Universiteit Utrecht

A framework for de-*novo* mutations discovery in Next Generation Sequencing data

Master Thesis Report – ICA 3779130

Authors:

Mircea Cretu Stancu

Supervisors:

Dr. A.J. Feelders

Laurent Francioli

Prof. Dr. Paul I.W. de Bakker

Universitair Medisch Centrum
Utrecht

Table of Contents

1. Introduction 4

1.1. Genome Structure 4

1.1.1. Chromosome 5

1.1.2. Gene 6

1.2. Inheritance 7

1.3. Evolution 9

1.3.1. Mutation 9

1.3.2. Natural Selection 9

1.3.3. Genetic drift 10

1.4. Importance of de-*novo* mutations 11

1.4.1. Mutation rates and distributions 12

1.4.2. Role of de-*novo* mutations in disease 13

2. Materials and Methods 14

2.1. Sequencing 14

2.2. Next Generation Sequencing 14

2.2.1. DNA Amplification 15

2.2.2. DNA Sequencing 16

2.3. Genome Analysis ToolKit – GATK 18

2.3.1. Walkers 19

2.4. Alignment 19

2.4.1. Base Quality Score Recalibration 20

2.5. Variant calling 22

2.5.1. Variant Quality Score Recalibration 24

2.6. Genome of The Netherlands 24

2.6.1. The X chromosome 25

2.7. PhaseByTransmission 26

2.7.1. Pedigree information 26

2.7.2. The Model 27

2.7.3. Extending for the X chromosome 28
2.7.4. Further extensions 30
2.7.5. Machine Learning post-filtering 31
2.8. HaploidWriter Walker 31
2.9. De*Novo*MutationPowerCaller 34
2.9.1. Motivation 34
2.9.2. Simulation data 34
2.9.3. Model description 35
2.9.4. Selection of the model and training 36
3. Results and Discussion 42
3.1. PhaseByTransmission 42
3.1.1. Initial Runs 42
3.1.2. Processing Results 43
3.1.3. The Allele Frequency prior 44
3.1.4. Comparison of Results..45
3.1.5. Explaining Results 46
3.1.6. Validation sets 50
3.1.7. Complete Genomics Data 51
3.1.8. More sensitive initial calls 52
3.1.9. Father Age Effect 53
3.2. HaploidWriter 54
3.3. De-NovoMutationPowerCaller 58
3.3.1. Adjusting the bins 58
4. Conclusions 60
5. References 61

Research Question

For the purpose of this thesis project, we address the problem of accurately and efficiently identifying de-*novo* mutations in the human germline. More precisely, how can we detect de-*novo* point mutations on the sex chromosome X in a robust yet sensible manner? What are the challenges that arise from the quality of the available data for this chromosome? What is the pattern of de-*novo* events on this chromosome, compared to the rest of our genome?

The challenge of devising a discovery method for such events comes from their rarity relative to the error rates of the underlying technology involved in DNA reading. We discuss the relevance of this research in the light of our increasing understanding of evolution and our genetic code's structure and function, as well as its practical applications of finding genetic disease risk factors.

We present the field's currently most used analysis methods and technologies, and describe each step that influences the design and/or performance of the model we implement. We present a straightforward yet efficient general model of de-*novo* mutations discovery and then show how the model needs to be adapted in order to correctly capture the particularities of the X chromosome. Furthermore we illustrate what information can be explained by our model and where we still need to apply domain knowledge to correct the output.

Finally, we show how the model is integrated in the complex and modular analysis pipeline used in the community. Furthermore, we create additional tools that enable this integration and/or profile our model's behaviour under different conditions.

1. Introduction

The DNA code is a basic element describing any living organism morphologically as well as functionally. Research of variation in this code within and across species is continuously adding to our knowledge of its structure and functional implications on the carrier, from a bottom-up perspective. Heritability in this context refers to the extent by which DNA explains a living organism.

Under the hypothesis that human traits are genetically tractable, the study of heritability revolves around mapping genetic variation to observable and/or measurable human traits. The incremental chemical and biological set of processes, by which such heritable components manifest on the complex organism level, are collectively defined as a biological pathway. One specific example is the class of DNA sequences that are directly translated to proteins. The chemical reactions that these proteins undergo further define or influence functions of the organism or traits that are observable and interpretable by humans. These traits, assumed to be influenced by variation in the genetic code, are called phenotypes. They may be pathological, if they are proven to be related to disease, or non-pathological.

Considering the enormous size of our genetic code, the complex low-level interactions within it as well as the complex and stratified mechanisms that transform our DNA code into observable and meaningful outcomes, the variation space is prohibitively large for an exhaustive, analytical understanding of its underlying structure and function.

Diseases have been found to have a heritable component and they affect millions of people around the globe. It is therefore important to investigate variation that can be shown to be associated and/or causal to disease. By causing a disease, diagnosed as such clinically, on the basis of observed phenotypes, this type of genetic variation causes large (qualitative or quantitative) changes in subsequent biological pathways. By definition then, it is easier to single out the set of individuals carrying a disease as well as variation in their DNA w.r.t. other, healthy individuals that do not present the disease causing phenotypes.

1.1 Genome Structure

Symbolically, DNA can be represented as an arbitrary string or sequence, of four literals $\{A, T, C, G\}$ called bases. The human genome has an estimated average length of $3 * 10^9$ bases. Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) are the four nitrogen containing molecules called nucleobases (or simply bases). Due to their chemical properties, each base in the sequence forms base-pair hydrogen bonds with a corresponding base to form the chemically and geometrically stable structure of a double stranded helix (i.e.: DNA sequence) as depicted in Figure1, that is further folded for compactness.

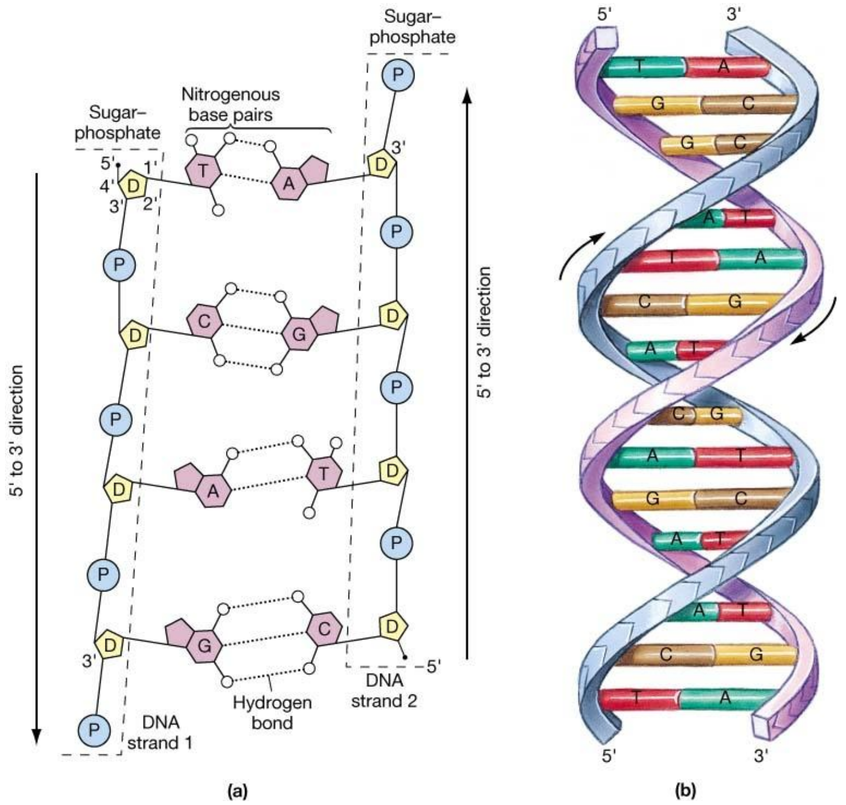

Figure 1: DNA Structure. (a) "unfolded" view of a double-stranded DNA molecule, showing the two chains of nucleotides, connected in the center by a series of hydrogen bonds between nitrogenous bases. (b) a schematic illustration showing the arrangement of the two strands in the double-helix configuration. The "backbone" on the outside is the sugar-phosphate chain, and the nitrogenous bases form the bridges across the middle. ©Welcometrust

The only stable base-pairs that can form are $A-T$ and $G-C$, therefore a double stranded DNA sequence is completely described by (either) one of the two base strings corresponding to one strand. The two strands are identifiable only through the structural sugar-phosphate (see Figure 1) backbone, that determines the unique (and opposite) direction along which either one of the strands can be read, when processed in our bodies. As such, the strand that is being processed at some point in time is regarded as the forward strand and its complement as the reverse strand.

1.1.1 Chromosome

A chromosome is a structured subsequence of the DNA. Human genomic chromosomes are duplicate (as opposed to simpler life-forms in which they are unduplicated), more precisely they contain two double stranded DNA sequences called chromatids that are geometrically joined in a region called the centromere. The human DNA contains 23 such chromosomal pairs. Out of these, 22 are called autosomal and the two respective chromatids are homologous, in the sense that they encode information about the same processes. The 23'rd chromosomal pair contains the sex chromosomes. The two chromosomes/chromatids, although paired, are no longer homologous and they determine an individual's sex. Females have an XX pair of chromosomes and males have an XY pair, at this position, where X and Y are the arbitrarily chosen names for these chromosomes. The 22 autosomal chromosome pairs are identified by their ordinal number (1, 2, …, 22) where the order was established according to their observed length (1 is the longest).

1.1.2 Gene

A gene is a structured element consisting of the contiguous DNA sequence necessary to generate a product of functional importance, a protein. Genes lie within chromosomes but vary greatly in size and content[1], ranging from a few hundred base-pair long genes to aprox. 2Mb (mega base-pair) long genes. Some structural patterns can be observed however, as depicted in Figure 2.

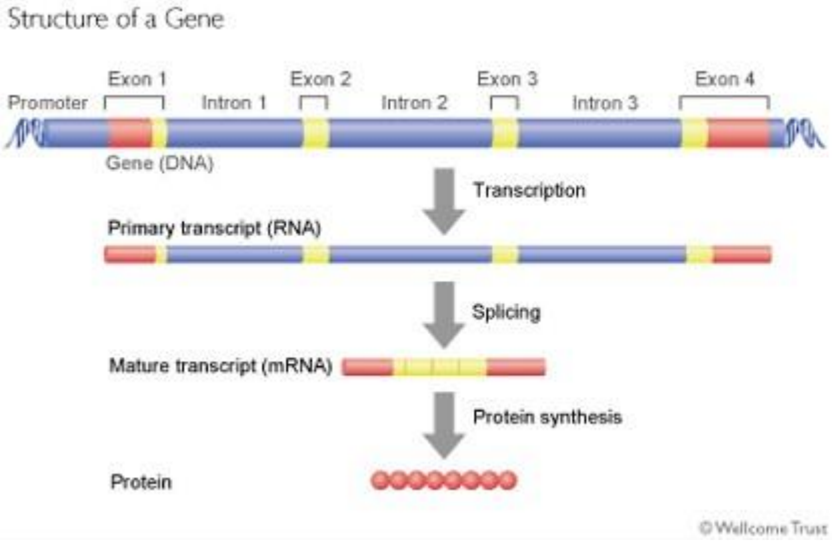

Figure 2: The generic structure of a gene and the process through which the information it encodes is transformed into a final product, a protein. The Promoter is a short specific sequence of DNA that lies before a gene, on the strand, and regulates how often the gene is read and processed. Whenever a gene is processed, a corresponding contiguous sequence of single stranded nucleotide chain is created, called RNA. Introns are contiguous sequences of DNA/RNA, in-between exons, that indicate to the molecules that process the gene where to cut. Exons are the actual "coding" part of a gene, as each 3 consecutive bases in an exon correspond to an amino-acid and the sequence of amino-acids resulting from a gene's exons further defines the protein.

