NTNU – Trondheim
Norwegian University of
Science and Technology

Relatedness estimation and pedigree reconstruction on large datasets

PROJECT THESIS

DECEMBER 2018

ALICE GUDEM
Department of Computer Science (IDI)
NTNU - Norwegian University of Science and Technology

Summary

The HUNT study has a biobank with DNA from about 70,000 people in Nord-Trøndelag. As a large part of the population has participated, it is expected that there is a high degree of relatedness within the population, and in such cases, it is normal to apply some relatedness estimation methods on the dataset.

There are many methods that exist for this purpose. However, mitochondrial DNA, which is directly inherited through the maternal line, doesn't follow the inheritance patterns of normal DNA. This raises the question of whether normal relatedness estimation is sufficient when considering mitochondrial DNA.

To approach this question, I have looked into what exists of relatedness estimation methods, especially those involving pedigrees, since such methods could be a natural starting point for estimating relatedness based on mitochondrial DNA. I have also tested the method DRUID on the HUNT data.

There are methods that do both relatedness estimation and pedigree reconstruction, methods that only do pedigree reconstruction, and methods that do relatedness estimation by leveraging reconstructed pedigrees. All of the methods considered in this report are IBD-segment based, and they utilize different approaches to improve performance and detection, which includes multi-way relatedness estimation, the use of composite likelihoods, and exploiting other factors such as mitochondrial DNA, Y-haplotypes, age and sex.

The results from DRUID showed a high degree of relatedness within a subset containing 1,000 individuals from the HUNT data, but there were some factors that added to the uncertainty of the results, namely the recombination rates and genetic mappings that were used.

The script, Druid-bakeoff, also crashed due to too much RAM usage when running on a subset of 7,000 individuals, so it is questionable whether DRUID could handle the complete set. That would nonetheless require a rewriting of the script.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"The Nord-Trøndelag Health Study" (HUNT) is a large health study first started in 1984 with about 120,000 participants providing health information, with DNA samples from almost 80,000 of the participants. The HUNT data is used in a wide area of research, and is especially useful for finding connections in the interactions between genetics, lifestyle and environment.The study has a high support, as a large part of the population in Nord-Trøndelag has participated (Krokstad et al., 2013). It is therefore expected that many of the participants will have relatives in the dataset.

In genetics, it is essential to know whether individuals are related or not when studying their DNA. At one end, relatives are needed in family association- and linkage-studies, and at the other end, studies like case–control association studies require individuals to be unrelated (Daly & Day, 2001).

To account for relatedness, one might use reported pedigrees (e.g. family connections taken from health registers), however these might hide cryptic relatedness. In addition, limiting samples to a geographical area can make cryptic relatedness inflate the false positives rate in case–control association studies (Voight & Pritchard, 2005). It is therefore normal to perform some sort of pedigree reconstruction or relatedness estimation on datasets of DNA samples to find the truest pedigree, and hence have knowledge of whether individuals are related or not.

Many methods exist for reconstructing pedigrees, some of them including PRIMUS, PADRE, DRUID and CLAPPER (Staples et al., 2014; Staples et al., 2015; Staples et al., 2016; Ramstetter et al., 2018, Ko and Nielsen, 2017). These methods utilize Identity-by-Descent (IBD) to estimate kinship between individuals, however, as the degree of relatedness increases, accuracy begins to drop. ERSA, a method for relatedness estimation using IBD, can detect up to 97% of 1st through 5th degree relationships within one degree, and then it drops to 80% for 6th and 7th degree (Huff et al., 2011). PADRE can discover over 50% of 13th degree relationships by combining ERSA and PRIMUS (Staples et al., 2016). When the generational degree increases, i.e. when the shared ancestor becomes more distant, so does the probability that a relationship is non-detectable between two individuals, i.e. the individuals don't share IBD segments that is long enough (Donnelly, 1983).

The pedigree in Figure 1.1 illustrates how autosomal, nuclear DNA is inherited, and a person will inherit roughly 25% of their grandparent's DNA, 12.5% of their great-grandparent's DNA, 6.25% of their great-great grandparent – and so it continues by halving the amount of DNA inherited in each generation. As stated in Donnelly (1983), the probability of detecting a 12th generation ancestor is 16% and it is 9% for a 13th generational ancestor. It is perhaps enough to set the limit at around 10-12 generations in research, when considering autosomal, nuclear DNA. At one point, two distantly related individuals will actually have independent, non-related DNA – the segments they have in common will be so small that they are negligible and their nuclear DNA will be unrelated in practice.

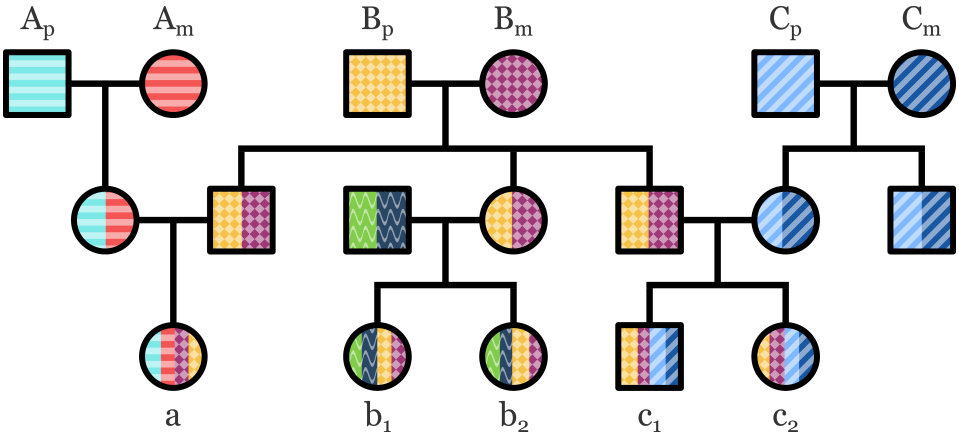

Figure 1.1: The figure depicts a pedigree with 18 individuals. Circles represent females and squares represent males. The color-patterns indicate the nuclear DNA of the founders, and how the DNA are inherited downwards through the generations.

However, with mitochondrial and allosomal DNA, inheritance works a little differently. Mitochondrial DNA (mtDNA) is inherited in its entirety through the direct maternal line and changes are due to mutations and heteroplasmy. Allosomal DNA is the DNA of the sex chromosomes, and a man's Y-chromosome has been inherited in its entirety through the direct paternal line. This essentially means that an individual might have more or less the same mtDNA or Y chromosome DNA as that of their direct female or male ancestor more than 10-12 generations back.

The difference between inheritance in nuclear DNA and mitochondrial DNA raises some questions, when considering large biobanks like that from the HUNT study - datasets with a large participation among a population that has a high degree of relatedness:

- Can it be the case that two individuals are closely related (e.g. they are cousins), but they have different, distant enough, female ancestors and hence independent mitochondrial DNA? What about seemingly unrelated individuals, can they have dependent mitochondrial DNA?
- When considering the degree of relatedness in mitochondrial DNA, where do we draw the line? Few methods can estimate relatedness above 10th degree with high

certainty, but it might be necessary to go farther behind with mitochondrial DNA because of its low mutation-rate.

Maybe doing a relatedness estimation on DNA samples might not be enough, when considering mitochondrial or allosomal DNA. Bodner, Irwin, Coble, and Parson (2011) proposes a quality control method where they identify close maternal relationships by looking at the autosomal DNA, and exclude samples if they find close relationships (mother/child or sibling-relationships) for use in forensic work. However, they also assume a randomly selected and relatively small sample set representing a big population - a sample set that might have a low degree of relatedness to begin with. This is not the case with the HUNT data, as it is not random and it has a high degree of relatedness.

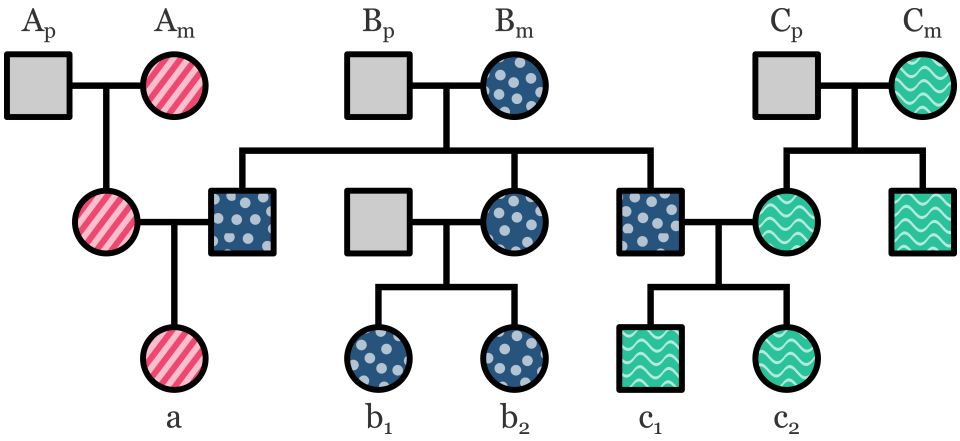

Figure 1.2: This is the same pedigree as in Figure 1.1, but showing mitochondrial DNA and its inheritance patterns instead of nuclear DNA. As can be seen, mitochondrial DNA is inherited through the maternal line.

