1126/science.1188021 doi: 10.1126/science.1188021

Günther, T., & Jakobsson, M. (2016, dec). Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective. *Current Opinion in Genetics & Development*, *41*, 115–123. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0959437X16301150 doi: 10.1016/j.gde.2016.09.004

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., ... Reich, D. (2015, mar). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, *522*(7555), 207–211. Retrieved from http://www.nature.com/doifinder/10.1038/nature14317 doi: 10.1038/nature14317

Harris, D. L. (1964, dec). Genotypic Covariances between Inbred Relatives. *Ge-*

netics, *50*(5), 1319–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14239792http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1210738

Heaton, M. P., Harhay, G. P., Bennett, G. L., Stone, R. T., Grosse, W. M., Casas, E., ... Laegreid, W. W. (2002, may). Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. *Mammalian Genome*, *13*(5), 272–281. Retrieved from http://link.springer.com/10.1007/s00335-001-2146-3 doi: 10.1007/s00335-001-2146-3

Hodder, I. (2007, sep). Çatalhöyük in the Context of the Middle Eastern Neolithic. *Annual Review of Anthropology*, *36*(1), 105–120. Retrieved from http://www.annualreviews.org/doi/10.1146/annurev.anthro.36.081406.094308 doi: 10.1146/annurev.anthro.36.081406.094308

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001, may). Ancient DNA. *Nature Reviews Genetics*, *2*(5), 353–359. Retrieved from http://www.nature.com/nrg/journal/v2/n5/full/nrg0501{_}353a.html{%}5Cnpapers://d2952c50-9509-4ba2-9a03-22fbc04267d4/Paper/p30060http://www.nature.com/doifinder/10.1038/35072071 doi: 10.1038/35072071

Jorde, L. (2001, jul). Consanguinity and Prereproductive Mortality in the Utah Mormon Population. *Human Heredity*, *52*(2), 61–65. Retrieved from http://www.karger.com/?doi=10.1159/000053356 doi: 10.1159/000053356

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014, dec). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. Retrieved from http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0356-4 doi: 10.1186/s12859-014-0356-4

Korneliussen, T. S., & Moltke, I. (2015, aug). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*(24), btv509. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/

10.1093/bioinformatics/btv509 doi: 10.1093/bioinformatics/btv509

Kılınç, G. M., Omrak, A., Özer, F., Günther, T., Büyükkarakaya, A. M., Bıçakçı, E., ... Götherström, A. (2016, oct). The Demographic Development of the First Farmers in Anatolia. *Current Biology*, *26*(19), 2659–2666. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0960982216308508 doi: 10.1016/j.cub.2016.07.057

Larsen, C. S. (1995, jan). Biological Changes in Human Populations with Agriculture. *Annual Review of Anthropology*, *24*(1), 185–213. Retrieved from http://anthro.annualreviews.org/cgi/doi/10.1146/annurev.anthro.24.1.185 doi: 10.1146/annurev.anthro.24.1.185

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., ... Reich, D. (2016, jul). Genomic insights into the origin of farming in the ancient Near East. *Nature*, *536*(7617), 419–424. Retrieved from http://www.nature.com/doifinder/10.1038/nature19310 doi: 10.1038/nature19310

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., ... Krause, J. (2014, sep). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, *513*(7518), 409–13. Retrieved from http://biorxiv.org/content/early/2013/12/23/001552.abstracthttp://www.ncbi.nlm.nih.gov/pubmed/25230663http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4170574 doi: 10.1038/nature13673

Lewontin, R. C. (1964, jan). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, *49*(1), 49–67. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17248194http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1210557

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009, aug). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352 doi: 10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010, sep). A survey of sequence alignment algorithms for

next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473–483. Retrieved from https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbq015 doi: 10.1093/bib/bbq015

