Mining sequencing and SNP array data for sex chromosome count

Svanhvít Sigurjónsdóttir

Faculty of Life and Environmental Sciences
University of Iceland
2017
MINING SEQUENCING AND SNP ARRAY DATA
FOR SEX CHROMOSOME COUNT

Svanhvit Sigurjonsdottir

60 ECTS thesis submitted in partial fulfillment of a
Magister Scientiarum degree in Bioinformatics

Advisor
Pall Melsted
Hannes Helgason

Faculty Representative
Arnar Palsson

M.Sc. committee
Pall Melsted
Hannes Helgason

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, May 2017
Abstract

The human sex chromosomes differ from the autosomes in many ways due to the fact that males have two different sex chromosomes instead of a pair. Because of this difference the sex chromosomes are viewed differently in genetic research and the Y chromosome often excluded from association studies. Defining high quality regions of the X and Y chromosomes could benefit association studies and genetic studies in general. In this thesis we attempt to refine the regions used to study the sex chromosomes, develop methods to determine the sex chromosome numbers of sequenced and SNP chip genotyped individuals. The sex chromosome count of individuals could then be used to fix incorrectly reported sex values. We use those sex chromosome numbers to find individuals with human sex chromosome aneuploidy to establish a data set of aneuploidy cases that is not biased against asymptomatic individuals. We used data at deCODE genetics, for 31,918 whole genome sequenced individuals and 162,797 chip genotyped. We compared our results against reported frequencies of three aneuploidy syndromes and at what frequencies different meiotic errors cause trisomy syndromes. The methods developed appear to determine the X chromosome count accurately, and the Trisomy X frequency in the data set is close to the previously reported values. The Y chromosome count appears to be underestimated using these methods and the Turner syndrome frequency is therefore higher and Klinefelter syndrome frequency is lower than expected.
Contents

List of Figures ix
List of Tables xi
Abbreviations xiii
Acknowledgments xv

1 Introduction 1
  1.1 Genetics ................................................. 1
    1.1.1 Chromosomes ........................................ 2
    1.1.2 Single Nucleotide Polymorphisms .................... 2
  1.2 Cell Division ........................................... 3
    1.2.1 Recombination ....................................... 3
  1.3 Human Sex Chromosomes .................................. 4
    1.3.1 Centromere ......................................... 5
    1.3.2 The Pseudoautosomal Regions ....................... 5
    1.3.3 X-transposed Region ................................ 5
    1.3.4 X-degenerate Region ................................ 6
  1.4 Human Sex Chromosome Aneuploidy ....................... 6

2 Background 9
  2.1 Next Generation Sequencing ............................. 9
    2.1.1 Sequencing Coverage ................................ 10
    2.1.2 Read Alignment ..................................... 10
  2.2 Alignment/Map Files ..................................... 11
  2.3 SAMtools ................................................. 11
  2.4 SNP Genotyping on BeadChips ............................ 12
    2.4.1 Log R Ratio ......................................... 13
    2.4.2 B allele Frequency ................................... 13
  2.5 K-means Clustering ..................................... 13
    2.5.1 Elbow Method ....................................... 14
  2.6 Logistic Regression .................................... 15

3 Methods 17
  3.1 Motivation .............................................. 17
  3.2 Establishing the number of sex chromosomes using sequencing data 18
    3.2.1 Chromosome X ...................................... 19
List of Figures

1.1 Representation of the 4 different nucleotides and their complementary bind-
ing. .......................................................... 1

1.2 Cell division in a diploid cell. .................................. 4

1.3 Overview of the different regions of the X and Y chromosomes. ........... 6

2.1 Hybridisation of nucleotides in chip genotyping. .......................... 12

2.2 The elbow method, for finding the optimal number of clusters. .......... 14

3.1 Ranges of normalised sequencing depth of the X chromosome. ......... 20

3.2 Ranges of normalised sequencing depth of the Y chromosome. .......... 21

3.3 The original and redefined regions of chromosome X. .................... 23

3.4 The original and redefined regions of chromosome Y. .................... 24

3.5 The comparison of SeDe between individuals with human sex chromosome
aneuploidy ....................................................... 25

3.6 Distribution of median LRR values ........................................ 30

3.7 Clustering of BAF values. 1 X chromosome ................................. 32

3.8 Clustering of BAF values. 2 X chromosome ................................. 32

3.9 Clustering of BAF values. 3 X chromosome ................................. 33

4.1 An individual that did not pass the confirmation step of the sequencing
method. .......................................................... 37

4.2 The 95% Confidence interval of Trisomy X frequency by chip type. ....... 43
LIST OF FIGURES

4.3 The 95% Confidence interval of Klinefelter syndrome frequency by chip type. 44

4.4 The 95% Confidence interval of Turner syndrome frequency by chip type. . 46

4.5 The 95% Confidence interval of the frequencies of the aneuploidy syndromes for both the chip and sequencing method. . . . . . . . . . . . . . . . . . . . . . . . . . . 47

4.6 The frequency of trisomy cases with parents of an advanced age. . . . . . . 50
## List of Tables

3.1 The normalised sequencing depths and their corresponding chromosome numbers. ...................................... 19

3.2 The changes in median sd and average sd of the median SeDe of all the sequenced individuals by altering the regions analysed. ........................................ 22

3.3 The allele ratios and the corresponding genotypes. ................................................................. 26

3.4 Parent of origin in Trisomy X. The genotypes of the individual with trisomy X along with their parents genotypes. The origin of the extra X chromosome can be determined from the genotypes. ........................................ 27

3.5 Which X chromosome is inherited in cases of Trisomy X where the extra X chromosome is inherited from the mother. The genotypes of the individual with trisomy X along with their parents genotypes. From the genotypes it can be determined whether the offspring has inherited two copies of the same X chromosome from their mother or one of each X chromosome. .... 27

3.6 Parent of origin in Klinefelter syndrome. The genotypes of the individual with Klinefelter syndrome along with their parents genotypes. The origin of the extra X chromosome can be determined from the genotypes. ........ 28

3.7 Which X chromosome is inherited in cases of Klinefelter syndrome where both X chromosomes is inherited from the mother. The genotypes of the Klinefelter case along with their mothers genotypes. From the genotypes it can be determined whether the offspring has inherited two copies of the same X chromosome from their mother or one of each X chromosome. .... 28

3.8 The number of X chromosomes determined by the optimal clustering of BAF values. If the results of the 2 step clustering matched a row where the last column in the table has a "-", the individual was removed from the data set. .................................................. 31

4.1 The number of individuals (N) in the seqDB and the number of cases of each of the aneuploidy syndromes (k). ................................................................. 37
LIST OF TABLES

4.2 The number of individuals analysed on each chip. For each chip type n is the number of individuals, and the following columns represent how many individuals were successfully categorised for the columns chromosome. . . . 40

4.3 Success rate at categorising the number of Y chromosomes using 2 different method. All values are in percentages. The first column for each method shows the percentage of individuals that could be categorised and the second column shows how many of those were correctly categorised in comparison to the seqDB. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

4.4 The number of individuals (N) in the chipDB from each chip (including those with no Y values in the Trisomy X row) and the number of cases of each of the aneuploidy syndromes (k). . . . . . . . . . . . . . . . . . . . . 41

4.5 The number of individuals (N) in the chipDB from each chip (including those with no Y values) and the number of Trisomy X cases (k). . . . 42

4.6 The number of individuals (N) in the chipDB from each chip (not including those with no Y values) and the number of cases of Klinefelter syndrome (k). 45

4.7 The number of individuals (N) in the chipDB from each chip (not including those with no Y values) and the number of cases of Turner syndrome (k). 45
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMA</td>
<td>Advanced Maternal Age</td>
</tr>
<tr>
<td>BAF</td>
<td>B Allele Frequency</td>
</tr>
<tr>
<td>BAM</td>
<td>Binary Alignment Map</td>
</tr>
<tr>
<td>bp</td>
<td>Basepair</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>GT</td>
<td>Genotype</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome Wide Association Studies</td>
</tr>
<tr>
<td>LRR</td>
<td>Log R ratio</td>
</tr>
<tr>
<td>Mb</td>
<td>Megabase</td>
</tr>
<tr>
<td>MI</td>
<td>Meiosis I</td>
</tr>
<tr>
<td>MII</td>
<td>Meiosis II</td>
</tr>
<tr>
<td>nSeDe</td>
<td>Normalised Sequencing Depth</td>
</tr>
<tr>
<td>PAB</td>
<td>Pseudoautosomal Boundary</td>
</tr>
<tr>
<td>PAR</td>
<td>Pseudoautosomal Region</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment Map</td>
</tr>
<tr>
<td>SeDe</td>
<td>Sequencing Depth</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>sd</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole Genome Sequencing</td>
</tr>
<tr>
<td>WSS</td>
<td>Within-Cluster Sum of Squares</td>
</tr>
<tr>
<td>X-deg</td>
<td>X-Degenerate Region</td>
</tr>
<tr>
<td>XTR</td>
<td>X-Transposed Region</td>
</tr>
</tbody>
</table>
Acknowledgments

Firstly I would like to thank deCODE Genetics and specifically Gíslí Másson for giving me the opportunity of doing the research for this MSc thesis, my advisors Páll Melsted and Hannes Helgason for everything they taught me and all the help in moulding this project and writing this thesis. Florian Zink and Patrick Sulem deserve a special thanks for answering all my questions and being incredibly helpful and patient.

I also want to thank Arnar Pálsson at the University of Iceland for all his help and advice.

I would like to thank Arnar Ingi Bragason for all his help, support and most of all patience, I would never have managed this without him. I would also like to thank my parents for all the support one could ever want or need.
1 Introduction

1.1 Genetics

The term DNA, an abbreviation of deoxyribonucleic acid is well known in today’s society and it is common knowledge that DNA is the hereditary material in humans and all other organisms. The DNA codes for all the different traits of an organism like for instance the eye and hair colour of humans. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide (NatureEducation, 2014).

Human DNA is double stranded, which means that the DNA bases pair up with each other by hydrogen bonds (figure 1.1), A with T and C with G, to form units called base pairs. Because of this non-random pairing the two strands are mirrors of each other, they can be read from opposite directions and will give the same information (Shaw, 2008).

![Figure 1.1: Representation of the 4 different nucleotides (A, T, G and C) and (a) their binding to form a single strand. (b) The complementary binding that (c) results in double stranded DNA (NatureEducation, 2014).](image)

An important property of DNA is that it can replicate, or make copies of itself. Each strand can serve as a pattern for duplicating the sequence of bases. This is critical in cell division because each new cell needs to have an exact copy of the DNA sequence from the parent cell.
1 Introduction

1.1.1 Chromosomes

Genome size varies greatly between species, the genome of *Pinus taeda* is the largest sequenced genome at 22.18 billion basepairs (Neale et al., 2014), while the smallest sequenced nuclear genome is that of the *Encephalitozoon cuniculi* at only 2.9 Mb (Corradi et al., 2010). The haploid human genome consists of about 3 billion nucleotides and each human cell contains two copies of the entire genome. For this to be possible the DNA must be tightly packaged into chromosomes. The number of chromosomes differs between species, humans are a diploid species which means that their chromosomes come in pairs of two (Alberts et al., 2008a).