1.1 An example

Consider Figure 1.2, which shows the same pedigree as Figure 1.1, but with the colors showing mitochondrial DNA inheritance as opposed to nuclear DNA. As can be seen, there are three groups of siblings – a, b_x and c_x. They are closely related to each other (they are all cousins). The figure showing nuclear DNA inheritance (Figure 1.1) shows that about half their nuclear DNA stems from the same two persons, grandfather B_p and grandmother B_m (that doesn't necessary indicate that they've all inherited the exact same DNA, but they might have parts in common). However, Figure 1.2 shows that they might have different mtDNA: If one assumes that their maternal grandmothers, A_m, B_m and C_m don't have a recent common ancestor, then their mtDNA might be different and hence independent – even though they're closely related.

On the other side, a might have a distant cousin d, say of 20th degree – outside the scope of what state-of-the-art pedigree reconstruction systems can infer – which actually

share a's maternal ancestor and hence a's mtDNA, making two individuals with independent nuclear DNA, actually sharing mtDNA.

1.2 Mitochondrial relatedness estimation

Assume you have a system – let's call it MITO for simplicity – that can infer mitochondrial relatedness in a big data set with a high degree of relatedness. How would the output look like, compared to a system that infers relatedness according to nuclear DNA (e.g. a pedigree reconstruction algorithm)? This might rely heavily on what threshold is chosen for the degree of relatedness. To illustrate, look at Figure 1.3. The figure shows three groups of relatives, denoted by the stroked shapes, that could be the output from some pedigree reconstruction algorithm, and the three following cases describe different scenarios.

Case 1 All of the mitochondria-groups are enclosed by the nuclear-groups, meaning that no individual shares mtDNA with anyone they're not related to.

By simply running a pedigree reconstruction algorithm, and tracing the maternal lines in the resulting pedigree would yield this result, i.e. the threshold of the degree of relatedness is the same for both nuclear DNA and mitochondrial DNA.

Case 2 The mitochondria-groups are crisscrossing the nuclear-groups, and there could be more mitochondria-groups than there are nuclear-groups, i.e. individuals in one family share a distant maternal ancestor with individuals in another family.

This could be a case where the relatedness-degree threshold for mitochondrial DNA is a bit higher than that for nuclear DNA, but not too much, e.g. the threshold is 10 generations for nuclear DNA and 20 generations for mtDNA.

Case 3 The mitochondrial-groups are considerably larger than the nuclear-groups, and they can both enclose whole nuclear-groups and crisscross them.

This could be the result of having a considerably higher relatedness-degree threshold for mitochondrial DNA compared to nuclear DNA, e.g. ten generations for nuclear DNA and 50 generations for mtDNA. It can also be the result of inbreeding farther back.

What consequences could this have? In case 2 and 3, it is easy to see that if you pick two individuals from the same nuclear-group, and assume they have related mtDNA, that might be wrong – although this can easily be solved if you have a pedigree and can ensure that the two individuals have a shared maternal ancestor. On the other side, if you choose two individuals from different nuclear-groups and assume their mtDNA to be unrelated, that might also be wrong, and this can lead to unaccounted for errors in further research. Furthermore, there might be fewer mitochondria-groups as seen in case 3 or there might be more, as seen in case 1 and 2.

Case 1 essentially describes the solution that is possible today - run a pedigree reconstruction method on the dataset and find the maternal paths in the pedigrees. There has been done little research on automatic relatedness estimation based on mitochondrial

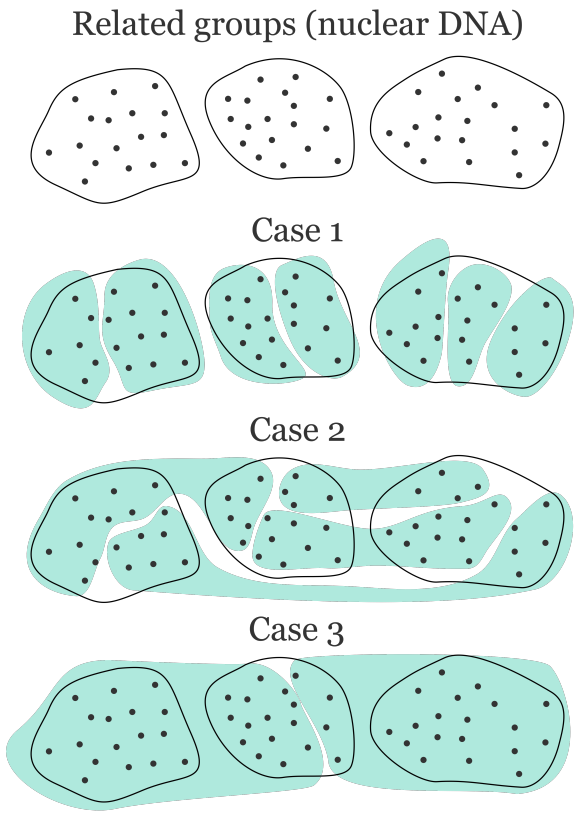

Figure 1.3: How could nuclear relatedness differ from mitochondrial relatedness? The dots represents individuals; the stroked shapes indicate relatedness between individuals based on their nuclear DNA, typically within ten degrees; and the filled shapes represents how the individuals could be related based on their mitochondrial DNA (which can be more than ten degrees relatedness).

DNA in huge DNA data sets, and with the increased effort in creating biobanks on a massive scale (Ramstetter et al., 2017), and with datasets like that of the HUNT study, this should be looked into.

1.3 The focus of this work

The example and the cases above gives the following motivation: To create a tool that is able to infer mitochondrial relatedness in datasets like that of the HUNT study, and further to compare the ”mitochondrial” relatives output from such a tool to ”normal” relatives in the dataset.

A tool that is able to infer mitochondrial relatedness might help researchers judge

whether two individuals' mtDNA is related or not, and assist them in choosing samples for studies. A daunting prospect would also be that the tool could be able to estimate the amount of maternal founders in a population.

In other words, a system like MITO is needed, and as far as my knowledge, such a system does not exist. My approach to this problem will be:

1. To first apply a pedigree reconstruction method on the data set.
2. Find individuals that share mitochondrial DNA within the discovered pedigrees
3. Estimate mitochondrial relatedness between the founders of the maternal lines.

Why start with pedigree reconstruction when looking at mitochondrial DNA relatedness? Why not just estimate mtDNA relatedness first? If you already have a pedigree that shows the relationships between individuals, some part of discovering the mitochondrial relatedness will already be done, as one can simply trace the maternal lines in a pedigree and thereby know who shares mitochondrial DNA within the degree that pedigree. As state-of-the-art pedigree reconstruction methods are able to detect low degrees of relatedness with a high certainty, this might yield a better result than simply looking at the mtDNA alone.

In this report, I will focus on point 1, looking into relatedness estimation methods that involve pedigrees (by either reconstructing or leveraging them). I have done a literature review on what exists of such methods and their approaches, which can be found in chapter 3. Furthermore, I have tested the pedigree reconstruction- and relatedness estimation-method DRUID, to see how it performs on a big data set like the HUNT study – ideally, I want a method that can detect as high a degree of relatedness as possible, handle big data sets and have a reasonable run time. The methods and results can be found in chapter 4 and chapter 5. Chapter 2 contains an introduction to concepts within genetics and biology that is necessary to know about, for the rest of the report.

Point 2 and 3 will be outside of the scope of this report, and will be a part of my master's thesis.

Chapter 2

Genetics

This chapter will contain a brief introduction to biology and genetics that is necessary to have a grasp of when reading the rest of the paper. Some parts will have a rather brief and superficial explanation, while others will go more in depth. It is also recommended to read this chapter sequentially for readers who are not familiar with concepts within these subjects.

Section 2.1 will explain relevant topics within biology and genetics, including the cell, DNA, mitochondria, gene and chromosomes to name a few. Section 2.2 will further explain concepts within heredity, including how DNA is inherited, what identity by descent means and the concepts degree of relatedness, pedigrees and cryptic relatedness.

2.1 Biology

2.1.1 Cells

A *cell* is the basic structural, functional and biological unit of all living organisms. It is enclosed within a membrane and contains several *organelles*, which is a specialized subunit that serves some special function.

Eukaryotic cells and the nucleus

An *eukaryotic cell* is a cell which has a nucleus. The *nucleus* is one of the larger organelles in a cell, and it contains the cell's genetic material, i.e. this is where most of an organism's DNA is located. The nucleus maintains the integrity of the DNA and controls the activities of the cell by regulating the gene expression. In other words, it works as the cell's control center.

A simplified eukaryotic cell is shown in Figure 2.1.