The coding part of the genome represents the set of all exons, as a clear correspondence can be made between the sequence of base-pairs read and the produced outcome, the chemical content of a protein. This amounts to 2% of the size of the genome. The rest of 98% is covered by noncoding DNA whose functionality falls under various degrees of understanding. There are noncoding parts of genes such as introns and promoters. Furthermore there are other structural elements such as pseudogenes[2],[3], which are genes that become inactive because they are no longer translated to proteins, regions that control the expression of actual genes (how often a gene is read so that the corresponding protein is created)[4], highly repetitive regions that serve as structural elements (telomers)[5], etc.

Given that information is encoded twice in organisms with duplicated chromosomes (once on each chromosome of a chromosomal pair), a homolog version of each gene is found on the other chromosome. Each of these two versions is called an allele. If the two corresponding alleles are identical then the individual is said to be homozygous for that allele, otherwise he is said to be heterozygous for the two alleles found. We note that the term allele refers specifically and exclusively to this relation of spatial pairing, on two different chromatids of a homologous DNA sequence. Intuitively defined for genes, an allele can however refer to any DNA sequence length, down to single base-pair positions, and it is used to distinguish between variants at some position.

Depending on the relationship between the two, possibly different alleles at a specific location on a chromosomal pair, the functional effect of one always "dominates" that of the other. Thus, between any two different alleles that can be found at a position, one is said to be dominant and the other recessive (or silent). Furthermore, the number of homologous base-pairs for each position determines the ploidy at that position. All the positions on the 22 autosomal chromosomal pairs are said to be diploid and their genotype is a combination of two alleles. The two sex chromosomes are haploid for males, as the genotype is completely defined by only one allele, whereas still diploid for females as they have two homologous X chromosomes.

1.2 Inheritance

Each organism grows and develops through repeated cell divisions from a single initial cell called zygote. In the case of humans and all other sexually procreating organisms, the zygote is the result of the merger between two special types of cells, one from each of the parents, called gametes (or germline cells, or reproductive cells). The gametes carry the genetic code that is inherited from each respective parent. During fertilization, when the zygote is created the gametes' genetic code is combined, creating the DNA molecule of the child (Figure 3).

The transmission of genes from parent to offspring is governed by the principles of inheritance formulated by Gregor Mendel before such detailed knowledge of the genome existed. His methodical experiments on breeding different pea species have allowed him to formulate the three principles of Mendelian inheritance as well as paving the way for the statistical analyses that are at the core of modern genetics[6],[7],[8]. The first principle is the principle of uniformity, which states that the offspring of two parents that differ in a trait should all have the same appearance. This means that two parents that are homozygous of different alleles, at some position, will pass on to the offspring one of their respective alleles and the children cannot be anything else than heterozygous on those two alleles. Between any two alleles however, a relation of dominance can be determined, as described earlier, thus all children will have the same appearance, the result of the dominant allele (Figure 3).

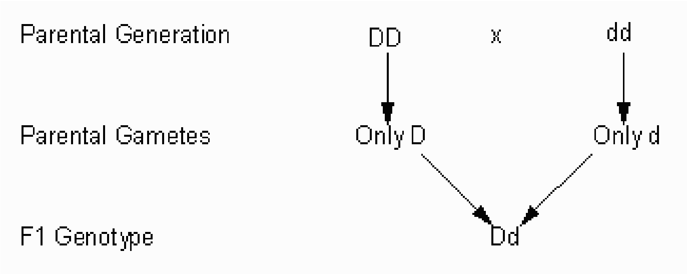

Figure 3: Illustration of inheritance and the first Mendelian principle – uniformity. The Parental generation contains 2 copies (in this case identical) of some gene; a gamete contains only one copy of the parental gene and the offspring (F1 generation) contains the union of the two gametes. Capital letters mark dominant genes, while non-capital letters mark recessive genes, therefore the offspring will manifest the effects that gene "D" produces.

The second Mendellian principle is that of segregation, which describes how pairs of gene variants, or alleles, are separated into reproductive cells. During the division process (meiosis) that creates the gametes, the chromosomal pairs are broken, each gamete cell holding the copy of only one chromosome from each of the 2^{23} 23 chromosomal pairs. This principle states that equal number of gametes are created, that hold either version of a parent's chromosome. In turn this implies that an offspring has equal chances of inheriting either one of the two (possibly different) alleles that a parent holds (see Figure 4).

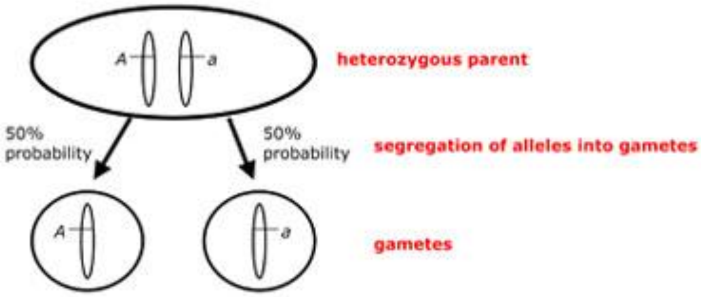

Figure 4: Illustration of the second Mendelian principle – segregation. Whenever gametes are created, they have an equal probability (50%) of holding either one of the two copies of a gene that the parent carries.

The principle of independent assortment finally states that each chromosomal pair of our DNA separates independently during meiosis. This means that alleles at some position segregate into gametes independently of alleles at other positions thus contributing to genetic diversity.

1.3 Evolution

In the study of evolution three constituting factors can be identified: mutation as a generating engine of variation, selection as a regulatory force and genetic drift as a measure of higher level factors that influence a species' evolving genetic code. Selection and genetic drift are also called evolutionary forces, as they exert influence over the genetic variation introduced by mutation processes.

1.3.1 Mutation

Mutation is a phenomenon that may happen naturally down to the single base-pair level of our DNA. Whenever a cell divides, in any living organism, its DNA code must be replicated so that each of the two resulting cells has an identical copy. This process is performed by a set of chemical molecules called polymerase enzymes that must perform a series of tasks sequentially. These molecules must unfold the DNA, each of the double stranded chains must be separated by breaking the base-pair connection then each single strand chain is read and copied by creating new base-pair connections. Since the base-pair connections that can form are unique, each strand of the DNA chain can act as a template for creating its complementary strand and achieving the stable structure. From one double stranded DNA sequence two, ideally identical ones, are thus obtained. These copying mechanisms may produce errors however, when handling the base nucleotides (either when reading the template chain or when adding the complementary bases) and the two resulting double stranded sequences may not be identical to the original one, at a number of positions each of them constituting a mutation. Depending on which step in the replication routine produces the error, different types of mutations may be produced. The three basic structural mutation types are insertion (of one or more nucleotide base-pairs) in the DNA chain, deletions (of one or more base-pairs) and substitutions, each of them generating a number of functional effects depending on the local genetic context in which they happen[9],[10],[11]. Deletions and insertions are collectively called indels. Their aggregation, among other reasons, is also because, as opposed to substitutions they may, and often do have lengths different than one.

1.3.2 Natural Selection

The process of selection, as the name suggests, influences what genetic variation is maintained in a population and transmitted throughout generations through reproduction. Given that mutations happen through bio-chemical processes and there is no intrinsic symbolic meaning behind them, selection then acts as a filter (on a very large timescale). The workings of natural selection is closely related to the notion of fitness, which measures an organism's ability to better adapt and respond to the environment that it lives in, which in turn increases its chances of contributing with more offspring to next generations. In consequence, a fit individual's alleles will be present in more individuals of the next generation thus increasing the respective allele frequencies within the population (positive selection), whereas a less fit individual's alleles will decrease in frequency in subsequent generations (negative/purifying selection). Attempts to generate a distribution of fitness effects that mutations have on organisms[9],[12] have revealed that the majority of them have different degrees of negative effects, some influence fitness measurements so insignificantly that they are regarded as neutral and only a very few result in an increase of fitness for the carrier. In this context, selection acts more as a long-term stability ensuring mechanism of a species' DNA. Furthermore, the stronger the effect an allele produces, the stronger the selection for/against it will be. Given that selection for different alleles is dependent on the fitness of the carrier, which is in turn strongly correlated to the environment that the carrier lives in (from climate conditions to other species present), the criteria that determine an individual's fitness may change over longer periods of time[13],[14]. Alleles that were not previously selected against may in time become subject to purifying selection, while other alleles will increase in frequency, in the population[15]. Variants of natural selection have been identified, such as balancing selection in which two or more alleles are kept at high frequencies in the population, disruptive selection where more alleles are again found at close frequencies but the homozygous genotype for either one results in higher fitness than the heterozygous genotype, etc. A complex example of positive as well as balancing selection is offered by genes related to the immune system. We intuitively expect that these regions would be under strong selective pressure, as their functions directly relate to an individual's healthiness, thus fitness. As different pathogen factors emerge or spread through different populations, alleles that are able to reliably identify them in an organism are positively selected for. On the other hand, different levels of a protein were found to influence our body's response to a pathogen[16],[17]; namely high levels were found to reduce magnitude of the inflammatory symptoms, while low levels of the same protein favour the elimination of pathogens from our body. The DNA sequence that regulates the expression of the gene producing the respective protein contains alleles that are under balancing selection, as both the described effects are essential to our immune system.

1.3.3 Genetic drift

While natural selection filters out alleles from a species' gene pool, according to the effects that they have on the carrier's fitness, the Mendelian principles of inheritance clearly suggest that reproduction is a matter of random sampling (respecting the defined rules) from the parents' genotypes. This process also has an effect on the gene pool of a specific population, which is captured by the notion of genetic drift.

Genetic drift acts as another natural force of regulating genetic variation of a species within a population. More specifically, it models how allele frequencies, at a generic position in our genome, change over generations due to the random sampling in the mating process. The underlying assumption is that the model of random sampling is correct, in the sense that an individual is not able to know/discern what allele (at some site of interrest) his/her partner has and base their mating option on this, a very reasonable assumption in practice. As opposed to natural selection, genetic drift cannot be directly related to criteria such as fitness. We consider a specific generation of a population of size N that is bi-allelic for some gene with allele frequency p for one of the alleles (A) and $1 - p$ for the other allele (B). The allele frequency for allele A within the next generation of our population can then be modelled as randomly sampled from the Binomial distribution $B(p, n)$, where n is the total number of alleles in the population, $n = 2 * N$ considering diploid individuals. The use of the binomial distribution is intuitively justified as follows: by the assumption of random mating (with respect to the genetic code) and equal probability of mating for each individual and furthermore assuming equal number of individuals of either sex and equal distribution of genotypes within each sex, this is reduced to randomly sampling, with replacement of two individuals as parents and build the offspring genotype as a combination of their alleles. Using the principles of Mendellian inheritance, namely the principle of segregation (stating that each genotype allele has equal probability of being transmitted to the child), we can further reduce this process by building the offspring genotype by randomly selecting, with replacement, 2 alleles from all the alleles of the population. Some properties of genetic drift can then be inferred from the properties of the underlying binomial distribution. Hence, the larger the population size, the smaller the variance will be from one generation to the next one thus the longer it will take for one of the alleles to reach 0 frequency and become extinct. Similarly, if the initial population allele frequency is close to 0.5, it will also take a larger number of generations for genetic drift alone to eliminate either one of the two alleles. It has been shown however, that even in the absence of natural selection (neutral selection), one allele is expected to reach fixation (allele frequency 1) in the population and expectations of this time, in number of generations, are given. For initial allele frequencies of 0.5 the time-to-fixation for one of the alleles was found to be $2.77 * N$ generations, where N is the population size[18]. This is a very large time-scale considering Earth's population, which would indicate that genetic drift's marginal effect is quite small. However, the whole population of Earth does not satisfy the assumptions of the model and must thus be applied to subgroups of populations and it becomes very important as effective population size gets smaller. An important contribution of genetic drift can be observed in populations where the gene pool is the result of founder effect[19],[20]. In these situations a small group, such as a religious group or colonists, of a larger population gets isolated and is only able to mate within the group. Allele frequencies within the group may then differ significantly from those within the original population and, through genetic drift, initially rare (with respect to the whole population) and sometimes deleterious alleles will increase in frequency or fix within the group. This was found to increase the incidence of rare disorders; for example the Bardet-Biedl syndrome has a prevalence of 1 in 17500 in Newfoundland population, a genetically isolated population, which is one order of magnitude higher than the incidence in the more admixed populations of Northern Europe.