Lipatov, M., Sanjeev, K., Patro, R., & Veeramah, K. (2015). Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. *bioRxiv*, 023374. Retrieved from http://biorxiv.org/content/early/2015/07/29/023374.abstract doi: https://doi.org/10.1101/023374

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., & Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC genomics*, *13*(8), S8. Retrieved from http://www.biomedcentral.com/1471-2164/13/S8/S8http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3535703{&}tool=pmcentrez{&}rendertype=abstracthttp://www.ncbi.nlm.nih.gov/pubmed/23281772http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3 doi: 10.1186/1471-2164-13-S8-S8

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010, nov). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq559 doi: 10.1093/bioinformatics/btq559

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., ... Reich, D. (2015, nov). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503. Retrieved from http://dx.doi.org/10.1038/nature16152http://www.nature.com/doifinder/10.1038/nature16152 doi: 10.1038/nature16152

Mathieson, I., & McVean, G. (2014, aug). Demography and the Age of Rare Variants. *PLoS Genetics*, *10*(8), e1004528. Retrieved from http://dx.plos.org/10.1371/journal.pgen.1004528 doi: 10.1371/journal.pgen.1004528

Monroy Kuhn, J. M., Jakobsson, M., & Günther, T. (2017). Estimating ge-

netic kin relationships in prehistoric populations. *bioRxiv*. Retrieved from http://www.biorxiv.org/content/early/2017/06/23/100297

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011, jun). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–451. Retrieved from http://www.nature.com/doifinder/10.1038/nrg2986 doi: 10.1038/nrg2986

Omrak, A., Günther, T., Valdiosera, C., Svensson, E. M., Malmström, H., Kiesewetter, H., ... Götherström, A. (2016, jan). Genomic Evidence Establishes Anatolia as the Source of the European Neolithic Gene Pool. *Current Biology*, *26*(2), 270–275. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S096098221501516X doi: 10.1016/j.cub.2015.12.019

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., ... Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. doi: 10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006, dec). Population Structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17194218http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1713260http://dx.plos.org/10.1371/journal.pgen.0020190 doi: 10.1371/journal.pgen.0020190

Pearson, J., Grove, M., Özbek, M., & Hongo, H. (2013, jun). Food and social complexity at Çayönü Tepesi, southeastern Anatolia: Stable isotope evidence of differentiation in diet according to burial practice and sex in the early Neolithic. *Journal of Anthropological Archaeology*, *32*(2), 180–189. Retrieved from http://dx.doi.org/10.1016/j.jaa.2013.01.002http://linkinghub.elsevier.com/retrieve/pii/S0278416513000044 doi: 10.1016/j.jaa.2013.01.002

Pemberton, J. (2008, mar). Wild pedigrees: the way forward. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1635), 613–621. Retrieved from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2007.1531 doi: 10.1098/rspb.2007.1531

Phillips, C., Fang, R., Ballard, D., Fondevila, M., Harrison, C., Hyland, F., ... Schneider, P. (2007, jun). Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Science International: Genetics*, *1*(2), 180–185. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1872497307000610 doi: 10.1016/j.fsigen.2007.02.007

Pickrell, J. K., & Pritchard, J. K. (2012, nov). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, *8*(11), e1002967. Retrieved from http://dx.plos.org/10.1371/journal.pgen.1002967 doi: 10.1371/journal.pgen.1002967

Pilloud, M. A., & Larsen, C. S. (2011, aug). "Official" and "practical" kin: Inferring social and community structure from dental phenotype at Neolithic Çatalhöyük, Turkey. *American Journal of Physical Anthropology*, *145*(4), 519–530. Retrieved from http://doi.wiley.com/10.1002/ajpa.21520 doi: 10.1002/ajpa.21520

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007, sep). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, *81*(3), 559–75. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524http://www.ncbi.nlm.nih.gov/pubmed/17701901http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1950838 doi: 10.1086/519795