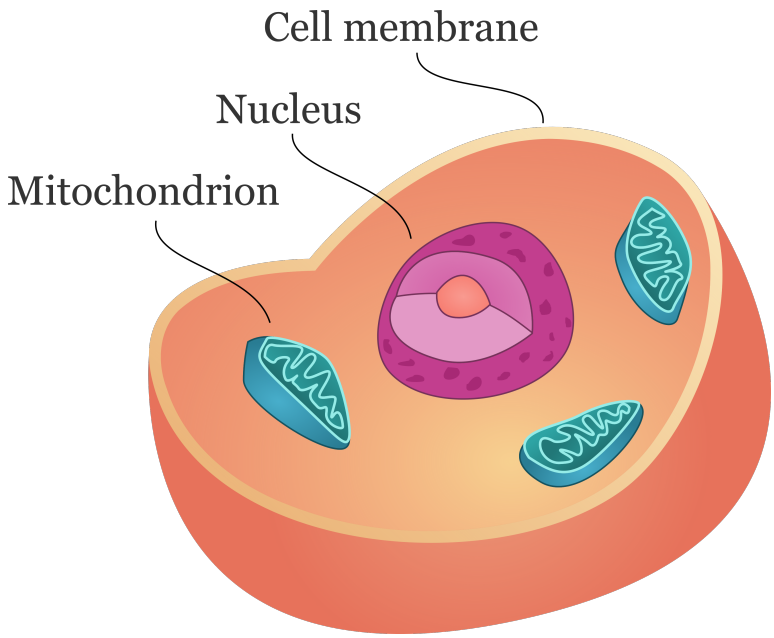

Figure 2.1: This figure shows a simplified eukaryotic cell – all the stuff that is not necessary for this report has simply been taken out, and what remains is the cell with its nucleus and its mitochondria, enclosed within the cell membrane.

Mitochondria

Another important organelle of the cell is the *mitochondrion* (plural *mitochondria*). It is often nicknamed the "Powerhouse of the Cell", mainly because it is responsible for supplying cellular energy by converting energy from food into a form that cells can use.

2.1.2 DNA and the genome

DNA is the hereditary material in humans and nearly all other organisms. All DNA in a human is referred to as the *genome*.

The DNA is arranged in two long strands of nucleotides that forms a spiral called a *double helix*. This double helix structure looks like a spiraling ladder as can be seen in Figure 2.2.

A *nucleotide* consists of a sugar molecule, a phosphate molecule and one of four chemical bases: Adenine (A), thymine (T), guanine (G) and cytosine (C). As shown in Figure 2.2, each base is paired up with another base in the two connected strands: Adenine is paired with thymine and cytosine is paired with guanine. One of the strands is therefore a "mirror" of the other.

The order of these chemical bases determines the information available for building and maintaining organisms, and hence DNA can be expressed as strings (e.g. "TATA" would be a sequence of nucleotides with the bases thymine, adenine, thymine and ade-

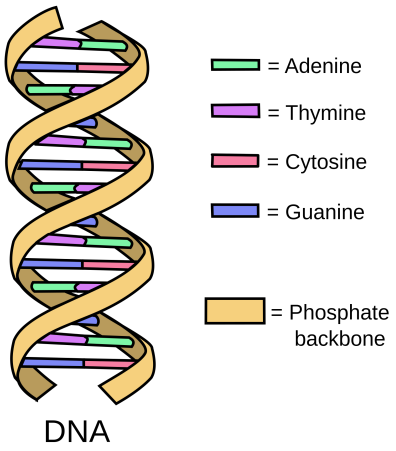

Figure 2.2: The image depicts two strands of DNA and how they are connected in a double helix with the four chemical bases. Source: user:Forluvoft / Wikimedia Commons / Public Domain

nine). A specific sequence (e.g. a gene) at a specific location in the DNA will therefore act as a code for some function (e.g. whether you have blue or brown eyes). Human DNA consists of about 3 billion bases, and more than 99% of these are the same in all people.

Nucleus DNA

The DNA located in the nucleus is called the *nucleus DNA* or simply *DNA*.

Mitochondrial DNA

The *Mitochondrial DNA* or *mtDNA* is a DNA found in the mitochondria. It is inherited directly from the mother, and can thus be used to trace the maternal line.

2.1.3 Loci

A locus (plural *loci*) is a fixed position on a chromosome, like the position of a gene or a genetic marker.

2.1.4 Genes

A gene is a sequence of the DNA string, and is the basic physical and functional unit of hereditary. A gene can be coding or noncoding, whereas the former acts as instructions to make protein sequences and constitutes about 1% of all human DNA (it is the proteins that do the actual work of the cell). The noncoding genes perform other tasks.

2.1.5 Chromosomes

In the nucleus of the cell, the DNA molecule is packed into thread-like structures called *chromosomes*. Each chromosome is made up of DNA coiled tightly together many times around proteins called histones that support their structure.

Each person normally has 23 pairs of chromosomes, where 22 of these pairs are called *autosomes* and look the same in both males and females. The 23rd pair, the *sex chromosomes* (or *allosome*), differ between males and females. Females have two copies of the *X-chromosome* while males have one X- and one *Y-chromosome*. The chromosomes are numbered after size in decreasing order.

Chromosome pair 1-22 are called *homologeous chromosomes* or *homologs*, meaning that they have the same size and code for the same genes – e.g. if one homolog contains a specific gene at a specific locus, the other homolog will have the same gene at the same locus, but they can have different alleles. In addition, a woman's X-chromosomes are homologs, but a man's Y and X-chromosomes are not.

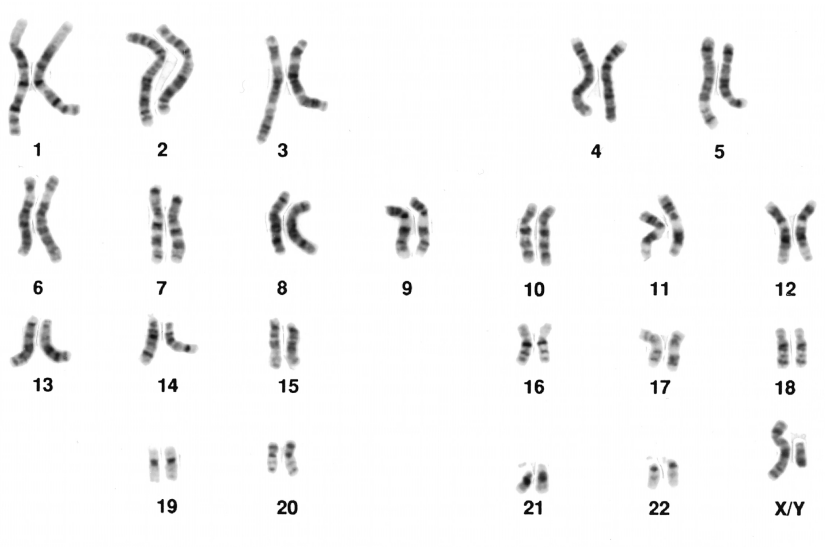

Figure 2.3: The chromosomes in a male individual. It shows the 22 chromosomes and a pair of sex chromosomes, numbered according to decreasing size. Source: National Cancer Institute / Wikimedia Commons / Public Domain .

2.1.6 Alleles, genotypes and phenotypes

An *allele* is the form a gene can take – where the gene is the variable, an allele is its value encoded in the DNA string.

A *genotype* is the alleles located at the same locus in a pair of homologs. More specifically, a genotype can be on the form Aa, where A is the allele on one of the homologs and a is on the other. A could be the allele for brown eye color, and a the allele for blue eye

color. The actual manifested physical trait is called the *phenotype*.

When individuals are genotyped (their DNA is samples), the genotypes are often unphased, meaning that one doesn't know which homolog has what allele. It is possible to infer the phasing of genotypes by looking at whole populations.

2.1.7 Haplotypes and haplogroups

A *haplotype* is a group of alleles that has been inherited together from a single parent, and hence will be located on one of the homologs. A *haplogroup* is a set of different haplotypes at different chromosome regions that are closely linked and are usually inherited together. Two examples of haplogroups is the Y-chromosome haplogroup and the mitochondrial haplogroup.

2.1.8 Single-nucleotide Polymorphisms

A *Single-nucleotide polymorphism*, often shortened SNP is a variation in a specific position on the genome where each variation occurs above some threshold for a population (e.g. more than 1%). For example, a SNP may replace the nucleotide cytosine (C) with thymine (T) on a certain stretch of DNA in about 20% of the population.

2.2 Heredity

As stated in subsection 2.1.5, each person has 23 pairs of chromosomes. For each of these pairs, one chromosome is inherited from the mother and one is inherited from the father.

This essentially means that each person has inherited roughly half their DNA from their mum and half from their dad.

2.2.1 Meiosis

So how is DNA from a parent duplicated to their offspring? It is done during *meiosis*, the process of creating *gametic cells* (gametic cells are either egg cells or sperm cells). As opposed to normal copying of DNA, which is called *mitosis*, where you want to copy the DNA exactly, gametic cells contains chromosomes that is a mixture of the two homologeous chromosomes in the original DNA, for each of the 22 pairs of chromosomes. As can be seen in Figure 2.4, four different chromosomes has been created based on one pair of chromosomes, and everyone is a little bit different. These four chromosomes could be represented as four siblings, and the process explains why siblings are different.

This also shows that a person will inherit roughly half of one of their parents DNA, and roughly half of that again will be from one of their grandparents.

Sex chromosomes

With the sex chromosomes, things are a bit different. The X-chromosome that each person inherits from their mother, will have been created the same way, with two homologs mixing, because a woman has two X-chromosomes that can be mixed.