1.4 Importance of de-*novo* mutations

Discriminating on when in an organism's lifespan and/or which type of cell a mutation occurs in, we distinguish between two different classes, namely germline and somatic mutations. Any mutations that arise in the reproductive cells or the produced gametes of an individual's parents are called germline mutations and they are present in every cell of the offspring, as he develops. Mutations that happen in subsequent offspring cell divisions are called somatic and they are found in various degrees of spread through the organism (from specific tissues to organs, etc.) depending on when and where in the individual's development they happen. Somatic mutations may influence an organism's fitness substantially, such as mutations related to cancer[21]. Germline mutations however, are the only candidates for transmission to further offspring throughout generations.

For the purpose of this project we focus on germline de-*novo* mutations (DNMs), in particular to their accurate discovery. We mainly treat single nucleotide variant (SNV) DNMs, although the extension to de-*novo* indels is quite straightforward. Single nucleotide DNMs are, in this context, mutations that form in the DNA replication process that is required when the gametes of both parents are created. The haploid

DNA molecules that come from each of the parents are joined to create the diploid DNA molecule of the offspring. A DNM can then be observed by looking at the genotypes of a mother-father-child trio at a position of interest. If the observed trio genotypes combination violates the Mendellian principles of inheritance then a mutation (or two) are present in the child. Given two symbolic alleles $A\ and\ B$ at a site in our genome, a straightforward example of a mother-father-child genotype combination containing a Mendellian violation is $AA-\ AA-\ AB$ and a combination resulting in two Mendellian violations is $AA-AA-BB$, as neither of the child's alleles is found in the parents. Two de-*novo* events at the same site in an offspring are extremely unlikely however, as we will see. We note that the combination $AA-BB-BB$ also contains a Mendellian violation. One of the child's B alleles is inherited from the father however, the other allele must come from the mother and she does not have any B alleles.

1.4.1 Mutation rates and distributions

All genetic variation within our genome arose as a DNM (at least once) that was then transmitted to further generations. It is therefore essential to our understanding of evolution as well as disease heritability, to understand the rates at which such mutations happen and certain biases that are involved. Rates of mutation were derived through two types of methods. Initial estimates were obtained through an indirect approach that would extrapolate a mutation rate by looking at known differences in the genome of two relatively close species (such as humans and chimpanzees[22]) for which knowledge about the time of separation from a common ancestor exists. Subsequent studies followed a more direct approach of finding DNMs in known genes thus obtaining local estimates from the human genome directly that were extrapolated to the whole genome[23],[24]. With technology making it possible to sequence the whole human genome, more recent studies have obtained these genome wide mutation rate estimates directly, by sequencing the whole genome of one, or many, trio families. The derived mutation rates seem to converge around the value $1.17 * 10^{-8}$ mutations per base-pair per generation[23].

Comparisons of the germline mutation rates between males and females revealed that the male germline is more susceptible to mutation, the main hypothesis being that the male germline cells undergo more divisions during an individual's life-time[25], thus offering more chances for mutation. As a direct consequence, it was found that the larger proportion of DNMs in an offspring come from the father, although the actual proportion is subject to large variance. Furthermore, a clear correlation between father's age at conception and number of de-*novo* mutations in offspring was observed however, the exact mathematical relation/magnitude is still subject to debate/improvement[23],[10].

Further profiling of DNM events have also yielded a non-random spatial distribution along the human genome. Thus, regions of the genome that are dense in C and/or G bases, GC – rich regions, show a higher mutation rate due to chemical instability, highly repetitive regions are also more error-prone, etc.. An enrichment of DNMs was also noted in coding regions partly explained by their high GC content, but also by other factors such as transcription associated mutations.

As genetic variation in our genome is continuously indexed, as a result of large projects that sequence groups of up to thousands of people correlations between variants/alleles is found; namely, it can be observed that the presence of an allele A at some position in our genome is correlated (to various degrees) with the presence of some other allele B at another position. These observations are quantified by the notion of linkage disequilibrium (LD). By making use of such patterns of LD around a respective position, it can be inferred whether an observed DNM originates from the paternal or the maternal germline.

1.4.2 Role of de-*novo* mutations in disease

Beyond their relevance to our understanding of evolution, DNMs were also found as good candidates for disease-causing variation[26],[27]. When attempting to explain the heritability of a disease, two general models are typically employed.

The "Mendellian disorder" model assumes that a disease's genetic causes are monogenic, in that they lie within the coding regions/exons of one gene, simple, in that the exonic variation produces visible effects in the generated protein and the ramifications of this can be further studied up to (ideally) the phenotype level, and rare, in the sense that variation in such highly functional regions is usually very deleterious thus these causal alleles will not typically propagate in the population. Also Mendellian diseases show very extreme phenotypes, often including various degrees of mental retardation and/or physical malformations. DNMs have been found to contribute in explaining the heritability of this class of disease with relevant examples such as Schinzel-Giedion syndrome[28], Bohring-Opitz syndrome[29], Kabuki syndrome[30] or KBG syndrome[31].

The "complex trait" model, by contrast, assumes that a disease's heritability is polygenic, in that causal variation is spread heterogeneously across the genome, complex, as causal variation may be found in regions for which function is known even less than for coding regions, and common, in that disease causing variation may segregate at higher frequencies within the population. Typically, many loci spread across the genome (up to hundreds of thousands[32]) are found to have small contributions through their causal alleles to the overall heritability of the disease. The same arguments make this model suitable for the study of some quantifiable human traits in general, such as height, and not only diseases. An underlying assumption of this model is the "common disease, common variant" assumption which would make DNMs counter-intuitive candidates for such analyses, as a DNM is unlikely (although not impossible) to result in an allele that already has high frequency in the population. DNMs have been found to contribute to the heritability of some common disease though[33], sometimes with larger effects than the other, common variants involved. They are sometimes considered a "trigger" variation for other underlying common variants[34],[35].

2. Materials and Methods

2.1 Sequencing

Sequencing is the complex and sequential process by which the DNA molecule of an organism is represented as a contiguous sequence of its nucleobases. Initial Sanger sequencing techniques, also termed “first generation” sequencing, date back to the 1970s[36],[37]. Similar to modern techniques, they are based on cutting the contiguous DNA molecule of 3 billion nucleotides into smaller manageable subsequences, called reads. These are individually sequenced by mimicking the DNA replication process that happens naturally each time a cell divides. Instead of adding normal complementary nucleotides to the template strand however, slightly modified nucleotides are added that, through chemical reactions, emit fluorescent light of different frequencies. Identification is then performed through simple spectrum discrimination of electrophoretic measurements between bases thus enabling their read-out, as well as a level of confidence for each outputted base. The method has been continuously researched and improved, reaching a very high empirical performance $> 99.99\%$[38],[39]. The use of this technology is very expensive and therefore rather impractical for genome-wide sequencing of individuals, or cohorts of individuals, which is desired for building statistical power for association studies.

2.2 Next Generation Sequencing

Improvements in most chemical and technical steps of the sequencing process, as well as major increase in post-sequencing processing power, methods and data storage have resulted in a number of related technologies collectively termed “Next Generation Sequencing” methods. These methods produce very large amounts of data, at a high throughput, and at significantly reduced costs. Instead of directly producing a sequence, they output a variable number of reads that cover a desired position, each with a corresponding confidence measure, and the actual sequence is derived by consensus in subsequent analysis.

All NGS technologies follow the basic workflow depicted in Figure 5. This involves cutting the DNA molecule into manageable size fragments, attaching elements that allow us to manipulate individual fragments, one or more “amplification” steps where we multiply the available DNA and the base reading process. Different NGS platforms implement these steps slightly differently[39],[40]. This creates a spectrum of technologies, each with specific advantages and limitations. Given different error modes for each NGS platform, the challenge remains to robustly integrate their sequencing output, for downstream analyses. We present the NGS workflow integrated in our analysis, as performed on an Illuminae platform, currently the most used technology.

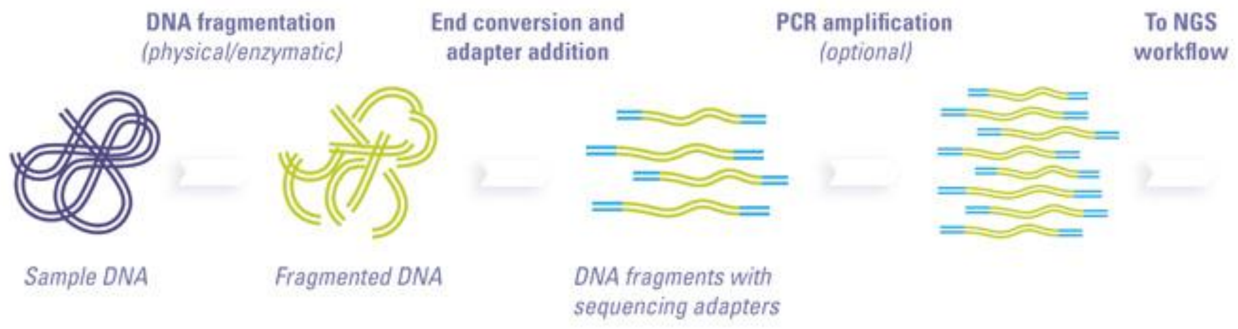

Figure 5: Overview of a Next Generation Sequencing process. An initial DNA molecule is first cut into short sequences, called reads, by the use of primers, in a special fluid. Sequencing adapters, also primers, are then added to the solution and they couple to both ends of each read. The PCR amplification step makes many copies of each read until a certain read concentration is obtained in the fluid. The reads can then be "read out" by the sequencer.

The sequencing process starts with library and template preparation. The DNA molecule to be sequenced is cut into small size reads. The Illumina platforms we used currently support a maximum reliable read length of ~ 150 bases. This is much smaller than the reads supported by Sanger sequencing methods, because of the slightly different electrophoretic calling methodology, but is also a source for the significantly higher throughput rate.

2.2.1 DNA Amplification

The emulsion PCR (Polymerase Chain Reaction) process is then performed on the joint set of reads. ePCR is a cyclic process that repeatedly makes "copies" of the fed set of reads generating exponentially increasing amounts of data. This step is generally performed in order to get a sufficient amount of DNA to sequence. Neither amplification nor the subsequent sequencing are deterministic. As a certain concentration of DNA reads is achieved and a number of other needed molecules are added the reaction (i.e.: PCR and/or sequencing) starts. The DNA of interest is put in a solution along with DNA primers, nucleotides, and polymerase. Within each cycle, the DNA reads are heated to a temperature of up to 98 degrees. This breaks the hydrogen bonds between base-pairs, effectively separating the two DNA strands. The single stranded reads are then bordered by small, typically $15 - 30$ base long, a-priori designed and synthetically created DNA sequences called primers. Additional primers are freely available in the solution/gel in which the reaction takes place. When the temperature is brought down, the single stranded primers attached to the reads form bonds with the complementary primer sequence available in the solution and mutant DNA polymerases (mimicking the structure and function of molecules in our cells) then bind to them and use the single stranded DNA reads that they are coupled to, to create the complementary strand (Figure 6). Primers are needed whenever we wish to artificially build a double stranded DNA sequence from a template because polymerases can only extend an existing DNA chain and not start one from zero.