**Human chromosomes**

Humans have 46 chromosomes or 23 pairs, 22 of these pairs are so called autosomal chromosomes and the last pair are the sex chromosomes.

Each autosomes are numbered and are usually referred to by their given number and a 3 letter abbreviation of the word chromosome so that the first chromosome is called chr1. The sex chromosomes are called X and Y, men have one X and one Y chromosomes while women have two X chromosomes (Alberts et al., 2008a).

There is much variation in chromosome length, the largest is chr1 at almost 249 Mb. The X chromosome is about 156 Mb which is a little above average compared to the autosomes. The Y chromosome is the second smallest chromosome around 57 Mb, the smallest is chr22 which is less than 47 Mb (NCBI, 2017a).

1.1.2 Single Nucleotide Polymorphisms

Individuals of the same species will have nearly identical genomes, that is the structure of their chromosomes and the order of nucleotides in their DNA is almost exactly the same. Single nucleotide polymorphisms (SNPs) are aberrations of the nucleotide sequence, similar to typos in a text. If a single nucleotide differs between members of species or between paired chromosomes in an individual that is defined SNP or a variant (Alberts et al., 2008a).

These different variants are called alleles, a variant is often biallelic which means that there at the position of the SNP two different nucleotides can be found within a population. The different nucleotides are the different alleles of the variants and in the case of biallelic variants are often referred to as allele A and allele B (LaFramboise, 2009).

For every SNP one can define the allele frequency, as the frequency of the rare allele in a population sample. Distribution of allele frequencies is one of the estimators of genetic
1.2 Cell Division

There are two types of cell division, mitosis in which a cell divides into two identical daughter cells and meiosis which produces four gametes (sperms or eggs). In meiosis a cell will divide into 4 cells that each have half as many chromosomes as the parent cell. In humans this means that a diploid cell (with 2 sets of chromosomes) divides into 4 haploid daughter cells that each contain only one set of chromosomes (Alberts et al., 2008b).

Meiosis involves one round of DNA replication and two rounds of cell division, meiosis I (MI) and meiosis II (MII). In mammals this process varies greatly between the two genders. In males the germ cells (the precursor to sperm) do not enter meiosis until after puberty. Only a part of the available germ cells enter puberty at any given time so there is constant production of new sperm cells throughout the life of the male. In females the germ cells (the precursor to eggs) start meiosis before birth but the process is arrested mid way. After puberty one germ cell, on average, will re-enter meiosis every menstrual cycle, this means that in females thousands of cells are frozen in the first round of meiotic cell division for decades (O’Connor, 2008b).

1.2.1 Recombination

In meiosis a process called recombination takes place. Recombination is the exchange of genetic material either between homologous chromosomes or within a chromosome. Recombination is a complex process which involves multiple proteins, many steps and is not fully understood. It occurs at the prophase I stage of meiosis, and it involves homologous chromosomes aligning, DNA strands being split and the crossing over of some genetic material. The crossover is called a chiasmata and can be seen in figure 1.2 (Alberts et al., 2008b; O’Connor, 2008b).

Recombination in meiosis ensures that even though gametes only contain one set of chromosomes they still contain genetic material from both sets of chromosomes from the parent cell, that is the offspring of two individuals will have alleles from all four of its grand-

diversity in a population and changes in allele frequency can be used to study the evolution of a population (Cornuet and Luikart, 1996).

For biallelic SNPs (where only two nucleotide variations are found in a population), the alleles can be arranged into three different genotypes. The AA and BB are called homozygous at that locus (position in the genome). If the individual has one copy of each allele (AB) he is heterozygous at that locus. The arrangement of alleles in a specific locus is called a genotype (gt), genotyping is the process of figuring out which alleles an individual has at a specific loci (Cornuet and Luikart, 1996).
1 Introduction

Figure 1.2: Cell division in a diploid cell. (a) Mitosis, a single gene replication round followed by one round of cell division. (b) Meiosis, a single gene replication round as in mitosis followed by two rounds of cell division producing 4 haploid cells. The chiasma (necessary for recombination) is represented in the third cell in figure b and the recombination of genetic material is visible in the following cells (O’Connor, 2008b).

The main purpose of recombination is to pair the corresponding chromosomes together. Errors in meiosis and recombination can lead to aneuploidy, that is the production of 2 cells that do not have the same number of chromosomes. For instance from such an error a daughter cell could have 24 chromosomes, one of each of the autosomes but both of the sex chromosomes and then the other daughter cell lacks both sex chromosomes and only has 22 chromosomes (Barzel and Kupiec, 2008).

1.3 Human Sex Chromosomes

Sex chromosomes have evolved from autosomes independently multiple times. The human sex chromosomes, X and Y, started diverging from each other 200-300 million years
1.3 Human Sex Chromosomes

The current model suggest that inversions on the Y chromosome suppressed recombination between the two chromosomes eventually leading to the accumulation of repeats and loss of gene content on the Y chromosome (Tomaszkiewicz et al., 2016; Trombetta and Cruciani, 2017).

The human sex chromosomes differ from the autosomes in various ways, the Y chromosome has the SRY gene which is the major sex determining gene in humans, in women one of the X chromosomes undergoes inactivation, and the lack of recombination subjects the sex chromosomes to unusual evolutionary forces. The human sex chromosomes have several structural features, some common and some unique to each that are described below (Bachtrog, 2013; Ross et al., 2005).

1.3.1 Centromere

All human chromosomes have a region called the centromere, which has an important role in chromosome segregation in cell division. The centromere has a key role in pairing the two identical chromosomes together in both meiosis and mitosis (O’Connor, 2008a). Centromeres are largely made up of DNA repeats and are impossible to sequence properly in humans with the current technology (figure 1.3).

1.3.2 The Pseudoautosomal Regions

The human sex chromosomes each have two regions of homology called the pseudoautosomal regions (PAR)(figure 1.3). These are situated on the ends of the chromosomes, PAR1 (2.7 Mb in length) is found on the short arms (Xp/Yp) and PAR2 (320 kb in length) on the long arm (Xq/Yq) of the sex chromosomes. The regions have a defined pseudoautosomal boundary (PAB) representing their end points on the chromosomes (Mangs and Morris, 2007).

There is at least one recombination per male meiosis in the PAR1 region, which is a high recombination rate for a region that is less than 3 Mb, compared to what is found on the autosomes. The recombination rate in the PARs is lower in females but still higher than the average autosomal recombination rate (Hinch et al., 2014).

1.3.3 X-transposed Region

A 3.4 Mb region of the X chromosome transposed to the Y chromosome 3-4 million years ago. This is called the X-transposed region (XTR)(figure 1.3) and it only contains 2 genes (Bachtrog, 2013). Recent reports suggest that recombination is possible between the two sex chromosomes in this region (Veerappa et al., 2013). The shared homology between
the X and Y chromosomes in this region can cause issues in genotyping (both sequencing and chip genotyping) since there is no way to know from which chromosome SNPs or regions originate when there are two identical regions on both chromosomes.

### 1.3.4 X-degenerate Region

Some parts of the Y chromosome contain regions that are well conserved amongst primates, these regions are referred to as the X degenerate region. It contains 16 single-copy genes that have homologues on the X chromosome (Bachtrog, 2013).

![Overview of the different regions of the X and Y chromosomes. Areas that have the same colour are homologous between the two sex chromosomes. The PARs and XTR are represented in pink and green respectively. The centromeres are represented as a tightening in the chromosome (Ross et al., 2005).](image)

**Figure 1.3:** Overview of the different regions of the X and Y chromosomes. Areas that have the same colour are homologous between the two sex chromosomes. The PARs and XTR are represented in pink and green respectively. The centromeres are represented as a tightening in the chromosome (Ross et al., 2005).

### 1.4 Human Sex Chromosome Aneuploidy

There are many known chromosome aneuploidy syndromes, most of them cause severe symptoms (i.e. Down’s syndrome, caused by aneuploidy of chr21) and they can often be fatal. The distinctiveness of the sex chromosomes makes sex chromosome aneuploidy different from autosomal aneuploidy. Factors such as the inactivation of one X chromosome
1.4 Human Sex Chromosome Aneuploidy

in healthy females and the low gene content of the Y chromosome can cause extra sex chromosomes to have less of an effect than an extra autosome has.

One of the most common factors of aneuploidy in humans is the advanced age of parents. Advanced maternal age (AMA) in particular, which is most commonly defined as 35 or older, is linked with an increase in maternally derived Klinefelter syndrome and Trisomy X (Hall et al., 2006; Kenny et al., 2013). The most likely cause of this is the waiting period in the meiosis of human egg development (chapter 1.2). This stasis can lead to errors in recombination which then lead to aneuploidy (Hassold et al., 2007).

Turner (45,X), Trisomy X (47, XXX) and Klinefelter (47, XXY) syndromes are all forms of sex chromosome aneuploidy. Turner syndrome (1/2500 live births) is often diagnosed in pregnancy or shortly after birth. Klinefelter syndrome(1/500 live births) is most often diagnosed at puberty, but can be diagnosed earlier especially in the most extreme cases. Trisomy X (1/1000 live births) is diagnosed at all ages but is thought to often go undiagnosed (Bojesen et al., 2003; Calogero et al., 2017; Culen et al., 2017; Lenroot et al., 2014; Nielsen and Wohlert, 1990).

Autosomal trisomies most often originate with the mother (Hassold et al., 2007) so that the offspring inherits one chromosome from their father and two from their mother. Previous reports show the same pattern for Trisomy X where 80.4% of studied cases have two X chromosomes from their mother (Hassold et al., 2007). However Klinefelter syndrome has a different pattern where about half the cases, 50.9% according to (Hassold et al., 2007), inherit both of their father’s sex chromosomes and one X chromosomes from their mother.

Errors in MI are more common for trisomies originating with the mother, about 63% of the total Trisomy X cases and 25.4% of the total Klinefelter cases are cause by MI errors whereas MII errors account for 17.4% and 15.2% respectively (Hassold et al., 2007). MI error cause the offspring to inherit one copy of each of the mothers X chromosomes while MII errors lead to a daughter cell with two copies of the same X chromosome.
2 Background

In this chapter we introduce the technical concepts necessary for the understanding of next chapters. Here we describe the methods used for analysing genomes (whole genome sequencing and chip genotyping), the file formats for storing the sequencing data and the statistical methods used for analysis in chapters 3 and 4.

2.1 Next Generation Sequencing

There are multiple different methods of sequencing available today that vary in complexity, cost and throughput. They all try to achieve the goal of determining the order of nucleotides in the genome, although some methods are more precise than others (Illumina Inc., 2015). Whole genome sequencing (WGS) is a method for analysing the entire genome of an individual. With these newer and more cost effective methods analysing the whole genome of multiple individuals is becoming possible for even small laboratories (Barnes and Breen, 2010).

Early sequencing methods such as the Sanger method and the capillary electrophoresis instruments used for the human Genome Projects were both costly and relatively slow. In 2005 the Genome analyser, the first of the Solexa analysers by Illumina, brought in the next generation of sequencing methods. These newer methods have significantly higher throughput and were more cost effective (Illumina Inc., 2015). The information below is Illumina specific since all the sequencing data used for this thesis was from Illumina machines.