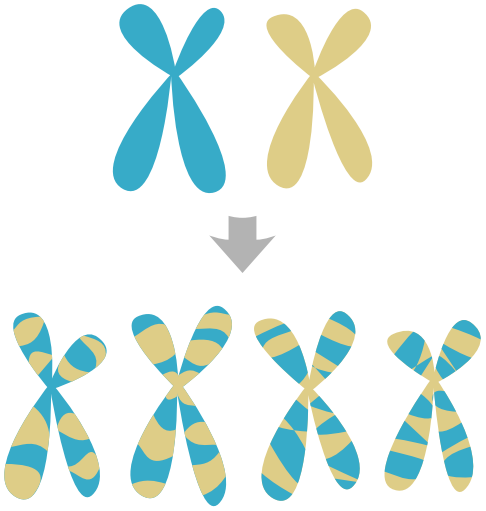

Figure 2.4: An example of how chromosomes are created based on two homologs during meiosis.

However, a man's X-chromosome and Y-chromosome, which are not homologous, are not compatible for this mixing process, and therefore they are passed unchanged to the next generation (roughly half of a man's sperm cells will have his Y-chromosome and the other half will have his X-chromosomes, also implicating that the sperm cells decides the gender of the offspring). In addition, this means that a man's Y-chromosome has been passed relatively unchanged through his paternal line of male ancestors.

Mitochondria

How then, is the mtDNA passed down the generations? Egg cells are a bit more complicated than sperm cells. In a way, egg cells are a kind of starter pack for a embryo to grow and develop, and one of these necessary items in the starter pack is the mitochondria, while sperm cells mainly contain DNA.

So the mitochondrion is passed directly and unchanged from the mother, during meiosis. This, in turn, means that the mtDNA has been passed relatively unchanged through the direct maternal line of a person's female ancestors.

Identity-By-Descent

When two people have inherited DNA from the same ancestor, and they have identical segments of DNA, those segments are said to be identiy-by-descent or IBD. Essentially, there has been no intermediate recombination of the segments through the generations. If two people have identical segments of DNA because of coincidences (i. e. they haven't inherited the same segment from a common ancestor), those segments are said to be identity-by-state (IBS).

Thompson (2013) states: "there is no absolute measure of IBD; IBD are always relative to some ancestral reference population". Which means that IBD are measured relative to

some point in time and the founder population at that point in time.

In addition, there are different IBD levels. IBD0 is the likelihood that two individuals are unrelated, IBD1 is the likelihood that the individuals are IBD on one haplotype and IBD2 is the likelihood that the individuals are IBD on both haplotypes (S. R. Browning & Browning, 2012).

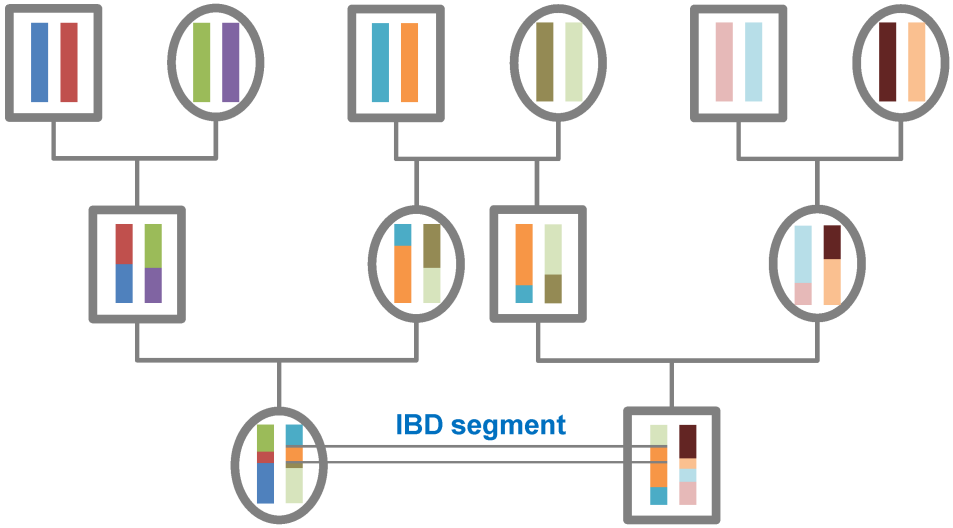

Figure 2.5: This is a pedigree with 12 individuals. Males are shown with squares and females with circles. The two vertical bars in each node represent two homologeous chromosomes (e.g. chromosome 1), and the colors indicate how these chromosomes are inherited. It shows how the pair of cousins in the bottom share an IBD segment inherited from their shared grandfather. Source: user:Gklambauer / Wikimedia Commons / CC-BY-CA-3.0 / GFDL

2.2.2 Degree of relatedness

The degree of relatedness follows the average sharing of DNA between two relatives, where first degree relatives share 50% of their genes on average, and these are parent/child- and full sibling-relationships. Second-degree relatives share about 25% of their genes, and includes and individual's grandparents, aunts, uncles, nephews, nieces and double cousins. Third degree relatives share about 12.5% of their genes and includes and individual's great-grandparents, great grandchildren and first cousins.

2.2.3 Pedigrees and cryptic relationships

A *pedigree* is a family tree or DNA tree. It might have been made based on reported relationships, but it can also be inferred from DNA data.

A *cryptic relationship* is an unknown relationship between two individuals. E.g. a reported pedigree on a dataset that hasn't included a specific relationship between two individuals.

Chapter 3

Relatedness estimation and pedigree reconstruction

Relatedness estimation is about inferring relatedness between individuals in DNA sample sets. Methods that infer relatedness typically output some number for each pair of individual in a dataset, indicating how related the individuals are. This can be the degree of relatedness, explained in subsection 2.2.2; a kinship coefficient, which is the probability that two alleles sampled at random are identity-by-descent (Manichaikul et al., 2010); or a relatedness coefficient which is two times the kinship coefficient (Ramstetter et al., 2017; Wright, 1922).

Pedigree reconstruction methods also attempt to build the most correct pedigree out of the sample data, which means that in addition to estimating the relationship between individuals (e.g. individual A and B have a 2nd-degree relationship or A and B has a aunt/nephew-relationhip), pedigree reconstruction methods also try to find out how they are connected and construct the tree (e.g. A is B's aunt on his mother's side). There are pedigree methods that only focus on the pedigree (Staples et al., 2014; Ko and Nielsen, 2017), methods that both construct the pedigree and infer more distant relationships (Ramstetter et al., 2018) and relatedness estimation algorithms that make use of estimated pedigrees (Staples et al., 2016).

The chapter will begin by explaining IBD-detection and what degrees of relatedness that are possible to infer, and further explain different approaches to pedigree reconstruction methods and relatedness estimation methods that make use of pedigrees. These include multi-way relatedness estimation, composite likelihoods, searching the pedigree space and exploiting other factors.

3.1 IBD-detection

The methods presented in this chapter all make use of pairwise IBD-estimates, and there are many different methods for estimating IBD-segments between individuals.

To discover IBD segments, haplotype frequency is a key aspect. If the frequency for a haplotype is small, its appearance in individuals might be due to IBD. However, one usually does not have haplotype data, but rather unphased genotypes, which is why many IBD detection methods also involve haplotype estimation (phasing) (S. R. Browning & Browning, 2012).

The different approaches to detecting IBD-segments, presented in (S. R. Browning & Browning, 2012), includes using a length threshold (i.e. a minimum length of shared DNA is needed for that segment to be IBD) on genotypes or haplotypes; taking a probabilistic approach, where linkage disequilibrium may or may not be taken into account; look at rare, shared haplotypes; or consider several individuals as opposed to considering pairwise relationship, especially when finding shorter segments.

As many pedigree reconstruction methods and relatedness estimation methods use IBD to detect relatedness, it is important that the IBD estimates are of good quality. As Staples et al., 2014 states, several factors can affect IBD estimates, including the number of genetic markers, population substructure, admixed populations and minor allele frequencies. A study testing IBD methods on french-canadian populations also found that the total length of IBD sharing increased, when they had inbred ancestors (Gauvin et al., 2014). If inbreeding in population is not taken into account, methods may wrongly classify individuals to be more related than they actually are, e.g. giving them a higher degree of relatedness.

It is also worth mentioning that IBD-detection-methods can be used in itself to detect relationships, an example being RefinedIBD. (B. L. Browning and Browning, 2013b; Ramstetter et al., 2017)

3.2 IBD and degrees of relatedness

When the relationship becomes more distant, the average proportion of shared DNA drops exponentially. More specifically, around 9th degree relatedness, the probability to detect IBD segments between individuals rapidly decreases (Donnelly, 1983). However, if two distantly related people do share an IBD segment, the segment is relatively long - an example being full fifth cousins, separated by 12 meioses and sharing both a male and female ancestor. They will on average share 0.1% of their genome or approximately 3cM, and when they do have an IBD segment in common, it is usually composed of a single segment with a mean length of 8.3 cM. In addition, individuals from the same geographic region may have many recent common ancestors, but may still have only one or two detectable IBD segments (S. R. Browning & Browning, 2012). This essentially means that it is difficult to detect distant relationships, especially based on pairwise comparisons of DNA alone.

The results from a benchmarking of different relatedness estimation methods done by Ramstetter et al. (2017) show that all tested methods perform well on 1st and 2nd degree relationships, with an accuracy of $>98\%$ and $>92.8\%$ respectively. However, the IBD segment-based methods perform better, and can also detect more than 76% of up to seventh degree relatedness correctly within one degree. One important aspect to note, is that it is not always necessary to detect the exact degree of relatedness with distantly related individuals, and that discovering relatedness within one degree might be highly

informative (Ramstetter et al., 2018).