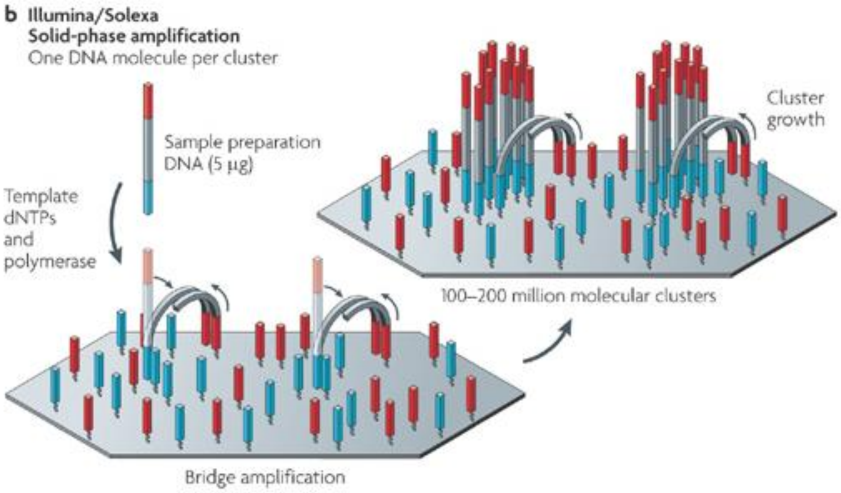

Figure 6: One solid-phase PCR amplification step, as performed on an Illumina machine. The blue and red sequences are primers. As a read is attached, the free primer at one end forms hydrogen bonds with its complement, freely available on the board. An enzyme can then attach to this end and use the single stranded read as template for creating its complementary strand thus obtaining a double stranded read. Hydrogen bonds break again through heating up and the process is repeated several times until (ideally) identical clusters of the same read are created.

The PCR cycles determine an exponential increase in available reads and are repeated until the expected value of reads covering any genomic position reaches the desired average coverage. During each ePCR cycle, there is an equal probability that any read chosen for amplification comes from either one of the two chromosomes, covering those positions, of a diploid organism. This then leads to a binomial distribution of reads (w.r.t. originating chromosome) at each individual position in the sequenced genome. This implies that, on average, the two alleles present at any site are equally represented in our sequencing data, enabling us to make correct calls of the underlying genotype. However, for each site there is a probability, proportional to the coverage at that site, that one or both of the alleles are under-represented or not captured at all. The impact of this distribution is greater for heterozygous positions since they require adequate coverage of both alleles.

2.2.2 DNA Sequencing

With a library of reads built as described above, the actual sequencing process is applied. One sequencing run as a whole is similar to one individual cycle of the amplification (see Figure 7). The building of the complementary strand however, is in turn cyclic so as to allow read-outs of the basses. The big dataset of reads is organized into clusters and more runs of the sequencing machine can be performed as needed. After the reads have been separated into single stranded DNA chains and primers were added, the sequencing cycles begin. During each cycle, dye-augmented (i.e.: each type of base has a distinct colour) single basses are added to sequencing solution and polymerases add the corresponding base to each read, the base complementary to the next position of the template single stranded read. After each adding step, all the “unassigned” dye-augmented bases are taken out of the substance and light is emitted in the gel by two lasers. Total reflection inside the gel creates a stationary wave that is then measured and the just-added bases are predicted with a corresponding level of confidence/probability of success. The dye on the just-read bases is eliminated and the cycle is repeated until the reads are completely built and read. Variants of this process exist where, for example, during one cycle only one type of base can be added to each/all reads ($A, C, G\ or\ T$). The such obtained reads can then be used to reconstruct the sequenced individual's contiguous DNA string, during alignment.

Alternatively, the method of cutting the initial DNA molecule into reads can be modified to produce paired-end reads, as opposed to single, independent ones. Paired reads are obtained by designing primers that cut the DNA in such a way that one cut produces two reads, for which we know that they are separated by a fixed and relatively constant nucleobase distance. For example a paired end read can be two 70-base reads, spaced such that there are aprox. 250 basses between them. These 250 "hidden" basses are not captured, neither amplified nor read. Having paired reads can improve downstream alignment for certain parts of the genome[41]. If one of the two ends falls within a region where alignment is hard (such as repetitive regions, indels, etc.) and its pair does not, we can use the confident alignment of one, to infer the other.

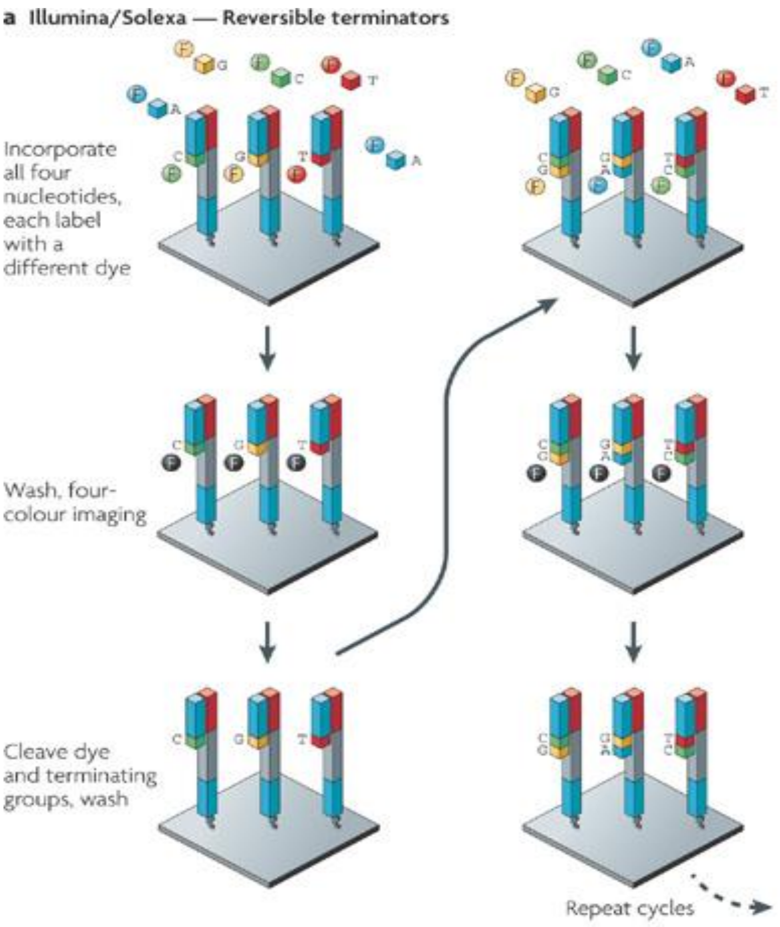
a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
A
C
G
T

Figure 7 : Two consecutive cycles of a sequencing process using reversible terminators, as performed on an Illumina machine. The reads to be "read out" are primed at both ends and an enzyme attaches; then free floating nucleobases of all 4 types are added to the solution, each labeled with a different dye color (up). The enzyme couples the right base to each read, then the remaining free bases are removed and the just added bases are read (middle). Finally, the dye on the added basses is removed and the cycle can be repeated (bottom).

The massive read output of a sequencing machine is stored in the .fastq format. Each read is stored as the ASCII sequence of bases that were called. Furthermore, a homolog sequence of quality scores corresponding to each base in the read is stored, represented by ASCII characters, that correspond to integer values. The quality representation used is Phred[42] scaled likelihoods of the probabilities $P(b),\ where\ b = called\ base$ outputted by the sequencing machine:

$$Q = -10 \log_{10} P(b' \neq b), \qquad where\ b' = true\ underlying\ base$$

A Phred quality of 10 then corresponds to a probability of 10% that the base-call produced by the sequencer is in fact wrong, etc. Fastq is the standard format for the output of most sequencing technologies and it offers a singular data representation format for subsequent analyses, although the quality values from different sequencers cannot be fully integrated yet.

2.3 Genome Analysis ToolKit – GATK

The Genome Analysis ToolKit (GATK) is a java based framework developed to enable and ease the manipulation and processing of genetic data. Developed and maintained by a group within the Broad Institute[43], the framework is continuously extended currently offering the means to easily run entire analysis pipelines, from raw sequencer output to final results. The framework as well as most of its tools is open-source and external contributors can participate in extending it.

One of the essential features GATK provides is its data access patterns, allowing developers to focus on processing of the information. Given the massive sequencing and/or alignment data (about 200Gb for an individual 12x coverage genome), it becomes clear that processing has to be done without loading an entire file into memory. The other defining design element is the map-reduce processing framework. Given that input is loaded sequentially, the processing must in turn be modular. Namely the processing is applied to each symbolic element that the traversal engine loads and all intermediate results are pooled together to compute the desired output

Two main types of formats are used to store genetic data, both relying on annotations. The .bam format is used to store raw alignment data, as described above and it contains genomic position wise, base-wise and/or read-wise annotations. The .vcf format is used to store variants computed from the alignment and containing higher end, variant annotations.

2.3.1 Walkers

A walker is a means of traversing genetic information by "walking" along one of its dimensions. For each of the positions along this dimension, the walker will load all relevant information from the data "tracks" passed to the engine. Such tracks can comprise of sequencing data (e.g. bam files), a reference sequence, a list of variants, etc. The GATK provides two basic types of walkers:

- ReadWalker traverses files by loading into memory each read plus all contextual information available for it. Such information includes the base sequence representing it along with all the base quality estimates, reference position(s) it aligns to, mapping quality, etc. the "map" function is called for each such loaded read where the developer can define the desired processing and compute a local result. The "reduce" method is subsequently called for the same read, with the result produced by the map method, where the developer can define how to integrate the local result in the final output of the walker.
- LocusWalker traverses the whole reference genome position by position. For each genomic position it loads data such as reference base at that position, reference bases in an adjustable window around the current position, all reads containing bases aligned at this position along all their read-specific data, etc.

Depending on the processing being done, additional files can be passed such as files containing information about the individuals being processed (sex, family structure, case/control status, etc.). GATK automatically incorporates and correlates this information and makes it available in the same map-reduce paradigm (i.e.: assigns sex of individual automatically), or makes it globally available.

2.4 Alignment

Using Next Generation Sequencing technologies we can sequence reads that collectively cover the whole human (or other organism) genome, but during this process we lose all information regarding where in the genome they originate from. Reconstructing the DNA molecule that was sequenced can be done by two approaches. One is *de novo* assembly, that attempts to reconstruct the complete contiguous DNA sequence based exclusively on the reads obtained, typically by exploiting overlaps between reads covering close regions. For humans, as well as for a few other organisms (i.e.: mouse, zebra-fish), a reference sequence for the DNA molecule was pre-built and the task is reduced to aligning the reads to this reference sequence.

The human reference sequence was developed and is continuously maintained by the Genome Reference Consortium. The reference sequence was produced by a more complex methodology that combines de-*novo* assembly of approximately 13 individuals with comparative genetics, namely studying the content and resemblance of closely related species that are assumed to have separated relatively recently, on an evolutionary timescale.

Many efficient algorithms have been developed that can tackle the task of aligning up to millions of $\sim 100bp$ reads to a $3 * 10^9$ reference sequence. Some algorithms are designed to align reads from more NGS platforms, such as Mosaik[44], while others are tailored to perform specifically well for specific technologies. As the most widely used platform, many algorithms exist for aligning Illumina sequenced reads, such as MAQ[45], SOAP[46], ELAND, Bowtie[47], BWA, etc. and approaches vary from hash-table based to tree-based methods. In our analysis pipeline we use BWA[48] which relies on the Burrow-Weeler method of indexing a string, in our case the reference genome, and performing a tree-search for each read to be aligned. The algorithm is further optimised for specific data patterns in the reads/reference data. For example, a higher number of mismatches can be allowed around sites where indels are known to be present, so that they are captured correctly, and the use of paired-end reads can be employed to increase performance in highly repetitive regions. A phred-scaled quality measure is produced for each read as the likelihood of the respective read being miss-aligned.