Next generation sequencing (NGS) methods produce thousands of short segments of DNA simultaneously, initially 25-36 bp long and currently around 150-250 bp. This requires a preparation step that cleaves the DNA sample into random smaller fragments. The fragments are amplified so there are multiple copies of each fragment, which are then sequenced. NGS methods use fluorescently labelled nucleotides that bind to their matching nucleotide. The fluorescent signals from the binding nucleotide can then be read to establish the order of the nucleotide in the fragment, the order read from this process is called a read (Illumina Inc., 2015).

A main feature of NGS is paired-end (PE) sequencing which involves sequencing from both ends of the DNA fragment producing 2 reads that can then be aligned. This increases
the accuracy of the method by increasing the length of reads and improving mapping and genotyping (Illumina Inc., 2015).

2 Background

2.1.1 Sequencing Coverage

NGS methods offer researchers the possibility of increasing or decreasing the coverage of sequencing. A coverage of 30X is often the goal in WGS which means that on average each nucleotide has been sequenced 30 times. This number is often referred to as the sequencing depth (SeDe) (Illumina Inc., 2015).

One reason greater SeDe is advantageous is because it increases accuracy. Reads can contain errors and the only way to distinguish between errors and mutations is having multiple reads over the same genomic regions. Typical error rates are about 1-2% per sequenced basepair. If the depth is around 30X and 28 of them contain the nucleotide A at a specific position and the other 2 reads contain a T, this would indicate that the reads that contain T have an error.

If the focus is on a small part of the genome a much greater coverage could be used, but with that comes increased complexity in the protocol used prior to sequencing. Time and cost are usually the controlling factor when the coverage is set.

The SeDe will be used in chapters 3 and 4 to estimate the numbers of chromosomes present in individuals.

2.1.2 Read Alignment

When performing WGS using NGS methods the output will be hundreds of short reads. These need to assembled before the output is useful for research. The problem of assembling sequencing reads is often compared to taking several copies of the same book, passing them through a shredder and then trying to recreate the original book by putting the shredded pieces back together. The output of the sequencing is a jumble of reads with no indication of where they belong in the genome, there could be errors in the reads and areas of the genome with no reads at all. By either assembling the reads de novo (without using any reference) or aligning them to the reference genome the reads provide the sequence of nucleotides in the genome. Some regions such as centromeres can not be sequenced with current methods so there will always be a certain number of gaps in the resulting genome sequence, these areas will cause more problems in de novo assembly than in alignment to the reference genome because the reference genome will show indications of where these problematic regions are (Chenna et al., 2003; Langmead and Salzberg, 2012; Polyanovsky et al., 2011).

Today reference genomes are one of the most important aids in assembling reads. Ref-
2.2 Alignment/Map Files

Reference genomes have been assembled for many species and are updated regularly. The reference genome is not a genome of any single individual, it has been created from the sequencing and assembly of multiple individuals (NCBI, 2017b).

The reference genome has one allele of a SNP (called the reference allele) and newly discovered SNPs are added to it so one can find out which other alleles (alternative allele) can be found at each position.

Alignment algorithms take into account the different polymorphism in the genome, SNPs, insertions and deletions, copy number variations, and structural rearrangements that can occur in the genome, nucleotides can be deleted and inserted from the genome or one base can be changed into another nucleotide (Chenna et al., 2003; Langmead and Salzberg, 2012; Polyanovsky et al., 2011).

2.2 Alignment/Map Files

The sequence alignment/map format (SAM) is often used to store aligned reads from sequencing. SAM files have a header followed by alignment section which stores one alignment per line. The SAM format is TAB delimited and has 11 mandatory fields which describe various read properties. The binary alignment/map format (BAM) is a compressed version of SAM. It is compressed in the BGZF format, which is focused on performance and high compression (Li et al., 2009).

2.3 SAMtools

SAMtools is a library and software package used for retrieving and processing SAM and BAM files. It can be used for sorting and merging alignments, calling SNPs and retrieving information on a per position basis by using a feature called pileup.

The pileup format displays all the reads mapped to a particular position. Each line in a pileup format has 5 columns that show the sequence identifier (such as the chromosome name), position in the genome, the reference allele, the SeDe, a string of the bases found at that position in the aligned reads and optionally a 6th column with the quality of the bases (Li et al., 2009). Base quality is a measure of the quality of the identification of a sequenced base and can be related to base-calling error probabilities (Ewing et al., 2005).
2 Background

2.4 SNP Genotyping on BeadChips

SNP genotyping is the process of determining variation of SNPs in genetic sequences. This can be done on so called SNP Chips (henceforth chips). A chip can capture variation of predetermined set of SNPs (determined by the manufacturer), usually hundreds of thousands of SNPs per chip. This technology relies on the complementary binding of nucleotides. Chips have probes that bind to the nucleotides in the sample being analysed and specialised equipment can detect a signal from this binding. These raw intensity signals are then processed to determine the genotypes of all the SNPs on the chip. The following description will focus on Illumina SNP BeadChips.

To measure the intensity signals Illumina scanners use green and red lasers to excite fluorescent dyes. The colour and the intensity of the signal indicates the genotype of each SNP (figure 2.1).

The output from such a scan will have an intensity reading for both possible alleles of a SNP which are then normalised and processed using software.

SNP genotyping is more cost effective and quicker than WGS with NGS technology but is of course limited to a specific set of SNPs which is a disadvantage compared to the WGS data which covers the whole genome and can detect other types of variants such as INDELs (insertions and deletions of single or multiple nucleotides in the genome) (Katsanis and Katsanis, 2013).

Below we describe two properties of SNP genotyping results. The log R ratio (LRR) and the B allele frequency (BAF) which are then used for the methods described in chapter 3. All chips described here are illumina chip and illumina provides cluster files that are used for the calculation of LRR values (Illumina Inc, 2010).
2.4.1 Log R Ratio

An R value is the intensity reading of a specific fluorescent probe on a chip. The logged ratio of the observed intensity (R) to the expected intensity is the (LRR) which is used to determine copy numbers of SNPs on a chip.

\[
LRR = \log\left(\frac{\text{observed}(R)}{\text{expected}(R)}\right)
\]

If there are two copies present for the SNP (meaning that the individual has two copies of the chromosome that the SNP is located on or that part of the chromosome at least) the LRR value is around 0, but it will be lower if there are fewer copies but higher if there are more than 2 copies (Illumina Inc, 2010; Illumina Inc., 2013).

2.4.2 B allele Frequency

The B Allele Frequency (BAF) is a value calculated for each SNP on a chip. The reference allele and the alternative allele are designated as either A or B and the BAF value represents the fraction of bases that are genotyped as B. It ranges from 0 to 1 and for an organism that is diploid like humans the fractions will all be around 0 (homozygous for the A allele), 0.5 (heterozygous) or 1 (homozygous for the B allele). If however the organism only has one copy of a SNP (like in the case of the X and Y chromosomes in males) the BAF will only have a value of 0 or 1, and in the case of 3 chromosomes (as for Trisomy X) the observed frequencies will be around 0, 0.33, 0.66 and 1 (Illumina Inc, 2010; Illumina Inc., 2013).

2.5 K-means Clustering

Clustering methods have the goal of dividing all observations of a data set into clusters so that each observation in a cluster is more similar to all the other observations in the same cluster than with any observation in any other cluster.

K-means clustering is a clustering method that divides observation into previously determined number (k) of clusters. It is an iterative approach that is repeated until it has minimised the variance within each cluster.

The first step of K-means clustering divides all the observation randomly into k clusters (k being determined by the user). The second step is to calculate the centre of each clusters as the mean of all the observations in each cluster. The observations are now reassigned to the cluster with the closest centre. Step 2 is then iterated until no observations are reassigned (James et al., 2013).
The success of $K$-means clustering can depend on the original random assignments to clusters. For this reason the algorithm should be run multiple times with different initial configurations and chose the best solution, defined as the one with the least variance within clusters.

Deciding the number of clusters can be difficult if the data is noisy for instance. Multiple methods have been developed to find the optimal number for $k$, one of those is the elbow method.

### 2.5.1 Elbow Method

The elbow method uses a value called "within cluster sum of squares" (WSS) which is the sum of the squared differences between each data point and the centre of its cluster. It is a measure of the variation within a cluster.

By running $K$-means for a range of $k$ the WSS values can be compared to estimate how many clusters are necessary to minimise variation. The elbow method gets it name from the plotting of the WSS values for the range of $k$ (figure 2.2). The elbow is point whereby increasing $k$ does not significantly reduce the WSS value (Kodinariya and Makwana, 2013).

![Figure 2.2: The WSS values resulting from the $K$-means clustering using a range of $k$. The red arrow points to the "elbow". In the dataset used for this plot 3 clusters are the optimal number since increasing the number to 4 will not significantly reduce the variance within the clusters (Analytics, 2013).](image)

In chapter 3 we describe how $K$-means clustering used LRR and BAF values to determine the numbers of X chromosomes among chip genotyped individuals.
2.6 Logistic Regression

Classification is the process of predicting categorical variables, for instance predicting whether a person has blue or brown eyes (eye colour being the variable, blue and brown being different categories) based on their genotypes. There are different methods of classification, one of the best known is logistic regression.

Logistic regression can be used on training dataset, to estimate parameters in a model that can be used to classify other data into specific categories. For the example below that would be the coefficients (for instance SNPs) used to categorise eye colour.

The logistic function is as follows:

\[ p(X) = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}} \]  

(2.2)

The training data is used to estimate the value of the coefficients \( \beta_0 \) and \( \beta_1 \) using the maximum likelihood method. For instance in the case of eye colour, if the colour blue is coded as 1 and the colour brown is coded as 0 the coefficients are estimated so that if they would be put into equation 2.2 the results for individuals that have an observed eye colour as blue would be as close to 1 as possible while the results for individuals with an observed eye colour of brown would be as close to zero as possible (James et al., 2013).

In most cases there is more than one variable used as a predictor, in that case multiple logistic regression is used and the logistic function changes to:

\[ p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}} \]  

(2.3)

Where \( X = (X_0, \ldots, X_p) \) is a vector of \( p \) predictors.

In chapter 3 we describe how we deployed logistic regression to determine the numbers of Y chromosomes among the chip genotyped individuals using the sequenced individuals as the training data set.
3 Methods

All data used for the development of the following methods is either whole genome sequencing or chip genotyping data collected, analysed and processed at deCODE Genetics. Only data from Icelandic individuals where utilised. They all have pedigree information in deCODE’s genealogical database including the sex of each individual, their year of birth (and death if applicable) and the codes for their parents.

3.1 Motivation

What the human sex chromosome aneuploidy syndromes have in common is that none of them have a fixed set of symptoms, in fact these syndromes can be hard to diagnose because of the varied symptoms patients display. For instance some women with Trisomy X are asymptomatic while others are infertile, have learning disabilities, experience seizures and multiple other symptoms (Otter et al., 2010).

When symptoms vary greatly there is a danger that research can become biased to the more extreme variations of these syndromes. It would be of great utility to be able to identify sex chromosome aneuploidy from genetic data instead of phenotypes to obtain a data set of individuals that includes those that show little to no symptoms. The methods described in the next chapter use genetic data to establish the number of sex chromosomes of individuals with the aim of offering an unbiased view at sex chromosome aneuploidy.