PADRE can detect 4th through 13th degree relatedness between individuals. More specifically, PADRE could detect over 50% of 13th degree relationships although 95% of these share no genetic material through their most recent ancestor – this was accomplished by the use of multi-way relatedness, which is explained further in section 3.3 (Staples et al., 2016).

Another method that is able to detect distant relationships is DRUID (Ramstetter et al., 2018). They reported detection of $>$79% 10th degree relatives correctly or within one degree of truth, and an improvement relative to PADRE in many cases. DRUID also makes use of multi-way relatedness to improve their detections.

3.3 Multi-way relatedness estimation

Many methods make use of pair-wise estimation, meaning that they compare DNA between pairs of individuals. However, using multi-way estimation can yield better results (Staples et al., 2016; Ramstetter et al., 2018). Multi-way estimation makes use of a set of individuals rather than just two, when inferring relationships.

An example is PADRE (Staples et al., 2016), a method that combines the use of PRIMUS (Staples et al., 2014) and ERSA (Huff et al., 2011), by running PRIMUS to find pedigrees (which also detects the first three degrees of relatedness) and by using the likelihood ratio test in ERSA on the founders of the pedigrees. The test compares the null hypothesis that two founders are unrelated to the alternative, that they are Nth related. If a relationship is found with ERSA, PADRE will find the best fitting relationship between the founders by using a composite likelihood framework between the two pedigrees, by comparing the DNA of all individiuals in one of the pedigrees with all of the individuals in the other pedigree. If PRIMUS yields several pedigrees for a given family network, PADRE will compare for all pedigrees. That way, all individuals in both pedigrees are leveraged to find the degree of relatedness between them.

In their result, when testing on real data with 169 individuals of a predominantly northern ancestry, PADRE and ERSA had the same accuracy when founders didn't have any 1st or 2nd-degree relatives (essentially yielding pedigrees with only the founder). However, when the number of 1st and 2nd degree relatives increased, PADRE's accuracy also increased. When founders had two 1st or 2nd degree relatives, ERSA detected 23% of 10th degree relatives while PADRE detected 39% (both with a 95% confidence interval).

Another method that leverages multi-way relatedness, is DRUID. More specifically, DRUID uses pairs of siblings to find and identify aunts and uncles. They leverage the fact that all relatives besides the descendants of an individual are related through at least one parent of each sample. More specifically, a pair of siblings has more information about their ungenotyped parent than only one of them, and the siblings' aunts or uncles will share a sizable amount of DNA that is IBD2 with their parent (i.e. IBD on both haplotypes) (Ramstetter et al., 2018).

3.4 Composite Likelihood Estimation

Some methods use composite likelihood frameworks to decrease computation times. A composite likelihood is an inference function derived by multiplying components' likelihoods (Varin, Reid, & Firth, 2011), as opposed to for example taking the maximum likelihood of the the whole set of components together, which oftentimes might be computationally exhaustive. Methods that use composite likelihood frameworks include CLAPPER (Ko & Nielsen, 2017) and PADRE (Staples et al., 2016).

Padre uses a composite likelihood framework together with multi-way relatedness estimation as explained in section 3.3. The full composite likelihood for a pair of pedigree structures is the product of the pedigree likelihoods from PRIMUS and all pairwise relationships calculated by ERSA across the two pedigrees. For two founders x and y that are Nth-degree related, the equation is as follows:

$$\hat{L}_{1i,2j}(x,y|N) = L_{Net_{1i}} L_{Net_{2j}} \prod_{\substack{\forall a \in Net1_i \\ \forall b \in Net2_j}} \hat{L}_{ab}(S_{ab}|D_{ab})$$

D_{ab} is the degree of relatedness between individuals a and b given by N, and S_{ab} is a set with the lengths of IBD segments a and b share. $\hat{L}_{ab}(S_{ab}|D_{ab})$ is the maximum likelihood that individuals a and b are D_{ab}-degree related. $L_{Net_{1i}}$ and $L_{Net_{2j}}$ are the composite pedigree likelihoods, which are the sum of the log likelihoods of all pairwise relationships within each pedigree (Staples et al., 2014).

Clapper is another method that uses composite likelihoods (Ko & Nielsen, 2017). Clapper can detect up to five generations, and do not estimate distant relatedness: It only focuses on the pedigree. The method assumes an outbred population and no cycles in the population, except when full siblings get children.

The method takes the composite likelihood of full pedigrees, where H is a pedigree, X_i is the genotype vector for individual i, $R_{i,j}$ is the relationship between i and j induced by the pedigree and k is the number of individuals in the pedigree:

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

If k is one, then the likelihood is simply the probability of observing the individual's genotypes (e.g. the probability of observing the genotype Aa at a specific locus is 0.20 in a population). If $k > 1$, then the likelihood is the product of the marginal pairwise likelihoods $P(X_i, X_j|R_{i,j})$, which is an approximation to calculating the full conditional likelihood:

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

And since the probability of each individual is calculated $n-1$ times, it is divided by the marginal likelihood of each individual $k-2$ times to obtain a more natural scaling.

3.5 Searching the space of possible pedigrees

For a given group of relatives, methods might output the true pedigree, but also untrue pedigrees, e.g. by misclassifying a grand-father as an uncle, since both are second-degree relationships. Which is why pedigrees might evaluate several pedigrees based on the same individuals.

PRIMUS, a pedigree reconstruction method that can detect up to 3rd degree relatives, essentially tries to search the entire pedigree space. It first identifies family networks based upon on IBD scores, where everyone in the network has to be related to at least one other individual in the network. Given a network, individuals are iteratively added in pedigrees, trying every possible solution, and pedigrees are rejected when pairwise relationship likelihoods between individuals within the pedigree is below a given threshold or when the pedigrees are incorrect (which is explained further in the next section). The method might end up with several pedigrees for a family network at completion, where each pedigree has a pedigree likelihood score, which is the log likelihood of all pairwise likelihoods within the pedigree (Staples et al., 2014).

This way of doing it might increase computation time, and Staples et al. (2014) report a runtime that increases exponentially with the pedigree size and missing links. For example, pedigrees of size 50 with 20% missing samples required > 36 hours, as well as $> 12GB$ of RAM.

CLAPPER (Ko & Nielsen, 2017) uses simulated annealing as an heuristic, to avoid searching the entire pedigree space. It instead finds initial pedigrees, and do changes to these pedigrees. The changes can be divided into three classes: Changing pairwise relationships between individuals, e.g. from grandfather/grandson to uncle/nephew; splitting or joining pedigrees, or a combination of these; and transitioning between similar pedigrees, when information like age or sex is missing. The method uses simulated annealing by always accepting changes that will increase the likelihood of the pedigree and sometimes accept changes that will decrease the likelihood, to avoid a local maxima.

3.6 Exploiting other factors

It is possible to exploit other factors to further improve relatedness detection. Some of these include age, sex, mitochondrial DNA and Y-haplotypes.

PRIMUS is an example that exploits mitochondrial DNA and Y-haplotypes (Staples et al., 2015) as well as age and sex (Staples et al., 2014) to decrease the search space of possible pedigree by rejecting invalid pedigrees. For example, if two of the same sex has a child or if the parent of a child is younger than the child, then the pedigree is rejected.

If the pedigrees have missing links, mtDNA or Y-haplotypes can be used to reject them. More specifically, for mtDNA, one can trace the maternal line in the pedigree and if the mtDNA for the samples on the lines are discordant, then the pedigree can be rejected. The same logic can be applied for Y-haplotypes, but only for males (Staples et al., 2015).

Figure 3.1 shows how mtDNA can be used to reject pedigrees. The pedigree has reported A as the grandmother of B, but there is a missing link between them - a non-genotyped individual. If A and B has concordant mtDNA, then this pedigree may be the

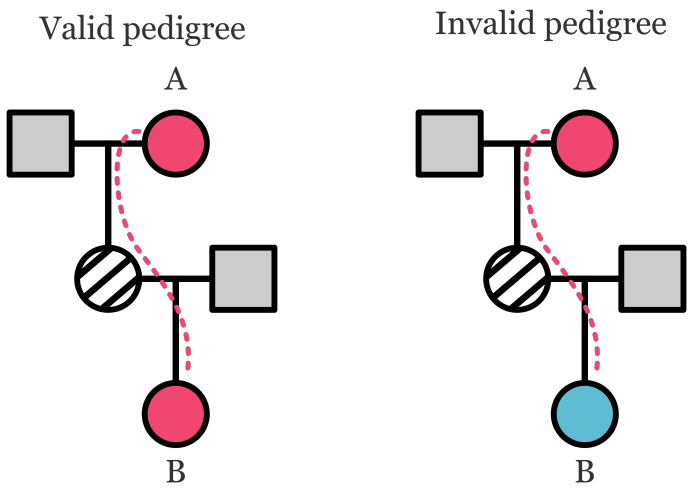

Figure 3.1: This figure shows two versions of a pedigree with 5 individuals, whereas one individual's genotype is missing, indicated by the striped pattern. Squares represent males, circles represent females, and the colors of A and B indicate their mtDNA, where they have the same mtDNA in the valid pedigree and different mtDNA in the invalid pedigree.

true pedigree. However, if A and B's mtDNA is disconcordant – there is no match – then one can safely assume that this is a wrong pedigree and reject it.

Chapter 4

Methods

This chapter contains a description of the programs used and flow of data when testing the pedigree reconstruction- and relatedness estimation-method DRUID (Ramstetter et al., 2018).