All information computed up until this point is stored in Sequence Alignment/Map (SAM) format files or their binary version, i.e.: BAM files. They encapsulate read level information such as the sample (/individual) it originates from, the group of reads it was sequenced together with, the position in the reference sequenced where it was aligned, mapping quality, etc., as well as the original base call quality for each individual base.

2.4.1 Base Quality Score Recalibration

Due to discrepancies observed between the base quality assignments at some of the intermediate steps and empirical error rates, recalibration of these scores needs to be performed. An empirical observation is that alignment around known indels is of lower quality. Failure to correctly align a deletion for example, by inserting an appropriate number of gaps in the aligned read, results in a higher number of mismatches (thus lower quality), but also in false evidence of variation at the respective site. Realignment is typically performed for sites around genomic positions where indels are known to occur in the population

Another important calibration step is the Base Quality Score Recalibration (BQSR)[43]. Given the aligned reads, we can empirically approximate a sequencer error rate by contrasting alignment matches and mismatches. A mismatch may originate from a sequencer error or it may be indicative of a variant allele present at that site. Knowing that current genetic databases include most of the variants present in any human genome, other sites where mismatches occur can be considered highly unlikely to have true underlying variants. Empirical error rates are estimated for initially assigned quality values with respect to known covariates for each base:

- Read group it belongs to – reads sequenced together show common error models due to shared chemical environment
- Assigned quality score – recalibration notwithstanding, a strong correlation is obvious between assigned and empirical quality scores
- Machine cycle producing the base – different runs of the sequencing machine show slightly different error models
- Dinucleotide (current base + previous base) – different chemistries between two adjacent bases generate different error probabilities

Plotting such computed error estimations against the sequencer reported quality scores, we observe systematic differences that should be corrected for (Figure 8a). Furthermore, observing the histogram of reported quality scores (Figure 9a), we note that there is little discrimination power between high quality bases. As it turns out from downstream analyses, the class of high quality basses is the most relevant in subsequent investigations, therefore, higher discrimination power on these naturally results in higher sensitivity for the respective analyses.

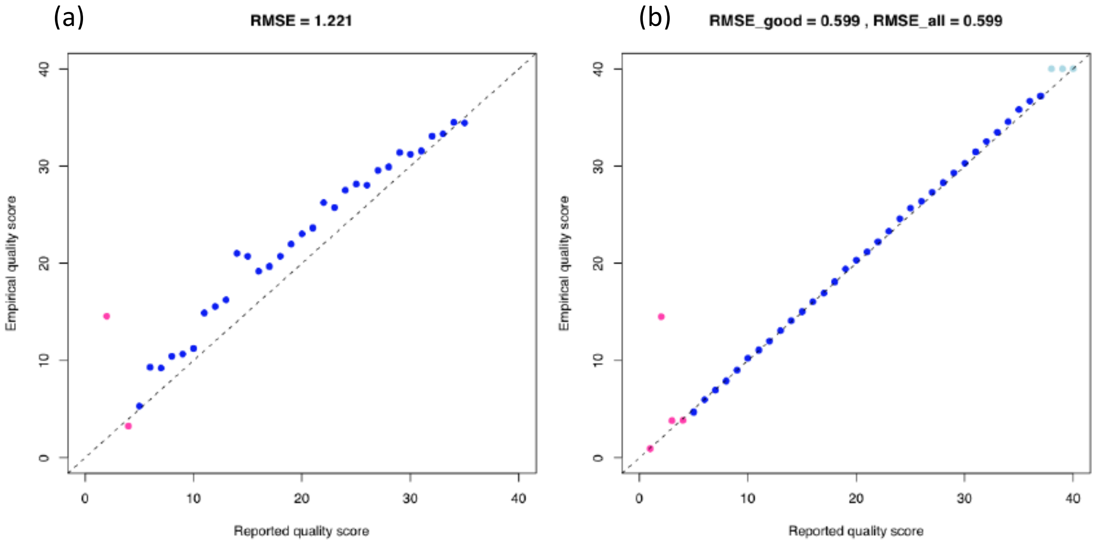

Figure 8 : Plots Empirically estimated error rates vs. sequencing machine reported error scores of base quality scores. (left): before BQSR; (right): after BQSR

The recalibration is then performed by estimating a model that best describes the empirically estimated error rates.

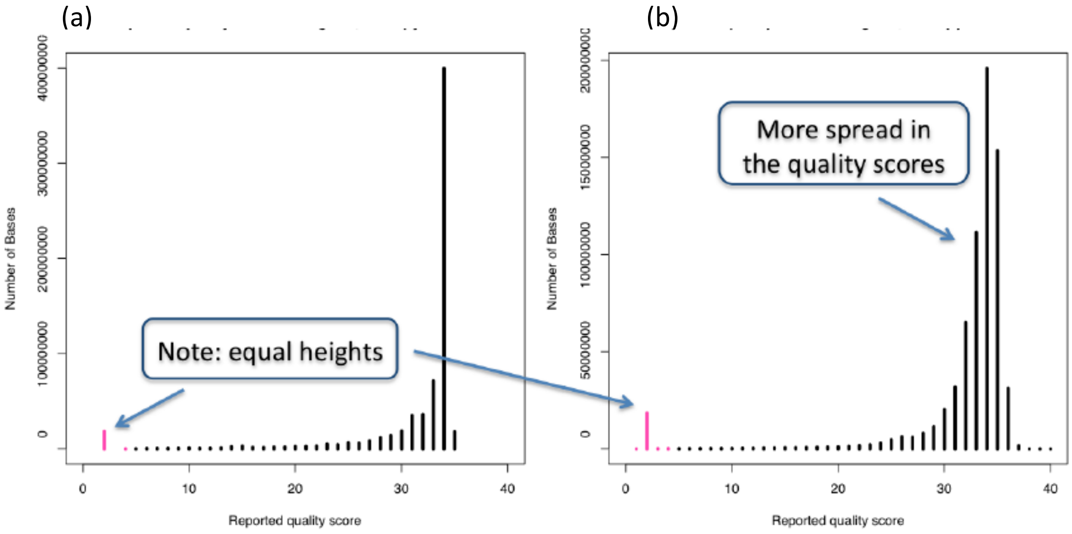

Figure 9 : Plots of the distribution of base quality scores. (a): before BQSR; (b): after BQSR.

The BaseRecalibrator locus walker is used to traverse the bam file that we wish to recalibrate; it computes the described features for all bases at each locus and summarizes the results in a report file. A second locus walker, PrintReads, is subsequently used on the same (or different) input bam file. This walker reads the report, computes the correction coefficients needed to recalibrate the base qualities on the fly and updates all quality values for all bases. Bases qualities are effectively recalibrated by using coefficients that estimate:

- The global difference between reported quality scores and empirical ones
- A quality bin specific shift
- A machine cycle and dinucleotide effect specific shift

After recalibration, the quality scores are much closer to the empirical observed values, as depicted in Figure 8b. Also, we notice much higher discrimination power, in the form of a higher spread of basses across the high quality bins of the spectrum (Figure 9b). BQSR thus allows for a statistically more robust usage of the alignment data in subsequent analyses.

2.5 Variant calling

All heritability studies attempt to explain genetic influence on observable phenotypes by investigating systematic genetic variation. Being able to confidently identify, or call, the alleles present at any site of interest in the genome is therefore essential.

In our analysis pipeline we use UnifiedGenotyper (UG), a LocusWalker that processes each genomic location and outputs the most likely genotype at that position, along with likelihoods for all other genotypes considered possible as well as other qualitative information relevant for subsequent analysis.

The UnifiedGenotyper is able to make full use of the available information when a group of individuals, in our case all the GoNL dataset, is being genotyped simultaneously at some position of interest. The UG then considers the alignment data for all found individuals jointly and derives the alleles present at that site, along with their respective allele frequencies, by using an Expectation-Maximisation (EM) algorithm[49]. The EM algorithm maximizes the joint likelihood of a set of parameters and a set of hidden, unobserved, data values, given a limited set of observed values. In our context, the observed values are the alignment data for each individual at the site of interest, the hidden values are the number of reference alleles of each individual (i.e.: the number of alternative alleles is straightforward given known ploidy) and the set of parameters of interest is the allele frequency spectrum at the respective site which in turn defines the set of present alleles.

The algorithm first computes the genotype for each individual in a maximum likelihood manner given the available alignment data:

$$P(G|D) = P(D|G) * P(G)$$

where G is the genotype and D is the available data, i.e.: the set of basses aligned to this position. The genotypes G considered by the likelihood computation are all the possible genotypes consisting of alleles identified by the EM algorithm. Each of these genotypes is considered equally likely, thus the $P(G)$ prior is assigned the likelihood value of 1. Evidence provided by each aligned read is considered to independently contribute, thus the likelihood function becomes:

$$P(D|G) = \prod_b P(b|A_1A_2)$$

where b is the base aligned to the current position, from each overlapping read and A_1A_2 is the allele representation of the underlying diploid genotype. The likelihood function can then be further expanded as:

$$\prod_b P(b|A_1A_2) = \prod_b P(b|A_1)\frac{1}{2} + P\,(b|A_2)\frac{1}{2}$$

where

$$P(b|A) = \begin{cases} 1-e, & if\ b = A \\ \frac{e}{3}, & if\ b \neq A \end{cases},$$

$$where\ A\ is\ a\ possible\ allele\ and\ e\ is\ the\ per-base\ error\ probability$$

The per-base error probability is derived from the recalibrated phred scaled quality score stored in the input alignment (.bam) file. The likelihood for each possible allelic combination at a site is computed and, following a normalization step, posteriors for each genotype are derived. The selected genotype is thus outputted, along with phred scaled likelihoods for all other possible genotypes. Considering a biallelic site the three possible genotypes are HOM_REF (homozygous on the reference allele), HET (heterozygous) or HOM_VAR (homozygous on the alternative allele). For a genotype called as HET an example of phred-scaled likelihoods (PLs) is 20,0,50. These likelihoods indicate that HET is the called genotype (the most likely genotype's PL value is forced to 0) and that there is a probability of 10^{-2} that the genotype is actually a HOM_REF and a probability of 10^{-5} that it is actually a HOM_VAR. The lowest of the PLs corresponding to not-called possible genotypes is denoted as the quality of the variant and filtering is typically performed on this value to select variants above a desired confidence threshold. The genotyper might conclude that either one, or both alleles of an individual cannot be confidently called, because of missing or highly inconsistent alignment data. The outputted genotype can therefore contain one or both alleles set to NO_CALL. If at least one allele is called as a NO_CALL, no PLs, or variant quality, can be produced.

The UG supports calling individuals assuming haploid genotype as well (as it is the case for the X chromosome in males), but the ploidy must be invariable across all samples that are called simultaneously. Since our dataset contains both males and females, the X chromosome is initially called as diploid (for all individuals) and adjusted further on in the pipeline.

Generated output is stored in a Variant Calling Format file (.vcf), which allows for standardized yet versatile generic representation of genetic variants. Along with position (in the form of chromosome and chromosomal-index), called genotype and PLs, other annotations can be outputted as needed for subsequent analyses. Typically subsequent filters are applied to the raw variant calls and information as to whether each variant passed is added. Some commonly used variant annotations, that are also useful for our analysis, are:

- Reference allele – the allele from the human reference genome, at this site
- Alternative allele – non-reference allele(s) found in at least one individual in the samples being called jointly
- Phred-scaled likelihoods (PLs) – as described above
- Genotype quality (GQ) – as described above
- Depth of coverage (DP) – total number of reads that overlap this position (for each individual)
- Allele depth (AD) – number of reads overlapping the position, that contain one/each specific allele

One line in a .vcf file corresponds to one genomic position. Each individual present in the analysis dataset represents a column and information describing each individual's variant at that position can be found on each line, allowing for missing data.

For practical reasons (i.e.: the 10^9 input order of magnitude) sites where all individuals are called as homozygous reference are typically not included in the output. For sites where at least one individual was found to have a non-reference allele (i.e.: variant sites), genotypes are produced for all individuals that are called together.