Another reason for determining the number of sex chromosomes for individuals that have been previously sequenced or chip genotyped is for comparison with reported sex. Having the knowledge of correct sex for individuals is necessary for genome wide association studies, especially if they are to include the sex chromosomes.

Determining the number of X chromosomes should be simple for the sequenced individuals. The X chromosomes sequencing data covers almost all the basepairs with no large gaps except for the centromere. This could be more complicated for the chip genotyped individuals since the chips do not cover the entire chromosome but a fixed set of SNPs.

We expect that determining the Y chromosome count to be more difficult for both the WGS and chip data. Large areas of the Y chromosome are highly repetitive which cause issues in sequence alignment. It is also possible that regions that share homology with the
3 Methods

X chromosome will be incorrectly mapped so that reads that should be mapped to the X chromosome are mapped to the Y and vice versa which could then inflate or decrease the SeDe of the Y chromosome. For the chip genotyped individuals the same issue of homology between the chromosomes could be present. Chips often have very few Y chromosome SNPs which causes further issues in the analysis of the data.

3.2 Establishing the number of sex chromosomes using sequencing data

Sequencing depth (SeDe) can be a good indicator of how many chromosomes individuals have of each chromosome pair. That is, if there is aneuploidy present in a chromosome pair the SeDe will be higher (if there is an extra chromosome) or lower (if there is a missing chromosome) than the SeDe of other chromosomes. This is because there are more (or less) copies of the relevant chromosome region. Using this information the number of sex chromosomes for each individual can be calculated by analysing the ratio between the SeDe of different chromosomes.

Since SeDe can be set during sequencing, different individuals will have different coverage. For this reason it is insufficient to only look at at the SeDe of the sex chromosomes. The SeDe needs to be normalised, for instance with values from the autosomes of the same individual.

It is possible to calculate the average SeDe over the entire autosome by dividing the total number of reads by the total number of basepairs for chromosomes 1-22. There could be deviations in the accuracy of the average autosomal SeDe if there is for instance an aneuploidy present in one of the autosomal chromosome pairs or if there was an error in the sequencing process. For this reason the SeDe of the sex chromosome was also compared to the SeDe of chromosomes 6, 7 and 8 (chosen because they are the autosomes most similar in length to chromosome X). The median SeDe of 6, 7 and 8 was calculated both over the entire chromosome and in 1Mb windows. There are two reasons why the median SeDe was calculated for the individual chromosomes while the average SeDe was calculated for the whole autosome. For one, variations in SeDe (for instance chromosome ends and centromeres often have a higher SeDe) will affect the average SeDe less when calculated for the entire autosome (a higher proportion of the total value) than it would for a single chromosome. The more significant reason is that the size of the autosome makes it difficult to calculate the median SeDe since the mean only needs the number of reads and the total basepairs of the autosome while the calculation of the median involves sorting the sequencing depth of all the basepairs and finding the median value. This is hardware demanding, especially when done for thousands of individuals.
3.2 Establishing the number of sex chromosomes using sequencing data

3.2.1 Chromosome X

The median SeDe of the X chromosome, excluding the PARs, the XTR, and the base pairs closest to the centromere, was calculated over the whole chromosome and in 1 Mb windows. The normalised SeDe (nSeDe) was then calculated using the average autosomal SeDe, and the median depths of chromosomes 6, 7 and 8 separately. Individuals with a standard deviation (sd) of the SeDe higher than 5 were removed from the data set.

The four different nSeDes were then compared to pre-defined ranges (table 3.1) if the average autosomal nSeDe and 2 of the other three fell into the same range the individual was categorised according to that range. Otherwise the individual was removed from the data set. The ranges are visualised in figures 3.1.

Table 3.1: The normalised sequencing depths and their corresponding chromosome numbers.

<table>
<thead>
<tr>
<th>nSeDe</th>
<th>Number of chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.4-0.6</td>
<td>1</td>
</tr>
<tr>
<td>0.9-1.1</td>
<td>2</td>
</tr>
<tr>
<td>1.4-1.6</td>
<td>3</td>
</tr>
</tbody>
</table>

3.2.2 Chromosome Y

A large portion of the long arm of the Y chromosomes, referred to as Yq12, is mainly heterochromatic causing some problems in the processing of sequencing data. The median depth of the Y chromosome was therefore calculated excluding the PARs, XTR and Yq12. The median depth of the Y chromosome was also calculated in 1Mb windows excluding the same regions. The categorisation of the number of Y chromosomes for individuals was performed in the same way as for the X chromosome, using the same range for the ratios (table 3.1, figure 3.2).

3.2.3 Redefining the chromosome regions used for analysis

After visualising the nSeDe of a number of individuals it became clear that it was necessary to redefine the regions used for the analysis on both chromosomes (figures 3.3 and 3.4). The coordinates of the original regions were from the GRCh38/hg38 reference assembly. It has been suggested that the definition of the PAB is conservative, that the rate of recombination is still unusually high well past the PAB indicating that the PAR regions may be larger than now thought (Cotter et al., 2016). The visualisation of the variation in the nSeDe shown in figures 3.3 and 3.4 shows an increased variation past the PAB of all four PARs. If there is recombination between the X and the Y chromosome past
3 Methods

(a) Distribution of sequencing depth for the X chromosome

Figure 3.1: (a) The SeDe of reads mapping to the chromosomes the X chromosome over the average autosomal sequencing depth for all the sequenced individuals. (b) The distribution of the nSeDe for the X chromosome. The values of n represent the number of individuals in that range. The coloured lines represent the ranges from table 3.1.

(b) Distribution of sequencing depth for the Y chromosome

the PAB this could lead to basepairs being mapped to the wrong chromosome. For this reason the regions used for analysis were reduced. Table 3.2 shows the differences in the median and average sd of the median SeDe of all the sequenced individuals using the whole sex chromosomes, the originally defined regions and the redefined regions. The median
3.2 Establishing the number of sex chromosomes using sequencing data

Figure 3.2: (a) The SeDe of reads mapping to the Y chromosome over the average autosomal sequencing depth for all the sequenced individuals. (b) The distribution of the nSeDe for the Y chromosome. The n values represent the number of individuals in that range. The coloured lines represent the ranges from table 3.1.

and average standard deviation decreases when the regions used are reduced from the whole chromosome to the originally defined regions and then again be using the redefined regions.
Table 3.2: The changes in median sd and average sd of the median SeDe of all the sequenced individuals by altering the regions analysed.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Whole chromosome</th>
<th>Original regions</th>
<th>Redefined regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median sd</td>
<td>average sd</td>
<td>median sd</td>
</tr>
<tr>
<td>X</td>
<td>1.99</td>
<td>2.10</td>
<td>0.76</td>
</tr>
<tr>
<td>Y</td>
<td>39.07</td>
<td>92.69</td>
<td>1.52</td>
</tr>
</tbody>
</table>

3.2.4 Confirmation of human sex chromosome aneuploidy

The individuals that had categorisation of their sex chromosomes in accordance with the previously described aneuploidy syndromes were subjected to confirmation step to establish that the increased or reduced nSeDe was stable over the entire chromosome and not just a sequencing error. For this the median depth in 1MB windows were used. The median SeDe in each window was normalised in the same way as the SeDe of the total chromosome but only using the average autosomal depth.

A comparison curve was then constructed by randomly choosing 100 females and 100 males (categorised as the same gender by the previously described method as was found in the reported information on the individuals). Based on the median SeDe in 1 MB windows they were normalised in the same way as the individuals with sex chromosome aneuploidy. After this normalisation the median of each 1Mb window was calculated, creating 1 curve for each gender, for both chromosomes.

Figure 3.5 shows this median of the X and Y chromosomes plotted with the comparison curves for individuals with clinically characterised trisomy X, Klinefelter syndrome and Turner syndrome.

The values of the median curve were then subtracted from the value of the individual in each window. If this subtraction yielded similar values for each window (mean = 0.5 ± 0.1 and sd < 0.05) the aneuploidy call was deemed successful.
3.2 Establishing the number of sex chromosomes using sequencing data

Figure 3.3: The original and redefined regions of chromosome X. The shaded areas are regions excluded from the analysis. Each plot point is the median normalised sequencing depth of a 0.1 Mb window of this randomly chosen individual. (a) shows the original definition of the regions and for both PARs and the centromere the increased variation in nSeDe is evident. (b) shows the redefined regions that now contain the windows with increased variance closest to the excluded regions.
Figure 3.4: The original and redefined regions of chromosome Y. As for figure 3.3 the variance is not contained within the shaded regions of figure (a) but is mostly contained within the redefined regions of figure (b).
3.2 Establishing the number of sex chromosomes using sequencing data

Figure 3.5: The comparison of SeDe between individuals with human sex chromosome aneuploidy (case) to a comparison curve created by the median SeDe of 100 individuals (control) of the same gender as the case in each figure. The control is shown in black and the case in red. Each plot point represents the normalised median SeDe of a 1 Mb window. A nSeDe of 1 represents two chromosomes present. An increase or decrease in the nSeDe by 0.5 means there is 1 more or 1 less chromosomes present respectively. Note that the Y chromosome is shorter than the X chromosome so the x-axis of the two are not comparable. This is the reason for fewer plot points for the Y chromosome. (A) A woman with Turner syndrome. The case has a genotype of 1X0Y and the depth change from the control is steady over the whole chromosome. (b) A man with Klinefelter syndrome, this case has a genotype of 2X1Y. (c) A woman with Trisomy X, this case has a genotype of 3X0Y.
3 Methods

3.3 Identifying the parent of origin of extra X chromosomes

For the individuals that can be categorised as having Trisomy X or Klinefelter syndrome we developed a method to identify from which parent the extra X chromosome was inherited.

For this method the trisomy X or Klinefelter individual had to have been sequenced using NGS and their parents had to have been either sequenced using NGS or chip genotyped.

Using a fixed set of SNPs it is possible to compare the genotypes of the individuals with aneuploidy and their parents to determine the origin of the extra X chromosome.

The fixed set of SNPs was created from the chip types the parents had been genotyped on. The set of SNP therefore contained SNPs from different chips so no chip typed individual had all the SNPs in the set. For each SNP the genotype of the offspring, the father and the mother (often referred to as a trio) was determined and the three genotypes compared. As previously explained in chapter 1.1.2 an individual should have one allele of each SNP on each of their chromosomes. So to determine the origin of the X chromosomes we need to establish how many copies of the reference allele and the alternative allele each individual of the trio has for each SNP available.

For the individuals that had been sequenced this was performed using the pileup program which is a part of the SAMtools package (chapter 2.3). Each SNP was called and the number of reads with the reference allele and the alternative allele counted to determine the genotype. The ratio of the alternative allele was calculated for each SNP, by dividing the number of occurrences by the SeDe of the position in question, and depending on the individual of the trio (father has 1 X chromosome, the mother has 2 and the offspring has 3) in question the genotype was determined (table 3.3). If the ratio did not fall within these ranges the SNP was removed.