Figure 4.1 illustrates the flow of data, as well as the sample size of the data. Five program were used, namely PLINK (Purcell, n.d. ; Purcell et al., 2007), BEAGLE (S. R. Browning & Browning, 2007), RefinedIBD (B. L. Browning & Browning, 2013a), Druid-bakeoff (Ramstetter, 2018) and DRUID (Ramstetter et al., 2018).

The HUNT data was provided in .bed format, which is PLINK's standard format. It contains DNA data from about 70,000 individuals, genotyped at about 350,000 markers. DRUID was run on a subset of the individuals and a subset of the markers. A description of the HUNT data can be found in the Appendix.

First, PLINK was used to create a .map file for Druid-bakeoff and DRUID, as well as converting the .bed file to .vcf format for BEAGLE.

BEAGLE was used to phase the genotypes, and RefinedIBD took a subset of the phased data as input, and calculated pairwise IBD scores for the individuals in the subset.

The .ibd file from RefinedIBD were given to Druid-bakeoff, together with the .map file from PLINK, to get .ibd12- and .seg-files for DRUID. At last, the files from Druid-bakeoff, together with the .map-file were given to DRUID, which reconstructed a pedigree and inferred relationships between the individuals in the set. The results are shown in chapter 5.

4.1 Equipment

The HUNT cloud was used to test the method, and the machine had 4 cores and 32 GB RAM as its configuration.

Table Table 4.1 shows the versions of the programs and their last edit date.

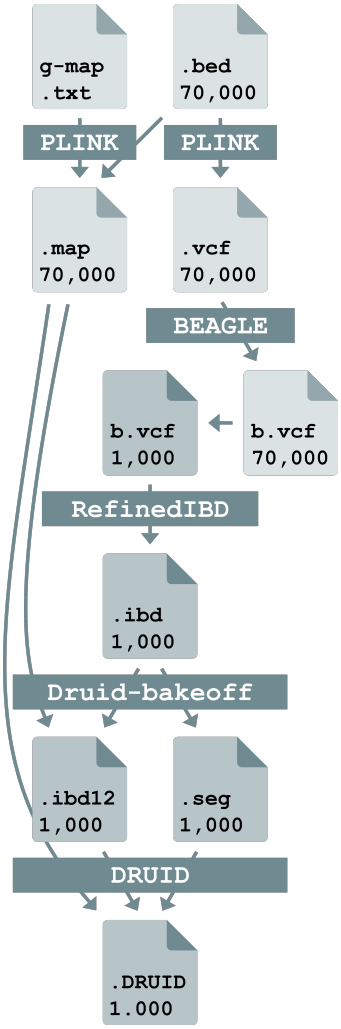

Figure 4.1: The data flow that was required to run druid. The numbers on the files indicate the sample size, i.e. the number of individuals.

Table 4.1: Program versions

Program	**Version**	**Last changed**
PLINK	1.90p 64-bit	March 25th, 2016
BEAGLE	5.0	September 7th, 2018
RefinedIBD		July 12th 2018
Druid-bakeoff		August 29th 2018
DRUID	1.02.1	October 23rd 2018

4.2 Execution

DRUID requires three files for running: A .map-file, a .seg-file and a .ibd12 file.

.map file

The following PLINK command was issued to get the .map file:

```
$ plink2 --bfile genotyped.bed --recode --cm-flag genetic-map-file.txt
    --out output-prefix
```

Listing 4.1: Creating .ped and .map files in PLINK

It converts a .bed file into a .map file with the `recode` flag. The `cm-map` flag needs a file that has a mapping from physical position to genetic position of the variants. If the `cm-map` flag is not set and the genetic mapping not provided, all genetic positions in the .map file will be set to zero, which yields wrong results in DRUID.

The mapping from physical positions to genetic positions were taken from Pickrell (2014), which contains interpolated map positions for physical and genetic positions of genes according to the CEU population (Utah residents with a northern and western european ancestry). The file-format where changed according to the SHAPEIT format for genetic maps ("The Genetic Map", 2018). An example of such a file and format is shown in Table 4.3.

The recombination rates were set to 1.0 centi-Morgans per Mega-base (cM/Mb) for all positions, as that is the approximate average rate of recombination in humans (Milo & Phillips, 2016) and it was also used as a default value in BEAGLE and RefinedIBD.

In addition, all rows that contained the same genetic position as the last row were removed from the file. Table 4.2 and Table 4.3 shows an example of this.

Table 4.2: A genetic map-file with duplicates of genetic positions. Duplicates are greyed out.

Physical position (bp)	Recombination rate	Genetic position
72765	0.12455	0.00266
94172	0.12458	0.00266
94426	0.12461	0.00269
95949	0.12461	0.00269
98087	0.12460	0.00315

Table 4.3: A genetic map-file where duplicates of genetic positions has been removed.

Physical position (bp)	Recombination rate	Genetic position
72765	0.12455	0.00266
95949	0.12461	0.00269
98087	0.12460	0.00315

.ibd12 and .seg files

As shown in Figure 4.1, DRUID requires IBD scores to be calculated, and the authors recommended BEAGLE 4.1 with refinedIBD run three times with different seeds. BEAGLE 5.0 has since been released, with refinedIBD taken out as a stand-alone program, which were instead chosen because of the improved performance.

BEAGLE requires input in VCF-format, which was done with PLINK:

```
$ plink2 --bfile genotyped.bed --recode vcf
```

Listing 4.2: Creating .vcf files in PLINK

BEAGLE was run the following way:

```
$ java -Xmx24g -jar beagle.28Sep18.793.jar gt=/path/to/dataset.vcf out=/
    path/to/output/prefix seed=1564 ne=100000
```

Listing 4.3: BEAGLE

The flag `seed` was set to 1564, and `ne`, effective population size, was set to 100,000 for Nord-Trøndelag.

The output from BEAGLE was the input to refinedIBD. The new version of refinedIBD doesn't have `seed`-argument anymore, so it was run once. As noted above, BEAGLE does have a `seed`-argument, but it took almost two weeks to run and running it two more times were not an option, considering time restraints.

BEAGLE crashed during chromosome 20 because of how PLINK 1.9 structures sex chromosomes during VCF file generation, yielding alleles without separators – a must for BEAGLE. At this point, a subset of 1,000 individuals was chosen from the output of BEAGLE (also in VCF-format) by taking the 1,009 first columns – 9 fixed columns and 1,000 columns with genotypes from the 1,000 individuals. The following command was used to subset the dataset:

```
$ cut -f 1-1009 /path/to/beagle-phased-dataset.vcf > /path/to/beagle-
    phased-subset.vcf
```

Listing 4.4: Subsetting individuals

To run RefinedIBD, the following command was issued:

```
$ java -Xmx32g -jar refined-ibd.12Jul18.a0b.jar gt=/path/to/beagle-phased-
    subset.vcf out=ibd-scores
```

Listing 4.5: RefinedIBD

The final step to get files in .ibd12 and .seg formats, where to run the script Druid-bakeoff using the output of refinedIBD and the .map file from above.

However, both the script and the .map file had to be modified due to BEAGLE crashing in the middle of chromosome 20 because of keyerrors in Druid-bakeoff. All variants that were not considered by BEAGLE had to be removed from the map file. This was done by finding the highest variant position in chromosome 20 that had been considered by BEAGLE, and removing all variants after this variant from the map file, since the map-file was sorted after chromosome number and in increasing order of variant positions. In

addition, a couple of for-loops in the Druid-bakeoff script that iterated over chromosomes 1-23 had to modified to iterate over chromosomes 1-20 to get the correct total genome length.

Druid-bakeoff and DRUID were run the following way:

```
$ python getIBD.py -f ibd-scores.ibd ibd-scores.ibd ibd-scores.ibd -m map-
    file-from-plink.map -s 1 -t 1 -o ibd12-and-seg

$ python DRUID.py -o output-prefix -i file.ibd12 -s file.seg -m file.map
```

Listing 4.6: Druid-bakeoff and DRUID

Note that instead of giving three different ibd-files to Druid-bakeoff, the same file was instead given thrice.

Chapter 5

Results

This chapter contains both the results from running the programs itself (i.e. their runtime and memory usage) and the output from DRUID.

5.1 Issues with big datasets

DRUID was chosen because it can detect large degrees of relatedness, but due to limited time and limited computing power as well as a large dataset, this proved to be a more difficult task than anticipated. To run DRUID, IBD-scores are needed and the authors recommended using RefinedIBD. However, to run RefinedIBD, samples need to be phased, which can be done with BEAGLE 5.0. Running the complete dataset in BEAGLE took almost two weeks and it crashed during phasing of chromosome 20, which prompted me to reduce the set considerably.

The dataset was first reduced to 7,000 individuals, but unfortunately, the Druid-bakeoff script couldn't handle the size of the IBD-file and crashed due to too much RAM usage, which lead to yet another reduction in sample size, from 7000 individuals to 1000 individuals. The set was also reduced to consider only chromosomes 1 to 20.