Ramel, C. (1997, jun). Mini- and Microsatellites. *Environmental Health Perspectives*, *105*(SUPPL. 4), 781. Retrieved from http://www.jstor.org/stable/3433284?origin=crossref doi: 10.2307/3433284

Ringnér, M. (2008, mar). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–4. Retrieved from http://www.nature.com/doifinder/10.1038/nbt0308-303http://www.ncbi.nlm.nih.gov/pubmed/18327243 doi: 10.1038/nbt0308-303

Rohrer, G. A., Freking, B. A., & Nonneman, D. (2007, jun). Single nucleotide polymorphisms for pig identification and parentage exclusion. *Animal Genetics*, *38*(3), 253–258. Retrieved from http://doi.wiley.com/10.1111/

j.1365-2052.2007.01593.x doi: 10.1111/j.1365-2052.2007.01593.x

Seitz, A., & Nieselt, K. (2017, apr). Improving ancient DNA genome assembly. *PeerJ*, *5*, e3126. Retrieved from https://doi.org/10.7287/peerj.preprints.2383v1https://peerj.com/articles/3126 doi: 10.7717/peerj.3126

Shapiro, B., & Hofreiter, M. (2014, jan). A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. *Science*, *343*(6169), 1236573–1236573. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24458647http://www.sciencemag.org/cgi/doi/10.1126/science.1236573 doi: 10.1126/science.1236573

Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., ... Jakobsson, M. (2012, apr). Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science*, *336*(6080), 466–469. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1216304 doi: 10.1126/science.1216304

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014, feb). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229–2234. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24469802http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3926038http://www.pnas.org/lookup/doi/10.1073/pnas.1318934111 doi: 10.1073/pnas.1318934111

Slatkin, M. (2008, jun). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. Retrieved from http://www.nature.com/doifinder/10.1038/nrg2361 doi: 10.1038/nrg2361

Speed, D., & Balding, D. J. (2014, nov). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, *16*(1), 33–44. Retrieved from http://dx.doi.org/10.1038/nrg3821http://www.nature.com/doifinder/10.1038/nrg3821 doi: 10.1038/nrg3821

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012, dec).

Improved Heritability Estimation from Genome-wide SNPs. *The American Journal of Human Genetics*, *91*(6), 1011–1021. Retrieved from http://dx.doi.org/10.1016/j.ajhg.2012.10.010http://linkinghub.elsevier.com/retrieve/pii/S0002929712005332 doi: 10.1016/j.ajhg.2012.10.010

Stoneking, M., & Krause, J. (2011, aug). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, *12*(9), 603–614. Retrieved from http://www.nature.com/nrg/journal/v12/n9/full/nrg3029.html{%}5Cnhttp://www.nature.com/nrg/journal/v12/n9/pdf/nrg3029.pdfhttp://www.nature.com/doifinder/10.1038/nrg3029 doi: 10.1038/nrg3029

Theunert, C., Racimo, F., & Slatkin, M. (2017). Joint Estimation of Relatedness Coefficients and Allele Frequencies from Ancient Samples. *Genetics*, *206*(2). Retrieved from http://www.genetics.org/content/206/2/1025.article-info

Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J. M., Pertoldi, C., Kristensen, T. N., ... Bendixen, C. (2009, oct). Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. *Heredity*, *103*(4), 326–332. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19623210http://www.nature.com/doifinder/10.1038/hdy.2009.73 doi: 10.1038/hdy.2009.73

Wang, J. (2016, feb). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, *107*(xxxx), 4–13. Retrieved from http://dx.doi.org/10.1016/j.tpb.2015.08.006http://linkinghub.elsevier.com/retrieve/pii/S0040580915000842 doi: 10.1016/j.tpb.2015.08.006

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006, oct). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, *7*(10), 771–780. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16983373http://www.nature.com/doifinder/10.1038/nrg1960 doi: 10.1038/nrg1960

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*,

20, 557–585.