<table>
<thead>
<tr>
<th>Chromosomes</th>
<th>Ratio</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.060 &lt;</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>&gt; 0.940</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.060 &lt;</td>
<td>0/0</td>
</tr>
<tr>
<td>2</td>
<td>0.440 - 0.560</td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td>&gt; 0.940</td>
<td>1/1</td>
</tr>
<tr>
<td></td>
<td>0.060 &lt;</td>
<td>0/0/0</td>
</tr>
<tr>
<td>3</td>
<td>0.273 - 0.393</td>
<td>0/0/1</td>
</tr>
<tr>
<td></td>
<td>0.606 - 0.726</td>
<td>0/1/1</td>
</tr>
<tr>
<td></td>
<td>&gt; 0.940</td>
<td>1/1/1</td>
</tr>
</tbody>
</table>
3.3 Identifying the parent of origin of extra X chromosomes

### 3.3.1 Parent of origin in trisomy X

To determine the origin of the three X chromosomes the offspring has to be heterozygous for a SNP, for instance have 2 chromosomes with the reference allele and 1 with the alternative allele or vice versa. The mother on the other hand has to be homozygous for a different allele than the father has (table 3.4). The father only has one allele since men only have one X chromosome. Calculation of this is shown in table 3.4.

The origin of each SNP in the set, that met the necessary criteria, was determined. Then by combining the results of the whole SNP set the origin of the X chromosomes could be established.

**Table 3.4: Parent of origin in Trisomy X. The genotypes of the individual with trisomy X along with their parents genotypes. The origin of the extra X chromosome can be determined from the genotypes.**

<table>
<thead>
<tr>
<th>Offspring gt</th>
<th>Father gt</th>
<th>Mother gt</th>
<th>Parent of Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0/1</td>
<td>0</td>
<td>1/1</td>
<td>Father</td>
</tr>
<tr>
<td>0/0/1</td>
<td>1</td>
<td>0/0</td>
<td>Mother</td>
</tr>
<tr>
<td>0/1/1</td>
<td>0</td>
<td>1/1</td>
<td>Mother</td>
</tr>
<tr>
<td>0/1/1</td>
<td>1</td>
<td>0/0</td>
<td>Father</td>
</tr>
</tbody>
</table>

**Identifying which of the mother’s chromosomes were passed to the offspring**

If the mother is the parent of origin for 2 of the 3 X chromosomes then it is possible to establish if the offspring has inherited 2 copies of the same X chromosome of one copy of each of their mother’s X chromosomes (table 3.5)

**Table 3.5: Which X chromosome is inherited in cases of Trisomy X where the extra X chromosome is inherited from the mother. The genotypes of the individual with trisomy X along with their parents genotypes. From the genotypes it can be determined whether the offspring has inherited two copies of the same X chromosome from their mother or one of each X chromosome.**

<table>
<thead>
<tr>
<th>Offspring gt</th>
<th>Father gt</th>
<th>Mother gt</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0/0</td>
<td>0</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
<tr>
<td>0/0/1</td>
<td>0</td>
<td>0/1</td>
<td>Different chromosomes</td>
</tr>
<tr>
<td>0/1/1</td>
<td>0</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
<tr>
<td>1/1/1</td>
<td>1</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
<tr>
<td>0/1/1</td>
<td>1</td>
<td>0/1</td>
<td>Different chromosomes</td>
</tr>
<tr>
<td>0/0/1</td>
<td>1</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
</tbody>
</table>
3 Methods

3.3.2 Parent of origin in Klinefelter syndrome

Determining the parent of origin in Klinefelter cases is done in a similar way. If the mother is the parent of origin for both X chromosomes then the offspring needs to be homozygous for the SNP to be useful. If the father is the parent of origin for one of the X chromosomes then the offspring needs to be heterozygous for the SNP (table 3.6).

Table 3.6: Parent of origin in Klinefelter syndrome. The genotypes of the individual with Klinefelter syndrome along with their parents genotypes. The origin of the extra X chromosome can be determined from the genotypes.

<table>
<thead>
<tr>
<th>Offspring gt</th>
<th>Father gt</th>
<th>Mother gt</th>
<th>Parent of origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>1</td>
<td>0/1</td>
<td>Mother</td>
</tr>
<tr>
<td>0/0</td>
<td>1</td>
<td>0/0</td>
<td>Mother</td>
</tr>
<tr>
<td>1/1</td>
<td>0</td>
<td>0/1</td>
<td>Mother</td>
</tr>
<tr>
<td>1/1</td>
<td>0</td>
<td>1/1</td>
<td>Mother</td>
</tr>
<tr>
<td>0/1</td>
<td>1</td>
<td>1/1</td>
<td>Father</td>
</tr>
<tr>
<td>0/1</td>
<td>0</td>
<td>0/0</td>
<td>Father</td>
</tr>
</tbody>
</table>

Identifying which of the mother’s chromosomes were passed to the offspring

If the mother has been established as the parent of origin for both X chromosomes it is possible to determine which of the two maternal X chromosomes were inherited (table 3.7).

Table 3.7: Which X chromosome is inherited in cases of Klinefelter syndrome where both X chromosomes is inherited from the mother. The genotypes of the Klinefelter case along with their mothers genotypes. From the genotypes it can be determined whether the offspring has inherited two copies of the same X chromosome from their mother or one of each X chromosome.

<table>
<thead>
<tr>
<th>Offspring gt</th>
<th>Mother gt</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/0</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
<tr>
<td>0/1</td>
<td>0/1</td>
<td>Different chromosomes</td>
</tr>
<tr>
<td>1/1</td>
<td>0/1</td>
<td>Same chromosome</td>
</tr>
</tbody>
</table>

3.4 Establishing the number of X chromosomes using chip data

To develop a method for accurately determining the number of sex chromosomes an individual has using their chip genotyping data is more difficult than using the sequencing
3.4 Establishing the number of X chromosomes using chip data

data. Sequencing data is more reliable than the chip data but also more expensive.

For this reason the sequencing data is used as a gold standard for developing the method for the chip data. This depends on some subset of individuals to have been both chip genotyped and sequenced.

If working with more than one chip type the following two steps need to be performed separately for each chip type to avoid differences in chip types affecting the results.

The LRR and BAF values were calculated using software developed in-house at deCODE genetics.(chapter 2.4).

3.4.1 Step 1: Clustering individuals based on LRR values

The LRR value is the best indicator of copy number, individuals with 2 copies of a SNP will have a higher LRR value for that SNP than individuals with only 1 copy. This fact was used to roughly separate individuals into 1 of 5 categories. LRR and BAF values (calculated from the intensity readings) were available for all SNPs for each individual genotyped.

First all the SNPs position on chromosome X (excluding the PARs and the XTR) were extracted for each individual and the median LRR value of all those SNPs calculated.

Figure 3.6 shows the distribution of median LRR values among individuals chip genotyped on the same chip. The distribution is bimodal (has 2 peaks) since there is a normal distribution of median LRR values for the individuals with 1 X chromosome (lower LRR) and for individuals with 2 X chromosomes (higher LRR).

K-means clustering was used to cluster all the individuals into 2 clusters based on median LRR values. The sd of the median LRR values within each cluster was calculated and all individuals within 2 standard deviations of the center of a cluster were categorised as either having 1 X or 2 X chromosomes depending on which cluster they belonged to.

The other 3 categories were individuals that had a median LRR value that was either too high or too low to be categorised and finally the individuals that had a median LRR value between the two clusters. All individuals in these 3 outlier categories were then subjected to the second step of the analysis.

3.4.2 Step 2: Clustering SNPs based on BAF values

For this second step of the analysis all individuals that had been deemed outliers by the first step were analysed separately, since these outlier groups should contain all aneuploidy
Figure 3.6: Distribution of median LRR values from one randomly chosen chip type. Individuals clustered into the lower LRR value cluster are coloured red while the ones in the higher LRR value cluster are coloured black. The dotted lines represent the cut off values for categorisation into 1 or 2 X chromosomes. Individuals above, below or between the two peaks were categorised as high, low or middle outliers respectively.

individuals. For this step the focus was on the BAF values of all the SNPs of the X chromosome (using only the regions described in chapter 3.1.3).

BAF values can be easily clustered as explained in chapter 2.2.2, there are only a limited option of values for each SNP and the options (and with it the clusters) increase with increased number of chromosomes.

K-means clustering was used to cluster the SNPs based on the BAF value and individuals categorised as having 1, 2 or 3 X chromosomes based on how many clusters the BAF values were best clustered into. Figures 3.6-3.8 show the clustering of BAF values for 3 different individuals with different numbers of X chromosomes.

The BAF values were clustered using K-means clustering with $k$ ranging from 1 to 5 and then the elbow method was used to determine which $k$ resulted in the best clustering. Table 3.8 shows the categorisation of individuals based on their clustering.

The clustering was performed using the \texttt{kmeans} function in R, and for the elbow method the \texttt{tot.withinss} value (an output value from the \texttt{km} function) was used. It is a
3.4 Establishing the number of X chromosomes using chip data

Table 3.8: The number of X chromosomes determined by the optimal clustering of BAF values. If the results of the 2 step clustering matched a row where the last column in the table has a "-", the individual was removed from the data set.

<table>
<thead>
<tr>
<th>k</th>
<th>outlier group</th>
<th>Number of X chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>low</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>middle</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>middle</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>middle</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>3</td>
</tr>
</tbody>
</table>

Algorithm 1 Determining X chromosome number using the elbow method

1: \textbf{procedure} COMPARE \textsc{tot.withinss} VALUES
2: \hspace{1em} if $t_5 - t_4 > 15$ then
3: \hspace{2em} $X \leftarrow NaN$
4: \hspace{1em} else if $t_4 - t_3 > 20$ then
5: \hspace{2em} $X \leftarrow 3$
6: \hspace{1em} else if $t_3 - t_2 > 100$ then
7: \hspace{2em} $X \leftarrow 2$
8: \hspace{1em} else if $t_2 - t_1 > 100$ then
9: \hspace{2em} $X \leftarrow 1$
10: \hspace{1em} else
11: \hspace{2.5em} $X \leftarrow NaN$

The clustering results has to be used in conjunction with the median LRR value since individuals with significant loss of X (the degeneration of the X chromosome with age) can have BAF values that cluster in the same way as individuals with 3 X chromosomes but have a lowered LRR value. These individuals were removed from the data sets.

BAF values for SNPs are clearly defined so long as the chip genotyping was successful. For this reason determining the optimal number of clusters is simple for BAF values in comparison with data with more noise or more factors affecting each value for instance.
3 Methods

Figure 3.7: Clustering of BAF values. Each data point represents the BAF value of a SNP. The colours represent the different clusters. The figure shows the clustering of the BAF values of the SNPs of an individual with 1 X chromosome.

Figure 3.8: Clustering of BAF values. Each data point represents the BAF value of a SNP. The colours represent the different clusters. The figure shows the clustering of the BAF values of the SNPs of an individual with 2 X chromosome.
3.4 Establishing the number of X chromosomes using chip data

Figure 3.9: Clustering of BAF values. Each data point represents the BAF value of a SNP. The colours represent the different clusters. The figure shows the clustering of the BAF values of the SNPs of an individual with 3 X chromosome.
3.5 Establishing the number of Y chromosomes using chip data

The chip genotyping data for chromosome Y is not of sufficient quality to determine the number of chromosomes using the same method as used for the X chromosome. See table 4.3 for a comparison of success for the two methods for chromosome Y. For this reason a different method was developed. Using the results from the sequencing data we set out to create a logistic regression model to estimate the number of Y chromosomes an individual has based on which SNPs the individual has genotyped on the chip. A set of individuals that have been both sequenced and chip genotyped was used to get create a multiple logistic regression model.