I do not know how DRUID would perform on the complete set, but that would nevertheless require rewriting of the druid-script or creating a new script that is able to convert IBD-files from a single run to the wanted input-format of DRUID. Ramstetter et al. (2018) reported an average runtime of 44.8 hours for a datasets of 1,155 individuals that was enriched in 2nd-degree relatives (where DRUID can utilize its multi-way estimation). During my run, DRUID completed rather fast (it spent no more than 10 minutes), which might be because the subset didn't contain a considerable amount of second degree relatives. It is entirely possible that the complete dataset of 70,000 individuals does have a lot of those relationships, but on the other hand, it might not have that many missing links.

Table 5.1: The number of relationships found in the dataset of 1000 individuals, based on degree. The first column shows which degree it is, where 0 is no relatedness. Number of relationships indicate how many Dth degree relationships where found, and the last column indicate how many Dth degree relationships an individual has on average, which is simply the number of relationships divided by the number of individuals.

Degree of relatedness	Number of relationships	Average relationship per individual
0	30163	30.163
1	24	0.024
2	19	0.019
3	70	0.070
4	191	0.191
5	595	0.595
6	2528	2.528
7	19809	19.809
8	103525	103.525
9	175535	175.535
10	118951	118.951
11	47715	47.715
12	375	0.375

5.2 Output from DRUID

Estimating relationships between n people correspond to a complete graph, meaning that the results contain $r = n(n-1)/2$ relationships. When $n = 1000$, this correspond to $r = 499500$ relationships.

Table 5.1 shows the results from the DRUID run. As can be seen, DRUID has found a rather substantial amount of 7th through 10th degree relationships – it reports more of these kinds of relationships than unrelated ones (degree 0). There can be several reasons for this, the first being that there simply is a high degree of relatedness within the population, because of the geographic limitation and potential inbreeding. It can also be due the population being very homogeneous, and hence being considered as more related than it is (this can have been a result from BEAGLE, RefinedIBD as well as DRUID). Another important factor that might have contributed to this is the genetic mapping file provided when creating the .map file (see section 4.2): First, the map-file was based on the CEU-population, i.e. residents in Utah with a northern/western european ancestry – this population might simply not be a perfect fit for the norwegian population in Nord-Trøndelag. Secondly, the recombination rate was set to 1.0 cM/Mb for all genetic mappings and this value was also used for BEAGLE and RefinedIBD. If a recombination rate was provided as well, as opposed to a default, average value, the results might have been different, and it is also possible to provide such data to RefinedIBD and BEAGLE.

Figure 5.1 shows the visual presentation of the output, with lighter colors indicating more distant degrees of relatedness. As can be seen, it looks a little like a ball of yarn, which is natural considering the results from DRUID.

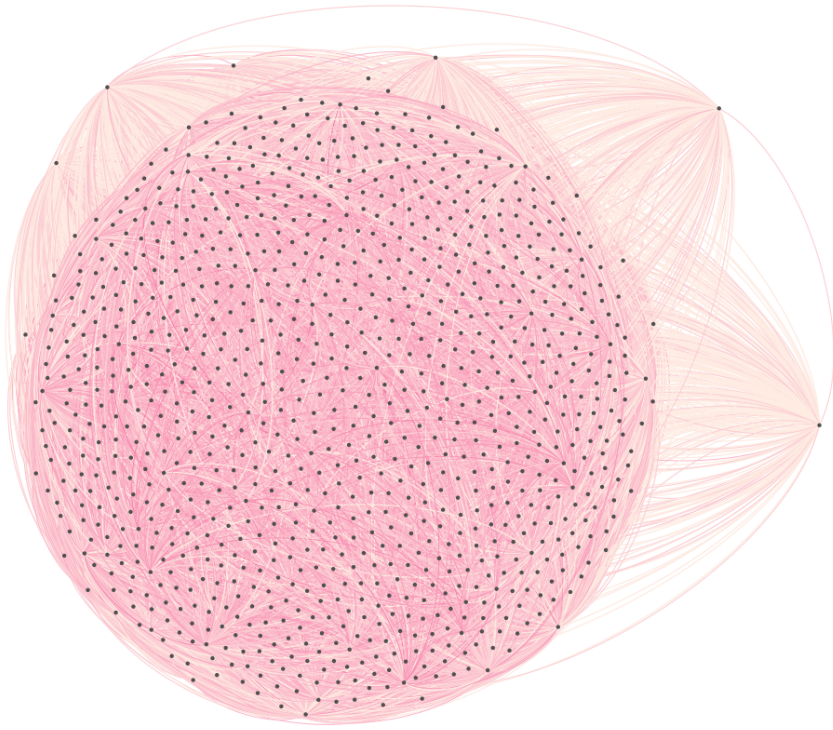

Figure 5.1: The figure shows a graph plot of the output from DRUID. It was created with GEPHI 0.9.2, using the Yifan Hu Multilevel layout, which uses a force-directed model. All edges that had weights of 0 (i.e. unrelated individuals) were removed from the graph.

Nonetheless, this is the results from a single run and cannot be used as a basis for the whole set, although it gives some pinpoints. Results taken from the average of several randomly chosen subsets would still be a much better indicator.

Chapter 6

Conclusion

In this report I have done a literature review on relatedness estimation methods where pedigrees are involved, finding that there are several kinds of methods: Some do both pedigree reconstruction and relatedness estimation; some do mainly relatedness estimation, but utilize pedigrees; and some mainly do pedigree reconstruction.

These methods usually utilize Identity-by-Descent-detection, where two individuals are IBD if they share non-recombinated segments of DNA from a common ancestor. IBD-based methods performs well compared to other approaches, and the tested IBD-based methods by Ramstetter et al. (2017) could detect $>76\%$ of up to 7th degree relationships within one degree of correctness. However, as the degree of relatedness increase, so does the difficulty of detecting IBD-segments. An approach to detect more distant relationships with higher accuracy, has been to leverage multi-way relationships, and looking at several persons DNA compared to eachother.

Some methods also use Composite Likelihood frameworks as opposed to calculating the Maximum Likelihood. The latter can be computationally expensieve, as well as difficult to calculate (e.g. when calculating conditional likelihoods). By using composite likelihoods, the calculations are easier to compute and performance increases, but it might also decrease the accuracy of detection.

When reconstructing pedigrees and searching the space of possible pedigrees, there are different approaches to decreasing the space. PRIMUS uses additional information to decrease the space, but it does nonetheless search the entire space of possible pedigrees, increasing computation time exponentially with pedigree sizes and missing links. CLAPPER uses simulated annealing as an heuristic to avoid having to search the entire space of pedigrees.

Furthermore, methods can utilize mitochondrial DNA and Y-haplotypes, as well as other factors like age and sex to improve their detections. Mitochondrial DNA, Y-haplotypes and sex can be used to reject invalid pedigrees, as is done by Staples et al. (2015) and sex can be used to resolve relationships (e.g. find out who is the father and who is the son in a parent/child relationship) (Ko & Nielsen, 2017).

I have also tested the pedigree reconstruction and relatedness estimation-method DRUID

on the HUNT data. Non-surprisingly, running programs on huge datasets is a very computationally intensive task and takes a long time. Because of memory issues and time-restraints, the data had to be subset considerably, from 70,000 to 1,000 individuals, and because of BEAGLE crashing due to data format issues, the DNA variants where also subset to consider chromosome 1 to chromosome 20.

The script Druid-bakeoff has not been written to handle huge files, which means that if one wants to try running DRUID on the whole dataset, the script either has to be rewritten to handle huge amounts of RAM or a new script would have to be created. However, running DRUID with the complete dataset might be infeasible because the complete dataset might have a higher amount of aunt/uncle–niece/nephew relationships.

The results from DRUID on 1,000 individuals based on chromosome 1 to 20 showed a huge amount of 7th through 10th relationships, which can be the result of geographic limitations and potential inbreeding, but the input files might also have affected the results in different directions. More specifically, the map file was based on the CEU population which might not be a perfect match for the population in Nord-Trøndelag, and all recombination rates were set to 1.0 cM/Mb.

Further work includes testing more methods that reconstruct pedigrees on the HUNT data, focusing on methods that can handle larger datasets (for example looking into methods that utilize composite likelihood frameworks or use other approaches to decrease computation time and memory usage). In addition, the potential inbreeding in the population should be looked into, and whether it is possible to account for this when doing e.g. IBD-detection. It is also interesting to see how potential inbreeding affects mitochondrial relatedness. In addition, it can be useful to see whether the techniques for relatedness estimation can be applied to the mitochondrial DNA as well, as the idea behind multi-way estimation might be applicable to mtDNA, as well as using composite likelihoods to improve performance.