Yao, F., Coquery, J., & Lê Cao, K.-A. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics*, *13*(1), 24. Retrieved from http://www.biomedcentral.com/1471-2105/13/24http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-24 doi: 10.1186/1471-2105-13-24

Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., ... Morton, N. E. (2004, dec). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proceedings of the National Academy of Sciences*, *101*(52), 18075–18080. Retrieved from http://www.pnas.org/content/101/52/18075.long{%}5Cnpapers2://publication/doi/10.1073/pnas.0408251102http://www.pnas.org/cgi/doi/10.1073/pnas.0408251102 doi: 10.1073/pnas.0408251102

APPENDIX A

SELECTION OF UNRELATED ANCIENT INDIVIDUALS

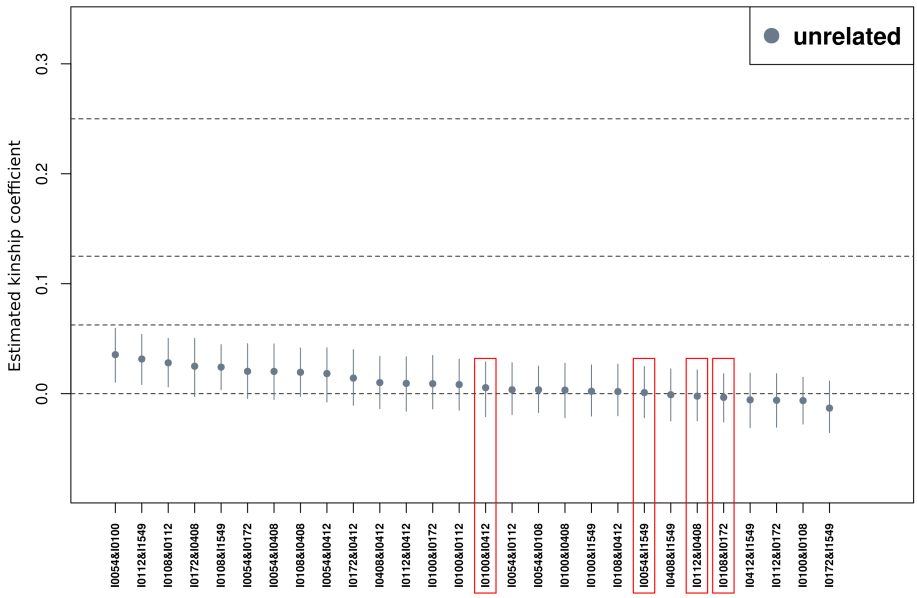

Figure A.1: Kinship coefficient (Θ) estimation for real ancient samples. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The red boxes show the four pairs (8 individuals) that were selected for in-silico pedigree formation.

APPENDIX B

WORKFLOW OF SNP FILTERING STEPS

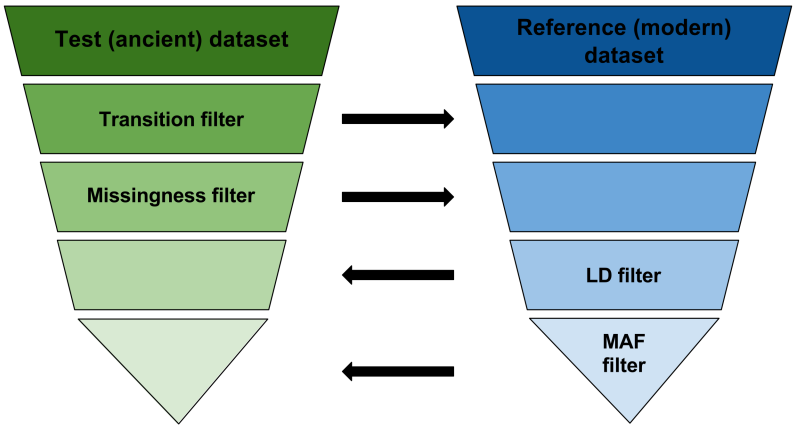

Figure B.1: Both the test (ancient) and the reference populations (modern) datasets are used for SNP filtering. The transition removal and missingness ($>$ 50%) filters are performed on test (ancient) dataset while LD (> 0.4) and MAF (< 0.1) filters are preformed on reference population dataset. After each filtering step, the same set of SNPs are filtered in the other dataset.