For each individual chip type a matrix was calculated for each individual. The matrix contained a 1 for all the SNPs that had been genotyped on the Y chromosome during the chip genotyping and a zero for the others. This matrix was limited to the SNPs positioned in the same region as used for the sequencing method described in chapter 3.1.3. Before creating the model noninformative SNPs are removed, as well as the most highly correlated SNP pairs (over 99% correlation) were determined and one SNP removed from each pair.

We then defined $A_{i,j}$ where $i$ is an individual and $j$ is a SNP. Each individual has either a 0 or a 1 value for each SNP depending on whether that chip has a genotype for that individual (1) or not (0).

$p(A_i)$ was then calculated for the sequenced individuals by determining $\beta_j$, coding the $y$ value for all the individuals that had been determined to possess a Y chromosome by the sequencing method as 1 and the others received a $y$ value of 0. Using the glm function in R a prediction model was created on a per chip bases. The model could then be used to calculate $p(A_i)$ for the chip typed individuals so that individuals with a $p(A_i)$ value higher that 0.5 were determined to have a Y chromosome while those with a lower $p(A_i)$ values do not.

After creating the model for the logistic regression the number of Y chromosomes can be estimated for all the individuals that have been chip genotyped using the model for their respective chip type.

This should be done separately for each chip type since there are differences in the set of SNPs on each chip type which would confound the results.
4 Results

The results of the experiments in this section were obtained from SNP chip and sequencing data from deCODE genetics (Gudbjartsson et al., 2015). The dataset contains 167,228 Icelanders with chip genotyping data and 31,918 with WGS data. All regions used are from the GRCh38/hg38 reference assembly.

4.1 Finding cases of sex chromosome aneuploidy using sequencing data

The SeDe was calculated from the BAM files of each individual. Some individuals have been sequenced more than once, in that case only one set of data was used preferably where the DNA had been extracted from a blood sample. If a blood sample was not an option (other samples were from buccal swabs) or applied to more than one data set, then the newest sequencing was chosen.

Certain regions were skipped on both chromosomes, discussed in chapter 3.1.3. The positions used in the analysis of chromosome X were 2.8-57.5 Mb, 63.5-88.0 Mb and 93.0-155.7 Mb. For the Y chromosome the positions used were 6.8-10.1 and 10.7-26.0 Mb.

The number of bam files was 31,918, 332 of those were removed since they represented the same individuals. The sd of the median SeDe of the X chromosome was calculated, and 59 individuals were removed from the data set because of an abnormally high sd (sd > 5). Finally 224 individuals were removed due to low median SeDe of the X chromosome (SeDe < 10).

We analysed the X chromosome status with the ratio comparison described in the methods chapter. 31,250 individuals could be categorised for chromosome X while 33 could not be classified. Of the ones that could be categorised for chromosome 30,730 could be categorised for the Y chromosome also while 520 could not.

The results described above were the outcome of the revised regions on both the X and Y chromosomes described in chapter 3.1.3. Before this redefinition of the regions used 283 individuals were removed due to their high sd and 144 individuals were removed because they had a median depth less than 10. This means 144 additional individuals could be included in the analysis by narrowing the used regions on the chromosome. For the
4 Results

redefinition of the Y chromosome the number of individuals that could not be categorised went down from 820 to 520.

At this point a database (called the seqDB) was created from the categorisation results, it contained information on each individual categorised for number of X chromosomes stating the number of X and Y chromosomes (if the individual could not be categorised for chromosome Y the database has an NA value), if the individual fits the criteria of Turner syndrome, Trisomy X or Klinefelter, which gender criteria they match and which gender is recorded in the genealogical database.

There were 16,418 individuals that fit the criteria of female (XX genotype) while 16,412 of those are supported by the genealogical database. 14,237 individuals fit the criteria of male (XY genotype) while 14,233 of those are supported by the genealogical database. There were 520 individuals that were not assigned a gender by the chromosome categorisation, these are the ones which could not be categorised for the Y chromosome, of those 510 are male according to the genealogical database and 10 are female. The X chromosomes of these 520 individuals all match the gender of the genealogical database.

4.1.1 Trisomy X cases in the sequencing data

As previously stated in chapter 1.4 individuals with Trisomy X have 3 X chromosomes, a genotype called 47,XXX. There were 16 individuals with 3 X chromosomes and no Y chromosome. There were no individuals categorised as having 3 X chromosomes that could not be categorised for the Y chromosome. This is a frequency of 0.098% which is close to the previously reported frequency of 0.1% (Lenroot et al., 2014). The expected frequency falls within the 95% confidence interval (table 4.1 and figure 4.5). All of them passed the confirmation step and they were all classified as female in the genealogical database.

4.1.2 Klinefelter syndrome cases in the sequencing data

Individuals with Klinefelter syndrome are males with 2 X chromosomes, a genotype called 47,XXY. There were 21 individuals with 2 X chromosomes and 1 Y chromosome, and none with 2 X and no value for Y chromosome that the genealogical database specified as male. These 21 individuals were all subjected to the confirmation step described in chapter 3.1.4, and as a result 2 were removed (figure 4.1 shows the nSeDe of one of them) this does not exclude the possibility that these individuals are Klinefelter cases, it does mean that their sequencing data is more varied than other individuals and therefore could be of lower quality.

This left 19 individuals with Klinefelter syndrome which is a frequency of 0.133%, the previously reported frequency is 0.2% (Bojesen et al., 2003). The previously reported
4.1 Finding cases of sex chromosome aneuploidy using sequencing data

frequency falls within the 95% confidence interval of the frequency of the cases (table 4.1 and figure 4.5).

Figure 4.1: The normalised sequencing depth of an individual that was identified as having Klinefelter syndrome by the sequencing method but failed the confirmation step (described in chapter 3.1.4) due to the erratic distribution of the nSeDe of chromosome X. Each plot point represents the median depth of a 1 Mb window, the red represent the removed individual and the black are the male X chromosome comparison curve (described in chapter 3.1.4). Only the regions used for analysis are shown and the horizontal lines show the nSeDe range for 2 X chromosomes.

4.1.3 Turner syndrome cases in the sequencing data

Individuals with Turner syndrome are females with only one X chromosome that is also their only sex chromosome, they have a genotype of 45,X. There were 2 individuals with 1 X chromosome and 0 Y chromosomes whose pedigree information in deCODE’s genealogical database classifies them as female. However there are 22 additional individual with a 45,X genotype in the seqDB that are classified as male by the genealogical database. If there are 2 individuals with Turner syndrome that is a frequency of 0.012%, if there are 24 cases that is a frequency of 0.150%. The previously reported frequency is 0.04% it does not fall within the 95% confidence interval of the frequency for the cases in the data set (table 4.1 and figure 4.5) (Culen et al., 2017).

Table 4.1: The number of individuals (N) in the seqDB and the number of cases of each of the aneuploidy syndromes (k).

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>N</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trisomy X</td>
<td>16239</td>
<td>16</td>
</tr>
<tr>
<td>Klinefelter Syndrome</td>
<td>14285</td>
<td>19</td>
</tr>
<tr>
<td>Turner Syndrome</td>
<td>16239</td>
<td>24</td>
</tr>
</tbody>
</table>
4.1.4 Individuals with Y chromosome aneuploidy

There were 14 individuals in the seqDB with a genotype of 47,XYY. This is a recognised human sex chromosome aneuploidy but due to the fact that the method used to estimate the number of Y chromosomes for the chip data is a trained method based on the sequencing data, and the fact that these 14 individuals were almost evenly distributed among chip types (never more than 2 per chip type) there was not sufficient statistical power to train the logistic regression to find individuals among the chip typed individuals with an extra Y chromosome. Therefore these 14 individuals were not used for the training of data nor for the comparison of methods.

4.2 Parent of Origin

We set out to determine the parent of origin of the X chromosomes in trisomy cases and the number of trisomies caused by errors in meiosis I and meiosis II. We compared the genotypes of trisomy cases and their parents. With this comparison we attempted to determine from which parent the X chromosomes of the trisomy case originated. In the cases where two X chromosomes were inherited from the mother we performed a second comparison of genotypes in an attempt to determine if the offspring had inherited two copies of the same chromosome or one copy of each of the mothers X chromosomes.

4.2.1 Parent of origin in Trisomy X cases

Of the 16 individuals with trisomy X in the seqDB 5 had parents that had been either sequenced or chip genotyped.

These 5 women had 196-240 SNPs that fit the criteria necessary to determine parent of origin. The mother was determined to be the parent of origin (offspring inherited 2 X chromosomes from their mother and one from their father) in 4 of the 5 cases. In the fifth case the father was the parent of origin (2 X chromosomes from the father and 1 from the mother).

Determining which of the mothers sex chromosomes are inherited

The four cases where the mother was the parent of origin were then analysed again this time searching for SNPs that could determine which of the mothers chromosomes had been inherited. The women had 172-384 SNPs that fit the criteria. For one case the origin could not be determined. Two of the cases had inherited 1 copy of each X chromosomes
4.3 Finding cases of sex chromosome aneuploidy using chip data

from their mother (MI error) while 1 case had inherited 2 copies of the same chromosome (MII error).

4.2.2 Parent of origin in Klinefelter syndrome cases

Of the 19 individuals with Klinefelter in the database 9 had parents that had either been sequenced or chip genotyped.

The 9 cases had between 253-750 SNPs that fit the criteria necessary to determine the parent of origin. Five of the cases had inherited both copies of the X chromosome from their mother while 4 had inherited 1 copy from each parent.

Determining which of the mothers sex chromosomes are inherited

Of the 5 cases where the mother was the established origin of both X chromosomes 2 individuals had inherited 2 different X chromosomes (MI error) while the other 3 had inherited 2 copies of the same chromosome (MII error).

4.3 Finding cases of sex chromosome aneuploidy using chip data

All the chip genotype data files (n=167,228) were put through quality control steps, removing the ones with less than 100,000 SNPs genotyped.

They were grouped by chip type and the SNPs positioned on the X and Y chromosomes were extracted. For each chip the methods previously described in chapters 3.3 and 3.4 were used to analyse how many X and Y chromosomes each individual has. That is we used the LRR and BAF value clustering methods to determine the X chromosomes counts and logistic regression trained on the sequencing data for the Y chromosome counts. Table 4.1 shows how many individuals were analysed on each chip, and for how many it was possible to estimate the number of X and Y chromosomes.

Of the 19 chip types used for the X analysis only 8 were used for the analysis of the Y chromosome. This is because the chip types either have no SNPs on the Y chromosome or had too few (analysis was not done for chip types with less than 100 useable SNPs). Table 4.3 shows the difference between using the same method as is used for the X chromosome and using logistic regression to categorise the Y chromosome number.

All in all 162,797 entries could be categorised for at least one of the sex chromosomes.
4 Results

Table 4.2: The number of individuals analysed on each chip. For each chip type \( n \) is the number of individuals, and the following columns represent how many individuals were successfully categorised for the columns chromosome.