2.5.1 Variant Quality Score Recalibration

The choice for the variant quality threshold made in calling determines a trade-off between sensitivity and specificity. Typically we want to produce a highly sensitive set of variants from the available alignment data, that we then filter. Variant Quality Score Recalibrator (VQSR)[43] builds a model to increase discrimination power between true underlying variants and sequencing artefacts, using features that are not directly considered when the initial call is made. The features used can be computed after the initial variant calls were made and added to the VCF file as annotations. Their description is beyond the purpose of this project. VQSR builds a Gaussian Mixture model that estimates the covariance between the computed features and the probability that the called variant is a true variant. A set of variants known to be true must be supplied for training, typically from public databases. The variants from our dataset are then projected on to the model and a confidence threshold is selected to separate the data into low and high confidence. Typically only high confidence sites will be considered for subsequent, downstream, analyses.

2.6 Genome of The Netherlands

The Genome of the Netherlands (GoNL) is a large sequencing project, in which 250 parent-offspring families across the Netherlands were sequenced. Through this representative sample of families, the project aimed at characterizing the genetic structure of people living in the Netherlands. The intermediate average depth of coverage used, $> 12x$, allowed for robust systematic detection of > 20.4 million single nucleotide variants (SNPs), as well as other types of genetic variation: short insertions and/or deletions, structural variants, etc. The large dataset, coupled with good sequencing coverage allowed for robust detection of variants that are rare within the population, i.e.: occur in less than 5% of the individuals, along with common genetic variation. This in turn enabled the fine-scaled characterisation of genetic structure across the country, supporting a demographic model, migrations, admixture with neighbouring populations, etc..

The familial structure of the dataset makes it an ideal candidate for the development and testing of computational methods for identifying and, further on, characterizing de-*novo* events. The 250 GoNL families have the following familial structure:

- 231 single offspring trios
- 8 dizygotic twin quartets
- 11 monozygotic twin quartets

2.6.1 The X chromosome

Applying the above described variant calling pipeline resulted in a VCF file containing 834,651 sites genotyped for each of the 767 individuals. The size of the reference sequence against which chromosome X reads are aligned is 155,270,560 (~155Mbp).

Given that the PCR amplification step produces reads originating from the sequence of a diploid chromosome following a binomial distribution, with an average coverage of 12x per autosomal, diploid, site, we expect a lower average coverage of 6x in males, as the X chromosome is not paired for them. Secondly, it has been shown that regions towards both ends of the X chromosome do in fact pair, in an autosomal/homologous manner with corresponding regions at both ends of the Y chromosome[50],[51]. These two regions are called pseudo-autosomal and the X (and/or Y) chromosome is diploid for men as well, within their boundaries. They are thought to have arisen through local recombination of the two different sex chromosomes and to serve a function in meiosis.

The alignment data, on which the genotype calling and any subsequent analysis is done, within the pseudo-autosomal regions is completely unreliable. Ideally, one reference sequence corresponds to each of these pseudo-autosomal regions (as for any other autosomal region) and PCR resulting amplified reads, from the corresponding X and Y regions, are aligned against it. For practical reasons however, X and Y chromosomes are treated independently as a whole and aligned against their own respective reference sequence. We thus have no control over where each of the reads covering these regions ends up, therefore producing an unpredictable and unreliable distribution and subsequent calling.

The two pseudo-autosomal regions span the two ends of the chromosome, below base-pair 2,699,520 and above base-pair 154,931,044 respectively. They are therefore roughly 2.6Mbp long and 340Kbp long respectively, cumulatively accounting for less than 2% of the X chromosome.

2.7 PhaseByTransmission

Because NGS technologies produce error rates much higher than the estimated probability of a de-*novo* mutation (DNM) event, accurate calibration of evidence supporting a DNM must be contrasted with evidence supporting Mendelian inheritance of a child's alleles. A miscalled genotype in either parents or in the offspring leads to a false negative or false positive DNM call.

Currently used variant callers such as UG (see above) make use of available individual sequencing data as well as, optionally, multiple individuals sequencing data, to estimate genotype likelihoods for each possible genotype at some specific site. They incorporate and minimize the uncertainty from NGS alignment data. This leaves much room for further improvement in the case of DNMs however, as the UG produces a number of 4.5M Mendelian violations within the 269 offspring of the GoNL dataset. Considering an expected number of mutations of 63.2 per individual from previous estimation endeavours, the expected number of Mendelian violations (MVs) is then much smaller, $\sim 16{,}300$. To this end we implemented the PhaseByTransmission.

2.7.1 Pedigree information

The UG computed likelihoods (i.e.: $(D|G)$) are independent per sample and no prior is used when deriving the posterior probability of a genotype. By incorporating transmission priors, derived from available information about the familial relationship between samples we are able to aggregate the available evidence in a more statistically robust manner.

As opposed to computing an individual's genotype in a likelihood based manner, we compute a trio's genotype combination (i.e.: mother – father – child). Considering the bi-allelic case that PBT currently supports, there are 27 possible trio genotype combinations. Out of these, 15 are consistent with Mendelian inheritance patterns, 10 imply one Mendelian violation (i.e.: one DNM), and 2 imply 2 Mendelian violations (see Table 1). By taking into account each individual's sequencing data through UG computed genotype likelihoods, a prior de-*novo* mutation rate and allele frequency in the population, the most likely trio genotype is computed, in a likelihood based manner.

	mother	father	child	#DNMs	transmission prior
1	AA	AA	AA	0	1
2	AA	AA	AB	1	1,00E-07
3	AA	AA	BB	2	1,00E-14
4	AA	AB	AA	0	0.75
5	AA	AB	AB	0	0.75
6	AA	AB	BB	1	1,00E-07
7	AA	BB	AA	1	1,00E-07
8	AA	BB	AB	0	1
9	AA	BB	BB	1	1,00E-07
10	AB	AA	AA	0	0.75
11	AB	AA	AB	0	0.75
12	AB	AA	BB	1	1,00E-07
13	AB	AB	AA	0	0.25
14	AB	AB	AB	0	0.5
15	AB	AB	BB	0	0.25
16	AB	BB	AA	1	1,00E-07
17	AB	BB	AB	0	0.75
18	AB	BB	BB	0	0.75
19	BB	AA	AA	1	1,00E-07
20	BB	AA	AB	0	1
21	BB	AA	BB	1	1,00E-07
22	BB	AB	AA	1	1,00E-07
23	BB	AB	AB	0	0.75
24	BB	AB	BB	0	0.75
25	BB	BB	AA	2	1,00E-14
26	BB	BB	AB	1	1,00E-07
27	BB	BB	BB	0	1

Table 1: All the 27 possible autosomal trio-genotype combinations. Contains the symbolic genotype of the three individuals (where A and B denote possible alleles), the number of de-novo mutations that the combination contains and the corresponding transmission prior for each combination (where 10^{-8} is the per base probability of a DNM as estimated in literature)

2.7.2 The Model

In order to compute the posteriors for trio genotype combinations PhaseByTransmission takes as input, the genotype likelihoods of all individuals, as computed by the UG (see above). If multiple trios are used at the same time for calling, the allele frequency for each site can be estimated and, subsequently, the allele frequency prior for each possible genotype, P_{AF}^{G}:

$$P_{AF}^{G} = \begin{cases} p^2 \text{ ,} if\ G\ is\ HOM\_REF \\ 2pq \text{ ,} if\ G\ is\ HET \\ q^2 \text{ ,} if\ G\ is\ HOM\_VAR \end{cases}$$

where p and q are the allele frequencies of the 2 present alleles at the respective site, i.e.: $p + q = 1$, as estimated from the set of samples provided. This prior encapsulates the expected Hardy-Weinberg equilibrium, which is in turn consistent with the genetic drift model described in Introduction. The allele frequency prior is then simply the probability of an individual having a combination of the two possible alleles identified for the respective site, under a random sampling with replacement model, where each allele's probability is its estimated allele frequency. We compute the allele frequency prior for each parent, as they are assumed to be unrelated (i.e.: independently sampled from the population). This prior cannot be applied to the offspring genotype, as its alleles are completely determined by its parents' alleles through Mendel's laws of inheritance. The offspring are not used in the computation of the allele frequencies, for the same reason.

We then define the likelihood of the data D given a trio genotype combination as follows:

$$P(D|G_M, G_F, G_C) = P(D|G_M) * P_{AF}^{G_M} * P(D|G_F) * P_{AF}^{G_F} * P(D|G_C) * P_C$$

where G_M, G_F and G_C are the genotypes of the mother, father and child respectively, $P(D|G)$ is the genotype likelihood for each individual as computed by UG, $P_{AF}^{G_M}$ and $P_{AF}^{G_F}$ are the allele frequency priors computed for the mother and the father's genotypes respectively and P_C is the transmission prior as computed in Table 1. The trio genotype likelihood is computed for each of the 27 possible configurations and the posterior probability is then obtained by a normalization step:

$$P(D|G_M^x, G_F^x, G_C^x) = \frac{P(D|G_M^x, G_F^x, G_C^x)}{\sum_{i=1}^{27} P(D|G_M^i, G_F^i, G_C^i)}$$

where i/x corresponds to the ordinal number for one of the 27 trio genotype combinations. Finally, the most likely combination is outputted along with a phred scaled quality score of the transmission posterior(TP):

$$TP = -10 * \log_{10}(1 - P(D|G_M^x, G_F^x, G_C^x))$$

where x indicates the most likely configuration.

2.7.3 Extending for the X chromosome

The described model is correctly defined for the diploid regions of the genome, i.e.: the autosomal chromosomes. We note that for the sex chromosomes, females are diploid, having an XX pair of chromosomes, whereas males are haploid, having an XY pair of non-homologous chromosomes and the inheritance pattern differs both for male and female offspring. Female offspring inherit one X chromosome from each parent and, as the father has only one, the paternal inheritance is deterministic. This reduces the number of possible trio genotype combinations to 18, out of which 8 are consistent with Mendelian inheritance patterns, 8 are indicative of one DNM, and 2 are indicative of two DNMs, as listed in Table 2.

	mother	father	child	#DNMs	transmission prior
1	AA	A	AA	0	1
2	AA	A	AB	1	1,00E-07
3	AA	A	BB	2	1,00E-14
4	AA	B	AA	1	1,00E-07
5	AA	B	AB	0	1
6	AA	B	BB	1	1,00E-07
7	AB	A	AA	0	0.5
8	AB	A	AB	0	0.5
9	AB	A	BB	1	1,00E-07
10	AB	B	AA	1	1,00E-07
11	AB	B	AB	0	0.5
12	AB	B	BB	0	0.5
13	BB	A	AA	1	1,00E-07
14	BB	A	AB	0	1
15	BB	A	BB	1	1,00E-07
16	BB	B	AA	2	1,00E-14
17	BB	B	AB	1	1,00E-07
18	BB	B	BB	0	1

Table 2: All the 18 possible X-linked trio-genotype combinations for a female offspring. Contains the symbolic genotype of the three individuals (where A and B denote possible alleles), the number of de-novo mutations that the combination contains and the corresponding transmission prior for each combination (where 10^{-8} is the per base probability of a DNM as estimated in literature)

Male offspring inherit the Y chromosome from the father and one X chromosome, only from the mother. This implies a number of 12 possible trio genotype combinations, 8 of which are consistent with Mendelian inheritance and 4 of which correspond to one DNM, as listed in Table 3.

	mother	father	child	#DNMs	transmission prior
1	AA	A	A	0	1
2	AA	A	B	1	1,00E-07
3	AA	B	A	0	1
4	AA	B	B	1	1,00E-07
5	AB	A	A	0	0.5
6	AB	A	B	0	0.5
7	AB	B	A	0	0.5
8	AB	B	B	0	0.5
9	BB	A	A	1	1,00E-07
10	BB	A	B	0	1
11	BB	B	A	1	1,00E-07
12	BB	B	B	0	1

Table 3: All the 12 possible X-linked trio-genotype combinations for a male offspring. Contains the symbolic genotype of the three individuals (where A and B denote possible alleles), the number of de-novo mutations that the combination contains and the corresponding transmission prior for each combination (where 10^{-8} is the per base probability of a DNM as estimated in literature)

We note that since male offspring do not inherit any $X - linked$ genetic code from the father, only 6 of the trio combinations listed in table# are uniquely informative w.r.t. DNM discovery (i.e.: father's genotype on the X chromosome is not informative) and we could therefore use only the mother-child combination to detect DNMs in this case. Father's genotype contribution to the likelihood function of a trio combination however ($= P(D|G_F) * \ P_{AF}^{G_F}$), can only increase discrimination power by increasing the likelihood of the best mother-child combination and decreasing the likelihood of some of the other mother – child combinations. Considering the lower sensitivity we notice in detecting male offspring DNMs (see results) we decide to consider the whole trio combination as opposed to mother-child pair, for $X - linked$ male offspring cases. Furthermore, since PBT can also be used as a "genotype score recalibration" method, by taking into account familial relationships, considering the father's genotype in this case offers more sensible results.