APPENDIX C

STATISTICAL ANALYSIS AND ERROR RATES FOR ANCIENT SAMPLES

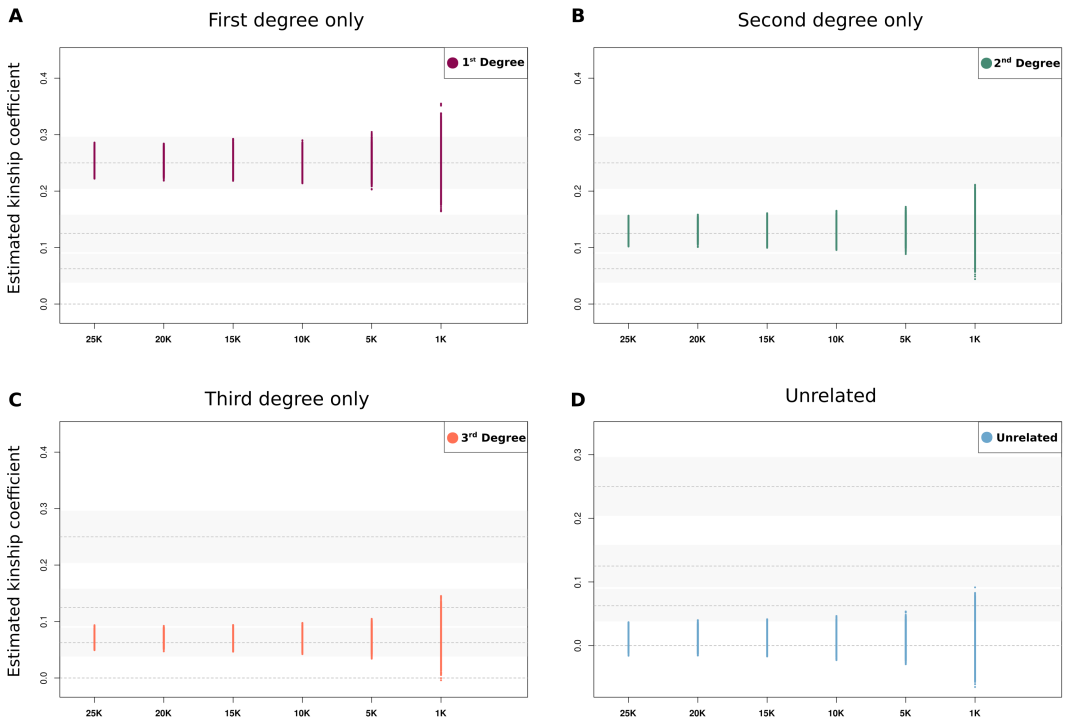

Figure C.1: Standard error calculations of kinship coefficient (Θ) for ancient samples. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5). The x-axis is the six different SNP numbers (25K, 20K, 15K,10K, 5K and 1K) used for Θ calculations. (A) For first-degree relatives only, (B) second-degree relatives only, (C) for third-degree relatives and (D) for unrelated individual pairs.

Table C.1: Error rates for first, second and third-degree related pairs. The five different statistical measures are true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided".