<table>
<thead>
<tr>
<th>Chip</th>
<th>( n )</th>
<th>( X )</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanOmniExpress-12v1-1 B</td>
<td>20,304</td>
<td>20,248</td>
<td>20,279</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1 H</td>
<td>31,366</td>
<td>31,135</td>
<td>32,528</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1-Multi H</td>
<td>2908</td>
<td>2,903</td>
<td>0</td>
</tr>
<tr>
<td>Human1Mv1 C</td>
<td>758</td>
<td>508</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni1M-Duov3 B</td>
<td>575</td>
<td>561</td>
<td>573</td>
</tr>
<tr>
<td>HumanOmni1-Quad v1-0 B</td>
<td>11,238</td>
<td>11,182</td>
<td>10,785</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-0 A</td>
<td>36,029</td>
<td>35,805</td>
<td>27,737</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-1 A</td>
<td>1398</td>
<td>1,393</td>
<td>1,390</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1 H</td>
<td>2727</td>
<td>2,575</td>
<td>2,687</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1-Multi H</td>
<td>428</td>
<td>426</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni2.5-8v1 A</td>
<td>4137</td>
<td>4,170</td>
<td>0</td>
</tr>
<tr>
<td>HumanHap300 (v1.0.0)</td>
<td>16,461</td>
<td>16,114</td>
<td>0</td>
</tr>
<tr>
<td>HumanHap300v2 A</td>
<td>7,005</td>
<td>6,888</td>
<td>0</td>
</tr>
<tr>
<td>HumanCNV370v1 C</td>
<td>14,540</td>
<td>14259</td>
<td>0</td>
</tr>
<tr>
<td>HumanCNV370-Quadv3 C</td>
<td>306</td>
<td>306</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni5-4v1 B</td>
<td>709</td>
<td>709</td>
<td>0</td>
</tr>
<tr>
<td>Human610-Quadv1 B</td>
<td>672</td>
<td>663</td>
<td>0</td>
</tr>
<tr>
<td>Decode OEx 8A</td>
<td>13,021</td>
<td>12,952</td>
<td>13,019</td>
</tr>
</tbody>
</table>

There were 7,762 individuals who had more than one entry, of those there were 3 who had differing numbers of X chromosomes estimated on different chip types, they were removed from the data set. For the ones that had the same X chromosome number in all entries only one entry was kept (the newest one), and the corresponding Y value was included so that the X and Y value always come from the same chip type.

There were 154,626 individuals in the final data set which was then compiled into a database (from here on called the chipDB) containing entries on the number of X and Y chromosomes, whether or not the individual fit the criteria for one of the three human sex chromosome aneuploidy syndromes, which gender their chromosome number matched and finally the gender according to the genealogical database. All entries had numbers for the X chromosome while 108,547 had a number for the Y chromosome.

There were 56,714 individuals classified as female by both the chipDB and deCODE’s genealogical database. There were 142 individuals classified as female by the chipDB but male in the genealogical database. There were 51,588 individuals classified as male by both the chipDB and the genealogical database and 103 classified as male by the chipDB but female in the genealogical database. Finally there were 46,079 individuals that received no gender classification in the chipDB due to the fact that they could not be categorised for the Y chromosome. Of those 26,289 are female according to the genealogical database (57.1%) and 19,790 male (42.9%).
### Table 4.3: Success rate at categorising the number of Y chromosomes using 2 different methods. All values are in percentages. The first column for each method shows the percentage of individuals that could be categorised and the second column shows how many of those were correctly categorised in comparison to the seqDB.

<table>
<thead>
<tr>
<th>Chip</th>
<th>Original method</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>categorised</td>
<td>correct</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1-1 B</td>
<td>84.2</td>
<td>52.2</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1 H</td>
<td>48.1</td>
<td>99.6</td>
</tr>
<tr>
<td>HumanOmni1M-Duv3 B</td>
<td>44.5</td>
<td>100</td>
</tr>
<tr>
<td>HumanOmni1-Quad v1-0 B</td>
<td>65.7</td>
<td>42.3</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-0 A</td>
<td>74.5</td>
<td>51.9</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-1 A</td>
<td>60.4</td>
<td>19.8</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1 H</td>
<td>97.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Decode OEx 8A</td>
<td>54.1</td>
<td>98.0</td>
</tr>
</tbody>
</table>

The chipDB contains a total of 83,106 individuals that are classified as female in the genealogical database and 71,520 males. This is the number used in chapters 4.3.1-4.3.3 when the frequency is calculated including the individuals that have no Y value, otherwise the number of individuals categorised as the relevant gender in the chipDB is used.

Table 4.4 shows the number of cases of each aneuploidy syndrome, all entries in the chipDB were used for Trisomy X but only the ones with a value for the Y chromosomes were used for Turner and Klinefelter syndrome. Note that only the individuals of the corresponding gender are used for frequency calculations so in table XX the number of individuals represent the number of individuals of the same gender as for the respective aneuploidy syndrome.

### Table 4.4: The number of individuals (N) in the chipDB from each chip (including those with no Y values in the Trisomy X row) and the number of cases of each of the aneuploidy syndromes (k).

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>N</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trisomy X</td>
<td>83106</td>
<td>55</td>
</tr>
<tr>
<td>Turner Syndrome</td>
<td>56856</td>
<td>98</td>
</tr>
<tr>
<td>Klinefelter Syndrome</td>
<td>51691</td>
<td>103</td>
</tr>
</tbody>
</table>

### 4.3.1 Trisomy X cases in the chip genotyping data

There were 37 individuals with 3 X chromosomes and no Y chromosomes. There were 18 more with 3 X chromosomes but no value available for the Y chromosome. All of these 55 individuals were female according to the genealogical database. This is a frequency of 0.065% using only the individuals with Y chromosome data available and of 0.066%
if the ones with no Y value are included which is substantially lower than the expected frequency of 0.1% (Lenroot et al., 2014).

Tables 4.4 and 4.5 show the total number of Trisomy X cases and the number of cases broken up by chip type. Figures 4.2 and 4.5 show the 95% confidence interval of the Trisomy X frequencies.

As can be seen in figure 4.2 the chip type HumanOmniExpress-12v1-H is the only one (of the chips with enough individuals typed to be expected to have Trisomy X cases) whose confidence interval does not cover the previously reported frequency of 0.1%, this chip type needs to be inspected further to establish why the methods of determining chromosome X count do not seem to work as well for it as for the others.

<table>
<thead>
<tr>
<th>Chip</th>
<th>N</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanOmniExpress-12v1-1 B</td>
<td>10113</td>
<td>6</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1 H</td>
<td>16773</td>
<td>5</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1-Multi H</td>
<td>1861</td>
<td>0</td>
</tr>
<tr>
<td>Human1Mv1 C</td>
<td>292</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni1M-Duov3 B</td>
<td>281</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni1-Quad v1-0 B</td>
<td>6427</td>
<td>4</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-0 A</td>
<td>17803</td>
<td>21</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-1 A</td>
<td>553</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1 H</td>
<td>1574</td>
<td>1</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1-Multi H</td>
<td>199</td>
<td>0</td>
</tr>
<tr>
<td>HumanOmni2.5-8v1 A</td>
<td>2096</td>
<td>1</td>
</tr>
<tr>
<td>HumanHap300 (v1.0.0)</td>
<td>8043</td>
<td>7</td>
</tr>
<tr>
<td>HumanHap300v2 A</td>
<td>4808</td>
<td>5</td>
</tr>
<tr>
<td>HumanCNV370v1 C</td>
<td>8059</td>
<td>4</td>
</tr>
<tr>
<td>HumanCNV370-Quadv3 C</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>HumanOmni5-4v1 B</td>
<td>213</td>
<td>0</td>
</tr>
<tr>
<td>Human610-Quadv1 B</td>
<td>283</td>
<td>0</td>
</tr>
<tr>
<td>Decode OEx 8A</td>
<td>3655</td>
<td></td>
</tr>
</tbody>
</table>

### 4.3.2 Klinefelter syndrome cases in the chip genotyping data

For the Klinefelter syndrome 3 different numbers need to be considered those who have 2X chromosomes and 1 Y chromosome (n = 103, class 1), those with that same genotype and were male according to the genealogical database (n = 77, class 2) and finally the individuals that have the 2X chromosomes, are male according to the genealogical database but have no value for the Y chromosome (n = 52, class 3).
4.3 Finding cases of sex chromosome aneuploidy using chip data

Figure 4.2: The 95% Confidence interval of Trisomy X frequency by chip type. The plot points represent the calculated frequency and the bars the 95% confidence interval. The red line represents the previously reported frequency of 0.1%. All values are in logarithmic scale.

Considering only class 1 this is a frequency of 0.199%, for class 2 the frequency is 0.149%. If we limit the analyses to male individuals based on the pedigree information, that is classes 1 and 3, the frequency is 0.217% and finally the frequency for classes 2 and 3 combined is 0.224%. The frequency for class 1 is used in table 4.3 since that is the frequency of all individuals determined to have a 47,XXY genotype. The expected frequency is 0.2% (Bojesen et al., 2003).

The 95% confidence interval for the different chip types and for the whole chip data set...
4 Results

(with Y chromosome values) can be seen in tables 4.4 and 4.6 respectively and in figures 4.3 and 4.5.

The chip called HumanOmni2.5-4v1-H is the only one whose 95% confidence interval does not cover the expected frequency and therefore needs further inspection.

Figure 4.3: The 95% Confidence interval of Klinefelter syndrome frequency by chip type. The plot points represent the calculated frequency and the bars the 95% confidence interval. The red line represents the previously reported frequency of 0.2%. All values are in logarithmic scale.
4.3 Finding cases of sex chromosome aneuploidy using chip data

Table 4.6: The number of individuals (N) in the chipDB from each chip (not including those with no Y values) and the number of cases of Klinefelter syndrome (k).

<table>
<thead>
<tr>
<th>Chip</th>
<th>N</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanOmniExpress-12v1-1 B</td>
<td>9785</td>
<td>21</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1 H</td>
<td>14125</td>
<td>19</td>
</tr>
<tr>
<td>HumanOmni1M-Duov3 B</td>
<td>270</td>
<td>1</td>
</tr>
<tr>
<td>HumanOmni1-Quad v1-0 B</td>
<td>4117</td>
<td>4</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-0 A</td>
<td>17644</td>
<td>34</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-1 A</td>
<td>739</td>
<td>3</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1 H</td>
<td>938</td>
<td>16</td>
</tr>
<tr>
<td>Decode OEx 8A</td>
<td>4073</td>
<td>5</td>
</tr>
</tbody>
</table>

4.3.3 Turner syndrome cases in the chip genotyping data

In the chipDB there were 98 individuals with 1X chromosome and no Y chromosomes, 34 of them are classified as female in the genealogical database. This is a frequency of 0.063% using only the female individuals otherwise the frequency is 0.177%. There are 49 individuals with 1X and no value for the Y chromosome that are female according to the genealogical database. If all the individuals with one X chromosome and 0 Y are combined with the 1X individuals that have no Y value but are female according to their pedigree data the frequency of Turner syndrome is 0.171%. The expected frequency is 0.04% (Culen et al., 2017).

The 95% confidence interval of the different chip types and for the entire chip data set (that has Y values) can be seen in figures 4.4 and 4.5 and the number of cases by chip type in table 4.7 and the number of total cases in table 4.4.