The second modification needed is that of the allele frequency prior, for the haploid genotype of the father:

$$P_{AF}^{G} = \begin{cases} p, & if\ G\ is\ HOM\_REF \\ q\ , & if\ G\ is\ HOM\_VAR \end{cases}$$

where p and q are again the estimated allele frequencies from the offered population of samples (excluding the offspring) and G is a haploid genotype.

2.7.4 Further extensions

Considering the properties of the sequencing data available and/or the nature of de-*novo* mutations, a number of further PBT extensions can be implemented, that would extend PBT's scope or detection power respectively.

Firstly, multi-allelic sites have been found in the population of GoNL as well as other large population sequencing projects (100 Genomes Project). The extension for this case would imply an increase in the number of possible trio genotype combinations, for which the defined likelihood function has to be evaluated. The number of possible diploid genotypes for n alleles is

$$n + \binom{n}{2} \ = \ \frac{n(n+1)}{2}$$

i.e.: n homozygous genotypes and $\binom{n}{2}$ unique heterozygous genotypes. The number of possible trio genotype combinations to be evaluated is then

$$\left(\frac{n(n+1)}{2}\right)^3$$

where the bi-allelic case corresponds to 27 unique genotype combinations. PBT thus has polynomial time complexity w.r.t. number of alleles, i.e.: $O(n^6)$.

Another considered modification is allowing for more complex family pedigrees horizontally and/or vertically. More specifically, as available familial information was used to estimate the joint probability distribution of a mother – father – child trio, additional familial information can be used to estimate the joint probability distribution of larger pedigrees; i.e.: including grandparents, uncles, siblings, etc. We note however that generalizing the above derived PBT complexity we obtain:

$$PBT(n,m) = \left(\frac{n(n+1)}{2}\right)^m = O(n^{2m})$$

pedigree combinations to evaluate, where n is the number of alleles and m is the size of pedigree considered. The exponential time complexity, w.r.t. pedigree size, makes the problem intractable for arbitrary sized pedigrees, all the more so considering the dimensionality of genetic data in terms of variant sites(i.e.: millions) and population size. By good use of domain knowledge however, some larger, fixed size pedigrees may be considered, or, alternatively, pruning of the pedigree combination search space may be applied. Siblings for example are good first candidates for pedigree extension and are more informative of one another's genotype, considering Mendelian rules of inheritance.

2.7.5 Machine Learning post-filtering

An initial run of PBT on the GoNL dataset produced a number of $60{,}968$ putative DNMs, still much higher than the $\sim 16{,}300$ expected number of DNMs. A number of sites were prioritized for validation (i.e.: accurate sequencing at much higher depth such that the new result can be considered underlying truth) and a training set of $2{,}265$ observations was produced, containing true positive DNM hits and false positive DNM hits.

This set was used to train a random forest model that would discriminate between true positives and false positives, thus fine-filtering the initial results. The model uses 22 features that were considered as possibly informative, including depths of coverage of individuals, various quality scores computed along the analysis pipeline, allele frequency, etc.. The trained model retained 12 of these features as informative and produced a test classification accuracy of 92.2%. Running the model on the initially found $60{,}968$ DNM hits, a set of $11{,}625$ high confidence hits was produced, for use in subsequent analysis.

2.8 HaploidWriter Walker

All of the GoNL data, including the X chromosome, was called using the diploid model described in the Variant Calling section. This is obviously not correct for the haploid X chromosome of male individuals in the dataset. While PBT can be implemented to use the appropriate information from a diploid called X chromosome, in addition to the adjusted inheritance pattern, incorrectly diploid genotyped samples still reduce PBT power significantly, mainly because of the wrongly computed PLs associated with each genotype.

The optimal solution would be to re-genotype the whole dataset using the haploid model of UG for the males. This is however practically cumbersome, as males and females would have to be called separately (i.e.: because the ploidy property must be identical for all samples pooled together for calling), and then merged together. Furthermore, the two data formats (i.e.: bam and vcf) are standalone and, typically, the genotyping is performed as an initial processing step when a new dataset is built. Genotype data for individuals at variant sites is typically all one needs for most downstream analyses. Given the variation, the rest of the genome can be simply read from the reference genome. The raw alignment data of all the individuals in the GoNL dataset for one chromosome would roughly be $\sim 625Gb$ whereas a vcf file containing all variation in the dataset for one chromosome is aprox. $12Gb$ large. This order of magnitude reduction in data size makes the use of vcf files preffered. Thus, it is not uncommon that datasets we wish to use are found in vcf format and the underlying alignment data is not available, thus not allowing us to re-genotype individuals. A method of inferring the haploid genotype from an initially called diploid genotype is then a useful tool for a pipeline analysing a haploid chromosome ($X\ or\ Y$).

To this end we implemented HaploidWriter, A RodWalker that takes as input some vcf file and a file specifying genomic intervals in the form of chromosome position and index start-end values, as well as a ped-file specifying familial relationships between vcf samples and/or sex. The walker traverses the genotypes of all individuals at every genomic positions. For genotypes of individuals that are males (as specified by sex), or for which the walker can infer that are males (from familial relationships), and that fall within the genomic positions specified by input, the walker transforms the respective genotypes to haploid (if they were initially called as diploid). The transformation is applied to the annotations needed by the PBT, namely genotype alleles and phred-scaled likelihoods, but it can straightforwardly be extended to new annotations, as they are found relevant for other analyses.

We use the case where two alleles are present in the population at a genomic position, although HaploidWriter can be applied to multi-allelic sites as well. For bi-allelic sites, the genotype is a combination of the two possible alleles and the likelihoods correspond to the three possible allelic combinations, in a fixed order: $PL_{HOM\_REF}, PL_{HET}, PL_{HOM\_VAR}$. A haploid genotype contains by definition only one allele, therefore it can only be homozygous, for some allele. If the initially called diploid genotype is homozygous for either one of the two alleles, then the haploid genotype is also homozygous for that allele. If the initial diploid genotype is heterozygous, the allele of the haploid genotype is determined by comparing the likelihoods of the diploid genotype. The smaller of the two homozygous alleles will be the allele assigned to the haploid genotype. For example an initially called genotype $/B$, where A is the reference allele, with $PL = 20{,}0{,}50$ will become an A haploid genotype. This is correct because the PL encode the amount of evidence, that was found in the alignment data, that supports each possible genotype. In the given example, the called genotype was HET, but the genotyper found that there is more evidence (or better quality) for the reference allele than for the alternative one, thus, knowing a-priori than the genotype cannot contain both alleles, we assign the reference one. Typically however, haploid genotypes that for which evidence of two (or more) alleles is found, are an indication of very bad quality data (i.e.: systematic alignment errors due to repetitive regions or indels, etc.).

The PLs are modified in a straightforward manner as follows; the inverse phred transformation is applied to obtain the probabilities of each original genotype

$$p_{GT} = 10^{-\frac{PL_{GT}}{10}}$$

where GT is each of the three possible genotypes and PL_{GT} is the corresponding likelihood. The probability of the initially called genotype will be the only one set to 1 after the transformation (due to the forced 0 PL value) and is recomputed

$$p_{calledGT} = 1 - \sum_{GT \neq calledGT} p_{GT}$$

The probabilities corresponding to homozygous genotypes only are kept (i.e.: in the bi-allelic case only the first and third), they are re-normalized to sum to 1 and the phred transformation is applied. For the above given example the haploid genotype would be $GT = A$, with $PL = 0{,}30$.

We expect the implementation to provide a good estimation, for all intended practical usages, of the correct haploid genotype and PLs of an initially called diploid genotype, but we also expect systematic differences when compared to the PLs computed by the haploid method of genotype calling of the UG. Intuitively, this stems from the likelihood based approach used by the UG. Namely, observing the same evidence b (i.e.: read containing this base) weights more to the final likelihood under an expected haploid genotype than under an expected diploid genotype:

$$P(b|A_1A_2) = P(b|A_1)\frac{1}{2} + P\ (b|A_2)\frac{1}{2}$$

and

$$P(b|A) = \begin{cases} 1 - e, & if\ b = A \\ \frac{e}{3}, & if\ b\ \neq A \end{cases}, where\ A\ is\ \ a\ possible\ allele$$

whereas

$$P(b|A_1) = \ P(b|A_1) = \begin{cases} 1 - e, & if\ b = A_1 \\ \frac{e}{3}, & if\ b\ \ \neq A_1 \end{cases},$$

$$where\ A_1\ is\ the\ underlying\ haploid\ genotype\ and\ e\ is\ the\ sequencer\ error\ probability$$

This difference would account for a systematic underestimation of the PLs from our HaploidWriter, w.r.t. UG. A less direct influence also comes from the Expectation Maximisation algorithm of UG that estimates the allele frequency spectrum at each position, prior to deriving the final genotype likelihoods. As the allele frequency is marginalized over each ploid (individual chromosome), having a lower number of chromosomes produces a different estimation.

2.9 De*Novo*MutationPowerCaller

2.9.1 Motivation

With the implemented PhaseByTransmission model in place, the following analysis pipeline can be used: We apply PBT to trio-based sequencing data and obtain a set of likely DNM candidate sites. The filtering method described eliminates some of the false positive hits producing a subset of high confidence hits. These are further confirmed through deep sequencing validation and the ones found to be true DNMs are used in subsequent analysis. We further wish to assess the detection power of PBT, namely how likely is PBT to detect a de-*novo* mutation at some locus in the genome, if one were present. Applying this to the whole genome we obtain a map of accessible versus inaccessible regions with respect to PBT de-*novo* detection. This is used as an indication as to what regions can be confidently assessed and what regions cannot, as we would be unlikely to detect any DNMs in that region. We are primarily interested in identifying and characterizing the regions where PBT has very low detection power (i.e.: what we are likely to miss).

To this end we build the DeNovoMutationPowerCaller (DNMPC) that outputs the probability of PBT detecting a DNM at any locus in the genome based on features of the sequencing data available for that region. We note that DNMPC provides information PBT's detection ability at a locus, under the assumption that one is present, and it gives no indication as to whether a mutation *is* actually present.

Multiple factors are likely to influence discovery power, including the sequencing depth of cover in each of the trio members, the quality of the mapping at the site, the genomic region's susceptibility to mutation, etc. It is not trivial to know which of these factors are most relevant for detection power, nor how they influence it. We therefore train a classification model to evaluate their joint discriminatory power in separating callable and uncallable regions. In order to build such a model we need a dataset comprising of sequencing data at sites that cover both detected and undetected DNMs. By definition, we cannot know apriori where false negative sites lie in the real sequencing data. Thus, an artificial dataset of "simulated" de-*novo* mutations is built from the GoNL sequencing data, so as to enable us to properly train a classification model.