Statistical Measures	**First Degree**	**SecondDegree**	**ThirdDegree**
TP_25K	100.0	100.0	99.7
TP_20K	100.0	99.9	99.5
TP_15K	100.0	99.8	98.3
TP_10K	100.0	99.5	97.1
TP_5K	99.8	97.2	92.2
TP_1K	84.6	68.9	58.8
FP_25K	0.0	0.0	0.0
FP_20K	0.0	0.0	0.1
FP_15K	0.0	0.0	0.2
FP_10K	0.0	0.0	0.3
FP_5K	0.0	0.0	1.8
FP_1K	0.0	0.0	16.4
FN_25K	0.0	0.0	0.0
FN_20K	0.0	0.0	0.0
FN_15K	0.0	0.0	0.0
FN_10K	0.0	0.0	0.0
FN_5K	0.0	0.0	0.0
FN_1K	0.0	0.0	0.0
IR_25K	0.0	0.0	0.1
IR_20K	0.0	0.0	0.1
IR_15K	0.0	0.0	0.2
IR_10K	0.0	0.0	1.2
IR_5K	0.0	0.0	4.8
IR_1K	0.0	8.8	26.4
UN_25K	0.0	0.0	0.2
UN_20K	0.0	0.1	0.4
UN_15K	0.0	0.2	1.5
UN_10K	0.0	0.5	1.7
UN_5K	0.2	2.7	2.9
UN_1K	15.4	22.3	14.7

APPENDIX D

STATISTICAL ANALYSIS AND ERROR RATES FOR 50X COVERAGE MODERN DATA

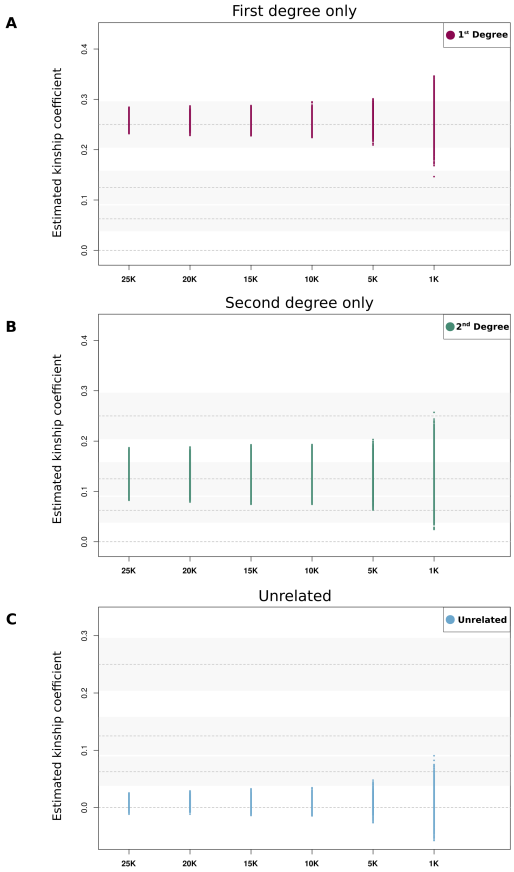

Figure D.1: Standard error calculations of kinship coefficient (Θ) for 50X coverage modern data. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5). The x-axis is the six different SNP numbers (25K, 20K, 15K,10K, 5K and 1K) used for Θ calculations. (A) For first-degree relatives only, (B) second-degree relatives only and (C) for unrelated individual pairs.

Table D.1: Error rates for first and second-degree related pairs. The five different statistical measures are true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided".

Statistical Measures	**First Degree**	**Second Degree**
TP_25K	100.0	70.5
TP_20K	100.0	70.6
TP_15K	100.0	69.9
TP_10K	100.0	68.7
TP_5K	99.6	66.9
TP_1K	81.7	56.5
FP_25K	0.0	0.0
FP_20K	0.0	0.0
FP_15K	0.0	0.0
FP_10K	0.0	0.0
FP_5K	0.0	0.0
FP_1K	0.0	0.0
FN_25K	0.0	0.0
FN_20K	0.0	0.0
FN_15K	0.0	0.0
FN_10K	0.0	0.0
FN_5K	0.0	0.0
FN_1K	0.0	0.0
IR_25K	0.0	2.6
IR_20K	0.0	2.7
IR_15K	0.0	4.0
IR_10K	0.0	5.1
IR_5K	0.0	7.0
IR_1K	0.0	17.7
UN_25K	0.0	26.9
UN_20K	0.0	26.6
UN_15K	0.0	26.1
UN_10K	0.0	26.2
UN_5K	0.4	26.0
UN_1K	18.3	25.8