Table 4.7: The number of individuals (N) in the chipDB from each chip (not including those with no Y values) and the number of cases of Turner syndrome (k).

<table>
<thead>
<tr>
<th>Chip</th>
<th>N</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanOmniExpress-12v1-1 B</td>
<td>10091</td>
<td>11</td>
</tr>
<tr>
<td>HumanOmniExpress-12v1 H</td>
<td>16767</td>
<td>25</td>
</tr>
<tr>
<td>HumanOmni1M-Duov3 B</td>
<td>281</td>
<td>13</td>
</tr>
<tr>
<td>HumanOmni1-Quad v1-0 B</td>
<td>6133</td>
<td>24</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-0 A</td>
<td>17803</td>
<td>3</td>
</tr>
<tr>
<td>HumanOmniExpress-24v1-1 A</td>
<td>553</td>
<td>3</td>
</tr>
<tr>
<td>HumanOmni2.5-4v1 H</td>
<td>1548</td>
<td>10</td>
</tr>
<tr>
<td>Decode OEx 8A</td>
<td>3655</td>
<td>9</td>
</tr>
</tbody>
</table>

45
Figure 4.4: The 95% Confidence interval of Turner syndrome frequency by chip type. The plot points represent the calculated frequency and the bars the 95% confidence interval. The red line represents the previously reported frequency of 0.04%. All values are in logarithmic scale.
4.3 Finding cases of sex chromosome aneuploidy using chip data

Figure 4.5: The 95% Confidence interval of the frequencies of the aneuploidy syndromes for both the chip and sequencing method. The plot points represent the calculated frequency and the bars the 95% confidence interval. The vertical lines represent the previously reported frequency values, (a) for Trisomy X or 0.1%, (b) for Klinefelter or 0.2% and (c) for Turner or 0.04%. All values are in logarithmic scale.
4 Results

4.4 Comparing the sequencing and chip data results

To validate the results of the chip method they were compared to the results from the sequencing method. There were 26,821 individuals that have been sequenced and chip genotyped and had values for the X chromosome. 26,813 of those had the same results for both methods. That is an error rate of 0.029%. Since the logistic regression model was trained on the Y chromosome sequencing data results it would not be accurate to simply compare the results of the two methods. For this reasons 1000 individuals that had been sequenced since the beginning of this project were not included in the sequencing data until after the development of all methods and all analysis had been concluded. They were then categorised for chromosome numbers using the sequencing method and compared to the chip results that had already been processed. Of those there were 2 individuals that had different numbers of Y chromosomes for the two different methods this is an error rate of 0.2%.

4.4.1 Aneuploidy comparison

For the three sex chromosome aneuploidy syndromes focused on here there was a double confirmation step in the database comparison. First the individuals that were classified as having a specific syndrome by the sequencing method were compared to the chip method checking whether they had the same chromosome numbers in the chip method. Then the same confirmation in the opposite direction. This was to confirm both that the chip method finds the same individuals as the sequencing method, that the chip method does not incorrectly classify individuals as having sex chromosome aneuploidy and that it does not miss any individuals with sex chromosome aneuploidy.

There were 7 individuals with Turner syndrome, 10 with Trisomy X and 8 with Klinefelter syndrome in the chipDB. For Trisomy X the sequencing method confirmed all the individuals found in the chip data and it contained no additional individuals with the syndromes. For Klinefelter all 8 individuals in the chipDB were confirmed by the sequencing method but the seqDB included 4 additional individuals with Klinefelter syndrome that all have an XY genotype in the chipDB. Finally for Turner syndrome 5 of the 7 individuals were confirmed by the seqDB while the other 2 have a Y chromosome according to the seqDB.

4.5 Parental age of human chromosome aneuploidy cases

For all the individuals determined to have Trisomy X or Klinefelter syndrome by the sequencing or chip methods we looked at the genealogical database to determine how many had mothers over 35 years of age at their time of birth.
Of the 55 cases of Trisomy X 54 had the mother’s year of birth in the data set. Of those, 14 had mothers of advanced age. The median age of the mothers among the Trisomy X cases was 38 and the average age was 30.15. As for the fathers age, 53 cases had the father’s year of birth recorded and of those there were 18 cases of the father being over the age of 35 when the Trisomy X case was born. That is 33.3% of the Trisomy X cases, the median age of the fathers was 30 and the average age was 32.43 years.

There were 107 cases of Klinefelter syndrome (103 from chip and the 4 additional from the WGS data), 106 of those had maternal age in the genealogical database and 105 paternal age. Of those 25 had mothers over the age of 35 at their birth, the median age of the mothers was 27 and the average age was 28.65 years. There were 38 cases of the fathers being over 35, the median age of the fathers when the Klinefelter offspring was born was 30 and the average age was 32.21 years.

To create a comparison frequency all Icelandic individuals in the deCODE data were compiled (excluding those with trisomies in the chipDB and seqDB) and the age of their parents calculated. There were 155,401 individuals with the mothers age available, and of those 27,965 had been born to a mother of AMA. This is a frequency of 18%. We also looked at paternal age since although it is not as highly correlated with human sex chromosome aneuploidy there were cases of the extra chromosome being inherited from the father in both syndromes (chapter 4.2). There were 154,533 individuals where we could calculate their fathers age at birth and of those 48,062 were born to fathers 35 or older which is a frequency of 31%.

The comparison frequency was used to create figure 4.6, where the 95% confidence interval of the frequencies of advanced parental age of the trisomy cases are visualised. Neither the maternal age nor the paternal age of the trisomy cases are significantly different from the rest of the data set.

Of the 14 Trisomy X and Klinefelter cases where the parent of origin of the X chromosomes could be determined (chapter 4.2) only 1 case had a parent of advanced age. That was a Klinefelter syndrome case, where the extra chromosome was determined to be inherited from the mother. The mother of that case was 35 when she gave birth, which is the youngest age considered as AMA.
Figure 4.6: Each plot point represents the frequency of cases with a parent 35 years or older. (a) Frequency of mothers 35 years or older and (b) the fathers. The red lines represent the frequency in the rest of the deCODE data set.
5 Conclusions

5.1 There is an underestimation of Y chromosome counts in both methods

When table 4.3 is considered the frequencies of Turner syndrome are the ones that deviate the most from the previously reported values. The methods we developed seem to genotype to many individuals as 45,X indicating that multiple individuals that have a Y chromosomes are determined to not have a Y chromosome. This is true for the sequencing method and since the chip method is trained on the sequencing data it has this same pattern. It is therefore necessary to refine the regions used on the Y chromosome analysis further or change some parameters for instance the nSeDe range to correctly determine the Y chromosome numbers for all individuals. This could be done by first examining all the individuals determined to have a 45,X genotype and train the sequencing method further on their Y chromosomes.

5.2 The chip method may underestimates the Trisomy X cases

The 95% confidence interval of the frequency of the Trisomy X cases discovered by the chip method does not cover the previously reported frequency of Trisomy X. It could be that individuals with 3X are being filtered out in the preprocessing of the chip data or the filters of the chip method. It should be repeated though that the chip method does find all the trisomy X individuals determined by the sequencing method. Another possibility is that there simply are fewer cases of Trisomy X in the chip data set. Individuals with symptoms of various disorders have been prioritised for sequencing and individuals collected as control cases have often only been chip genotyped. This could cause the chip data set to have a lower frequency of certain syndromes.
5 Conclusions

5.3 A redefinition is needed for the regions for sex chromosome studies

Human sex chromosomes differ from the autosomes in multiple ways. Since women have two X chromosomes and men have one X and one Y chromosomes the recombination is different between the sexes. For the Y chromosomes recombination is limited to the two PAR regions. By examining the median SeDe of over 20,000 individuals we have shown that the areas closest to the PAB still have a much higher variation in SeDe than most of the X and Y chromosomes. This should be considered when attempting any kind of association studies on the sex chromosomes, results in the regions around 1-2Mb from the PAB are possibly not to be trusted due to the inferior quality of the sequencing and chip data in that region.

Further use of the individuals that have been both chip genotyped and sequenced would be to refine the usable regions of each chromosome. By comparing the results of the two genotyping methods it could be possible to draw clear boundaries of reliable regions for use in future research such as GWAS. Which at this point the Y chromosome is most often not included in, and the X chromosome is included excluding the PARs.

5.4 Newer chip types yield similar quality of results as sequencing data

For the purpose of analysing the number of sex chromosomes an individual has, a more accurate answer can be reached using sequencing data than chip data. Also a higher percentage of individuals can be categorised using the sequencing data. This is because of the multiple reads of the sequencing method confirming most basepairs multiple times (depending on the coverage set for the sequencing), and also because the sequencing data gives a better overview of the entire chromosome rather than a fixed set of SNPs. However by modelling the method used to analyse the chip genotyping data on individuals that have been both sequenced and chip genotyped we managed to categorise the chip genotyped individuals with only 0.029% and 0.2% error rates for the X and Y chromosomes respectively. It is also worth mentioning that the efficiency of the categorisation of the chip genotyped individuals is higher in the newer chips, the Human Omni Express chips (table 4.1 and 4.2).
5.5 Determining sex chromosome numbers for comparison to reported gender

When performing association studies errors in reported sex can greatly affect the results. We demonstrate 256 individuals whose reported sex is not aligned with their sex chromosome categorisation. Having a tool like the methods described here to locate possible errors in reported sex can be valuable for further association studies. Having already demonstrated that there is an underestimation of the Y chromosome count there is a likelihood that a certain number of these individuals are males that are being categorised as individuals with Turner Syndrome. It would therefore be useful to refine the method for Y chromosome analysis before categorising them as having an incorrectly reported gender.

5.6 Different causes of recombination errors could be the cause of aneuploidy

While AMA is the leading known cause of human sex chromosome aneuploidy only 27.6% of the Trisomy X and Klinefelter syndrome cases found here have a mother 35 or older at birth. This means that there are other influential factors causing error in recombination. While the parent of origin results described in chapter 4.2 are aligned with previous reports (around 50% of Klinefelter cases originate from the father, while 80% of Trisomy X cases originate in the mother) (Hassold et al., 2007) the data used for this study does not have the power to separate cases of meiotic and mitotic errors causing the aneuploidy. This is because of a lack of cases where both parents have been sequenced.

Comparing the frequency of cases called in the deCODE dataset to previously reported frequencies is difficult because there are too few cases in the sequencing information and even fewer with genotypes for both parents which is necessary for the determination of parent of origin and the MI and MII error rates. To do a proper comparison of these frequencies a larger database of cases would be needed.

Since Klinefelter syndrome is the only trisomy disorder known to have a similar number of cases where the extra chromosome is inherited from the father as the mother and recombination of the X and Y chromosomes in male meiosis is restricted to the PARs it could be interesting to look into association studies of the PARs (particularly PAR1 which has the highest recombination rate of the human genome) and Klinefelter syndrome to determine whether the PARs are a contributing factor to the recombination errors that cause human sex chromosome aneuploidy. For this to be possible advancements need to be made in producing high quality genotype data of the PARs.
Bibliography


Hassold, T., Hall, H., and Hunt, P. (2007). The origin of human aneuploidy: Where we have been, where we are going. Human Molecular Genetics, 16(R2):203–208.


57