Bibliography

Bodner, M., Irwin, J. A., Coble, M. D., & Parson, W. (2011). Inspecting close maternal relatedness: Towards better mtdna population samples in forensic databases. *Forensic science international. Genetics*, *5*(2), 138–141. doi:10.1016/j.fsigen.2010.10.001

Browning, B. L., & Browning, S. R. (2013a). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*(2), 459–471. doi:10.1534/genetics.113.150029

Browning, B. L., & Browning, S. R. (2013b). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*(2), 459–471. doi:10.1534/genetics.113.150029

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. doi:10.1086/521987

Browning, S. R., & Browning, B. L. (2012). Identity by descent between distant relatives: Detection and applications. *Annual Review of Genetics*, *46*(1), 617–633. doi:10.1146/annurev-genet-110711-155534

Daly, A. K., & Day, C. P. (2001). Candidate gene case-control association studies: Advantages and potential pitfalls. *British journal of clinical pharmacology*, *52*(5), 489–499. doi:10.1046/j.0306-5251.2001.01510.x

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, *23*(1), 34–63. doi:10.1016/0040-5809(83)90004-7

Gauvin, H., Moreau, C., Lefebvre, J.-F., Laprise, C., Vézina, H., Labuda, D., & Roy-Gagnon, M.-H. (2014). Genome-wide patterns of identity-by-descent sharing in the french canadian founder population. *European journal of human genetics : EJHG*, *22*(6), 814–821. doi:10.1038/ejhg.2013.227

Huff, C. D., Witherspoon, D. J., Simonson, T. S., Xing, J., Watkins, W. S., Zhang, Y., … Jorde, L. B. (2011). Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome Research*, *21*(5), 768–774. doi:10.1101/gr.115972.110

Ko, A., & Nielsen, R. (2017). Composite likelihood method for inferring local pedigrees. *PLoS Genetics*, *13*(8). doi:10.1371/journal.pgen.1006963

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., … Holmen, J. (2013). Cohort profile: The hunt study, norway. *Int J Epidemiol*, *42*(4), 968–77. doi:10.1093/ije/dys095

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–73. doi:10.1093/bioinformatics/btq559

Milo, R., & Phillips, R. (2016). *Cell biology by the numbers*. Retrieved from http://book.bionumbers.org/what-is-the-rate-of-recombination/

Pickrell, J. (2014). Interpolated omni. Retrieved December 11, 2018, from https://github.com/joepickrell/1000-genomes-genetic-maps/tree/master/interpolated_OMNI

Purcell, S. (n.d.). Retrieved November 21, 2018, from http://pngu.mgh.harvard.edu/purcell/plink/

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559–75. doi:10.1086/519795

Ramstetter, M. D. (2018). Druid-bakeoff. Retrieved November 22, 2018, from https://github.com/MonicaRamstetter/bakeoff

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., … Williams, A. L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, *207*(1), 75–82. doi:10.1534/genetics.117.1122

Ramstetter, M. D., Shenoy, S. A., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., … Williams, A. L. (2018). Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. *American Journal of Human Genetics*, *103*(1), 30–44. doi:10.1016/j.ajhg.2018.05.008

Staples, J., Ekunwe, L., Lange, E., Wilson, J. G., Nickerson, D. A., & Below, J. E. (2015). Primus: Improving pedigree reconstruction using mitochondrial and y haplotypes. *Bioinformatics*, *32*(4), 596–598. doi:10.1093/bioinformatics/btv618

Staples, J., Qiao, D., Cho, M. H., Silverman, E. K., Nickerson, D. A., & Below, J. E. (2014). Primus: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *American Journal of Human Genetics*, *95*(5), 553–564. doi:10.1016/j.ajhg.2014.10.005

Staples, J., Witherspoon, D. J., Jorde, L. B., Nickerson, D. A., Below, J. E., & Huff, C. D. (2016). Padre: Pedigree-aware distant-relationship estimation. *American Journal of Human Genetics*, *99*(1), 154–162. doi:10.1016/j.ajhg.2016.05.020

The Genetic Map. (2018). Retrieved December 11, 2018, from http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#gmap

Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–26. doi:10.1534/genetics.112.148825

Varin, C., Reid, N., & Firth, D. (2011). *An overview of composite likelihood methods*.

Voight, B. F., & Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS genetics*, *1*(3). doi:10.1371/journal.pgen.0010032

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, *56*(645), 330–338. doi:10.1086/279872

Appendix

NTNU
K.G. Jebsen Center for
Genetic Epidemiology

ALL-IN

GWAS-data

SNPs

This document provides a description of the

handling of the all-in genotyped data.

Document version 2018-09-07

NTNU
K.G. Jebsen Center for
Genetic Epidemiology

Table of content

Changes

Documentation for genotyped data

This document provides a brief description of the handling of the all-in genotyped data from genotyping through QC. The purpose of the document is to provide background information for research using single or multiple SNP's which have been extracted and made available from the total dataset.

Background

From 2012-2015 the HUNT-Michigan (HUNT-MI) collaboration genotyped approximately 72.000 individuals from the HUNT biobank. The genotyping effort was a research collaboration between researchers at NTNU and the University of Michigan. Every individual with a DNA sample with a suitable DNA concentration was selected for genotyping. Samples were picked at random and genotyped in batches. All genotyping was performed at the Genomics-Core Facility (GCF) at the Norwegian University of Science and Technology, NTNU.

Contact list

Kristian Hveem	Leader, K.G. Jebsen center for genetic epidemiology	kristian.hveem@ntnu.no
Maiken E. Gabrielsen	Research coordinator, K.G. Jebsen center for genetic epidemiology	maiken.e.gabrielsen@ntnu.no
Anne Heidi Skogholt	Analysis coordinator, K.G. Jebsen center for genetic epidemiology	anne.heidi.skogholt@ntnu.no
Ben M. Brumpton	Senior Researcher, K.G. Jebsen center for genetic epidemiology	ben.brumpton@ntnu.no
Oddgeir L. Holmen	Leader, HUNT data center	oddgeir.l.holmen@ntnu.no

Acknowledgements

When using gwas - SNP´s in publications, please acknowledge the following:
The genotyping was financed by the National Institute of health (NIH), University of Michigan, The Norwegian Research council, and Central Norway Regional Health Authority and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU).
The genotype quality control and imputation has been conducted by the K.G. Jebsen center for genetic epidemiology, Department of public health and nursing, Faculty of medicine and health sciences, Norwegian University of Science and Technology (NTNU).

Quick overview:

Genotyping platform: Illumina

Chip: HumanCoreExome arrays:

- HumanCoreExome 12 v.1.0
- HumanCoreExome 12 v.1.1
- UM HUNT Biobank v1.0 (HumanCoreExome 24 with custom content)

Imputation: Human reference consortium (HRC) and custom panel including 2200 HUNT individuals with low pass WGS

About the data-set:

The data-set has been produced and quality controlled at the K.G. Jebsen center for genetic epidemiology, NTNU, in collaboration with Associate Professor Willer and Professor Abecasis at the University of Michigan. Below you will find the description of the handling of the original data-set.

All SNPs in the file have been imputed (including the also genotyped SNPs). Genotypes are coded as dosage. The value given is the dosage of the alternative allele.

Quality control

In total, DNA from 71,860 HUNT samples was genotyped using one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). Samples that failed to reach a 99% call rate, had contamination > 2.5% as estimated with BAF Regress (Jun *et al*., 2012), large chromosomal copy number variants, lower call rate of a technical duplicate pair and twins, gonosomal constellations other than XX and XY, or whose inferred sex contradicted the reported gender, were excluded. Samples that passed quality control were analysed in a second round of genotype calling following the Genome Studio quality control protocol described elsewhere (Guo *et al*., 2014). Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; http://genome.ucsc.edu) using BLAT (Dunham et al., 2012). Variants were excluded if (1) their probe sequences could not be perfectly mapped to the reference genome, cluster separation was < 0.3, Gentrain score was < 0.15, showed deviations from Hardy Weinberg equilibrium in unrelated samples of European ancestry with p-value < 0.0001), their call rate was < 99%, or another assay with higher call rate genotyped the same variant.

Ancestry/Population structures

Ancestry of all samples was inferred by projecting all genotyped samples into the space of the principal components of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from http://csg.sph.umich.edu/chaolong/LASER/) (Li et al., 2008; Wang et al., 2014), using

PLINK v1.90 (Chang *et al*., 2015). Recent European ancestry was defined as samples that fell into an ellipsoid spanning exclusively European populations of the HGDP panel. The different arrays were harmonized by reducing to a set of overlapping variants and excluding variants that showed frequency differences > 15% between data sets, or that were monomorphic in one and had MAF > 1% in another data set. The resulting genotype data were phased using Eagle2 v2.3 (Loh *et al*., 2016).

Imputation

Imputation was performed on the 69,716 samples of recent European ancestry using Minimac3 (v2.0.1, http://genome.sph.umich.edu/wiki/Minimac3) (Das *et al*., 2016) with default settings (2.5 Mb reference based chunking with 500kb windows) and a customized Haplotype Reference consortium release 1.1 (HRC v1.1) for autosomal variants and HRC v1.1 for chromosome X variants (McCarthy *et al*., 2016). The customized reference panel represented the merged panel of two reciprocally imputed reference panels: (1) 2,201 low-coverage whole-genome sequences samples from the HUNT study and (2) HRC v1.1 with 1,023 HUNT WGS samples removed before merging. We excluded imputed variants with Rsq < 0.3 resulting in over 24.9 million well-imputed variants.

References

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi:10.1186/s13742-015-0047-8

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nat Genet*. doi:10.1038/ng.3656

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., . . . Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:nature11247 [pii] 10.1038/nature11247

Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., . . . Long, J. (2014). Illumina human exome genotyping array clustering and quality control. *Nat Protoc, 9*(11), 2643-2662.

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., . . . Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet, 91*(5), 839-848. doi:10.1016/j.ajhg.2012.09.004

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., . . . Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104. doi:10.1126/science.1153717

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., . . . Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *bioRxiv*. doi:http://dx.doi.org/10.1101/052308

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., . . . Haplotype Reference, C. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet, 48*(10), 1279-1283. doi:10.1038/ng.3643

Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., . . . Abecasis, G. R. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nat Genet, 46*(4), 409-415. doi:10.1038/ng.2924