APPENDIX E

STATISTICAL ANALYSIS AND ERROR RATES FOR 10X COVERAGE MODERN DATA

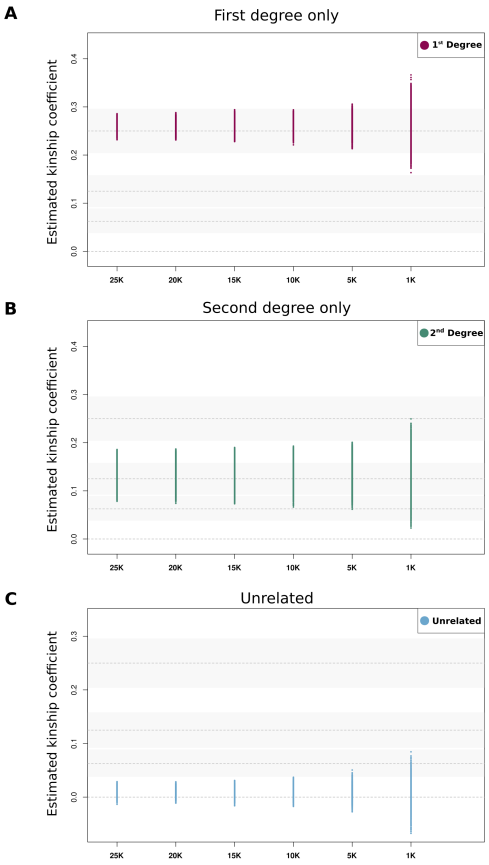

Figure E.1: Standard error calculations of kinship coefficient (Θ) for 10X coverage modern data. Dashed lines illustrate the theoretical Θ value for 1st, 2nd, 3rd degree and unrelated individuals. The gray shaded areas are the 95% CI for Θ (Table 2.5). The x-axis is the six different SNP numbers (25K, 20K, 15K,10K, 5K and 1K) used for Θ calculations. (A) For first-degree relatives only, (B) second-degree relatives only and (C) for unrelated individual pairs.

Table E.1: Error rates for first and second-degree related pairs. The five different statistical measures are true positive (TP), false positive (FP), false negative (FN), incorrectly related (IR) and undecided (UN). Pairs assigned to the wrong relatedness degree "incorrectly related", the pairs in between expected relatedness degrees "undecided".

Statistical Measures	**First Degree**	**Second Degree**
TP_25K	100.0	68.5
TP_20K	100.0	67.7
TP_15K	100.0	68.9
TP_10K	100.0	68.5
TP_5K	99.3	66.2
TP_1K	80.3	56.4
FP_25K	0.0	0.0
FP_20K	0.0	0.0
FP_15K	0.0	0.0
FP_10K	0.0	0.0
FP_5K	0.0	0.0
FP_1K	0.0	0.0
FN_25K	0.0	0.0
FN_20K	0.0	0.0
FN_15K	0.0	0.0
FN_10K	0.0	0.0
FN_5K	0.0	0.0
FN_1K	0.0	0.0
IR_25K	0.0	2.5
IR_20K	0.0	3.0
IR_15K	0.0	2.9
IR_10K	0.0	4.1
IR_5K	0.0	5.7
IR_1K	0.0	16.6
UN_25K	0.0	29.0
UN_20K	0.0	29.3
UN_15K	0.0	28.2
UN_10K	0.0	27.4
UN_5K	0.7	28.1
UN_1K	19.7	26.9