2.9.2 Simulation data

A set consisting of 100,000 simulated de-*novo* mutations was created. Firstly, the positions to insert an artificial DNM were sampled uniformly and uniquely across the whole reference genome. Sites that are known to be polymorphic were filtered out. Then each position was attributed to one GoNL family randomly, also following a uniform distribution (across available families). Finally, a mutation was artificially inserted in the child data by changing one of the genotype alleles such that the respective trio genotypes combination would result in a Mendellian violation. So as to be able to train a realistic model on the artificially created dataset, the mutations are inserted in such a way as to preserve the error rates and the error profiles as much as possible. To this end the modifications are made at the lowest level possible in the analysis pipeline, namely at individual read's level. The set of reads covering the target position in the child are selected and modified such that they should intuitively result in a different genotype call. For each read, the base aligned to the desired position is flipped to another "mutated" base with 0.5 probability to simulate the binomial sampling of two different alleles in a diploid context. The inserted mutated base for each position is selected randomly from the 3 possible alternatives so as to not introduce additional biases. Some reads covering the targeted locus may have misalignments for that position, namely bases other than the reference allele that are however insufficient evidence to suggest polymorphism; they typically come from errors in the sequencing technology. These bases are not altered to preserve local error profiles. On this modified trio sequencing data we run the UnifiedGenotyper to obtain new genotype calls. Subsequently we apply PBT to check for DNMs, as well as the random forest filtering step, for better sensitivity. Excluding the filtered DNMs and other exceptional cases resulting from the random sampling, such as positions outside the accessible genome, etc., we obtain a total set of 84819 data points, which we consider to be *true* underlying DNMs, that we can use for training and testing the model. For each such data point, a set of features is computed from the sequencing data.

2.9.3 Model description

The model that we build is to be implemented into the GATK platform and it should be practical, with respect to time resources, to run it on a whole genome, i.e.: input size in the order of 10^9. The resulting model should then have a simple allocation rule. Moreover, online computation of the features, at each locus, should be performed in, preferably, constant time.

A subset of features from our dataset, that are trivially computable within a GATK walker, was selected:

- father's depth of coverage – father_dp
- mother's depth of coverage – mother_dp
- child's depth of coverage – child_dp
- mapping quality – MQ
- GC content within a 200 base-pair window, on the reference sequence, centred on the evaluated position – GC-content
- the sequence of 3 consecutive base-pairs, on the reference sequence, centred on the evaluated position – Triplet

We may interrogate the GATK walker's traversal engine for the depth of coverage of any sample in the supplied alignment input files at any currently processed locus. Thus, the coverage of each individual family member (father_dp, mother_dp and child_dp) may easily be used.

Mapping quality (MQ) is computed as the average of the mapping qualities for all the reads that cover the locus(for the three individuals) and is thus computable in constant time, i.e.: $O(3 * AvgDP)$ where $AvgDP$ is the average depth of coverage across the genome that GoNL was sequenced at(i.e.: 12). We note that in creating the simulated mutations, we basically introduced a misalignment in aprox. half of the child's reads without correcting the mapping quality scores. This makes the Mapping Quality that we compute a slight overestimation. Considering that each DNM contains an equal expected number of modified reads(AvgDP/2), we consider this bias a, small, constant "global" shift of the true MQ values, that cannot influence discrimination power however. Mapping quality and the depths of coverage are natural choices for features as they relate directly to the confidence in genotyping a position, based on which we call DNMs.

GC-Content is computed as the percentage of basses that are either a C or a G in a window that spans 100 positions left and right respectively, of the current locus. This was also shown to influence the occurrence of mutations. GC-content is computed on the reference genome, in constant time.

Triplet represents the window of two bases immediately adjacent to the current locus, also on the reference genome. Chemistry between a base and its immediate neighbours was found significant in processes operating on the DNA molecule, therefore it can be informative for our model as well.

Running PBT and the subsequent random forest filtering on the built dataset, we find that 76% of the artificially created DNMs were detected as such, while the remainder of 24% of sites, although containing data suggestive of a mutation, were not recognized by PBT. These two sets represent the two target classes consisting of true positives (TP) and false negatives (FN) respectively. We first note the prior class distribution of 76% to 24% (TP to FN). This is an indication of good PBT performance; by the uniform sampling method this implies that a DNM call can be made at 76% of the sites that one may occur (i.e.: across the genome). This high calling rate on true DNMs makes the task of discriminating the non-callable ones harder.

2.9.4 Selection of the model and training

A default model of assigning the majority class to all points already results in a classification performance of 76%. However, we are primarily interested in obtaining a good performance on the false negatives. To this end, we perform a number of simulations, in R, training simple classification models in order to select the best one, with respect to accuracy as well as simple allocation rule. The data is split into two sets, for training and testing, comprising of 80% and 20% of the data-points respectively. The models considered are:

- K-nearest neighbours (knn)
- Linear regression
- Logistic regression
- Shrinkage discriminant analysis (sda)

Initial training of the models shows the following performances on the test-set, where performance is $\frac{\#correct\ classifications}{\#test\ points}$:

- Knn – 86%
- Linear regression – 37%
- Logistic regression – 52%
- Sda – 93%

The very good overall performance of the shrinkage discriminant model is misleading for our goals. The sda method estimates allocation priors based on the class distribution of the points in the training set.

This leads to very different within-class performances, namely to a performance of 16% on the false negatives (FN). Given that we are interested in obtaining similar within-class performances (FN and TP respectively), this is not satisfactory. Dropping the sda-estimated priors, the method behaves similarly to a linear regression model, with an overall performance of 42%.

In order to obtain a good performance of the knn model, we employed a local grid search for the model-parameters. The model obtaining the overall 86% performance (with almost identical within-class performances) uses a “neighbour set” of 20 random data points from the sets of FNs and TPs respectively and the allocation rule predicts a class by the majority vote of 15 closest neighbours. Considering the input size ($\sim 10^9$) and the size of the neighbour set needed for good performance (i.e.: 40 neighbours) we considered the model sub-optimal with respect to implementation. Furthermore, bootstrapping revealed, as expected, that the model is not sufficiently stable, by obtaining an average performance of 82% with a variance of 10, over 40 bootstrapping runs.

The $\sim 37\%$ overall performance of the linear regression model is the best obtained performance of several tried models including subsets of features as well as feature interactions. In addition to the low performance observed, a proper allocation rule has to be devised for proper usage (i.e.: the linear models regress the features provided to the discrete output space of 0 – FN and 1 – TP). We observed that approximating that prediction by approximating the regression output to the nearest class label (i.e.: 0 or 1) is not feasible. By computing the empirical averages of the model’s output on the FN and TP points respectively, when run on the training set, we computed the “empirical” labels, for FNs and TPs respectively, to be 0.57 and 0.91 respectively. The performance on the test set was thus computed by rounding to the derived “empirical” label values. From several bootstrapping runs however, we found these average values to have high variance and we thus considered this model to be unreliable.

The logistic regression also showed poor performance ($\sim 52\%$) however, the allocation rule is straight forward, as the output is the probability of a data point belonging to one of the two classes (i.e.: TPs). Furthermore, the computation of this probability is done in constant time once the model has been trained. We therefore decided to use logistic regression for subsequent simulations.

To further inspect the low performance of the logistic regression model, we plotted the within class histograms for all the selected features (Figure 10). We notice that there is little information with respect to discrimination power in any of them. This is mainly because of the disproportionate class distribution which makes the true positives dominate the false negatives across the spectrum. Plotting the within class distributions (i.e.: normalizing the histograms) we observe that discrimination power of each individual feature is increased significantly. In order to enforce this property in the training process, the training set was selected such that the two classes are equally represented (FN and TP). We selected 80% of the TP data points and included them in the training set. To these we added an equally large set of FNs obtained by sampling with replacement from a subset of all FNs, representing 80% of the FN data points available.

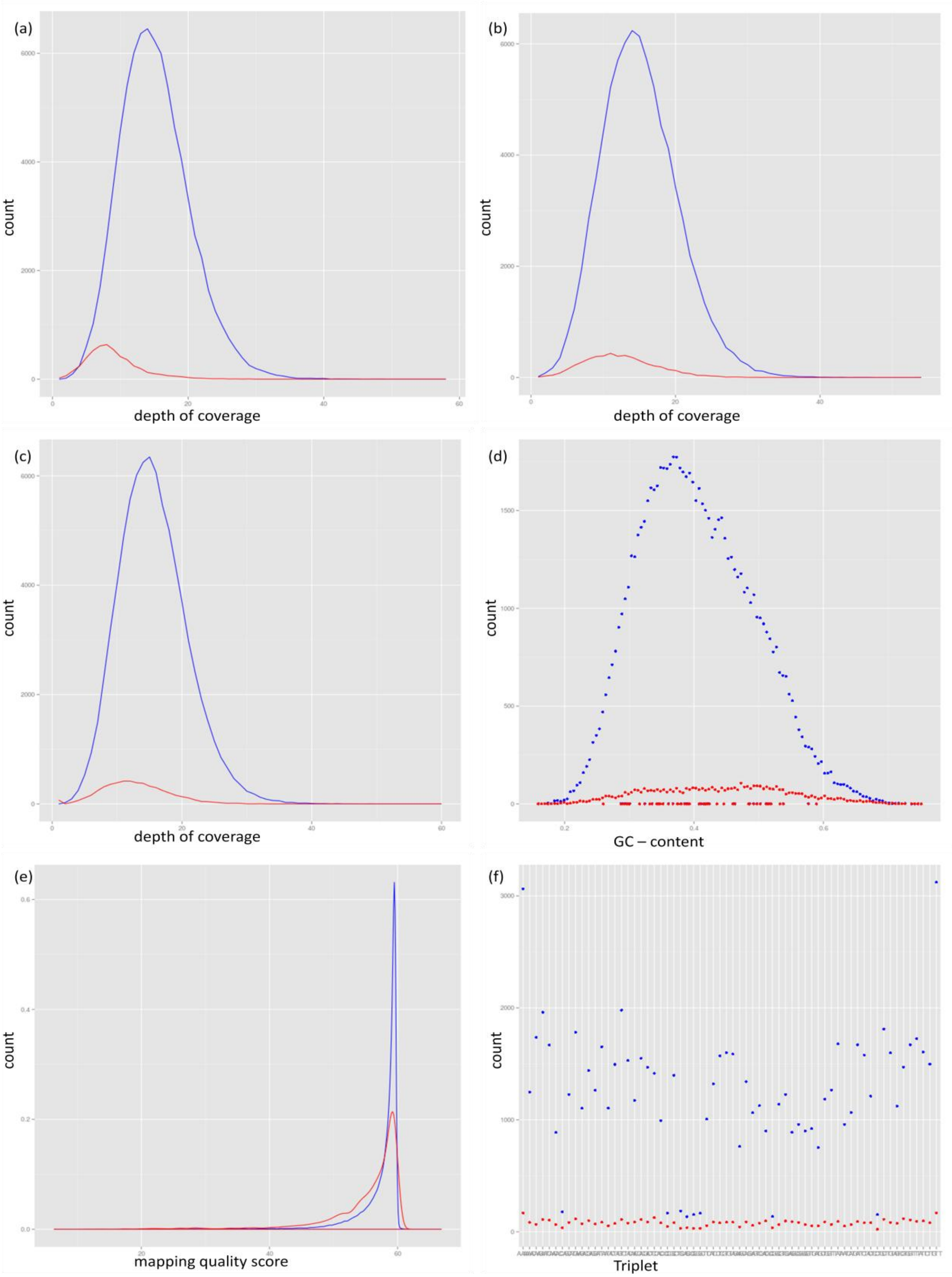

Figure 10 : plots of the histograms of each of the considered features. Color blue marks the class of true positives (TPs) and color red marks the class of false negatives (FNs). (a) : child's depth of coverage. (b) : father's depth of coverage. (c) : mother's depth of coverage. (d) : GC-content. (e) : mapping quality. (f) : Triplet

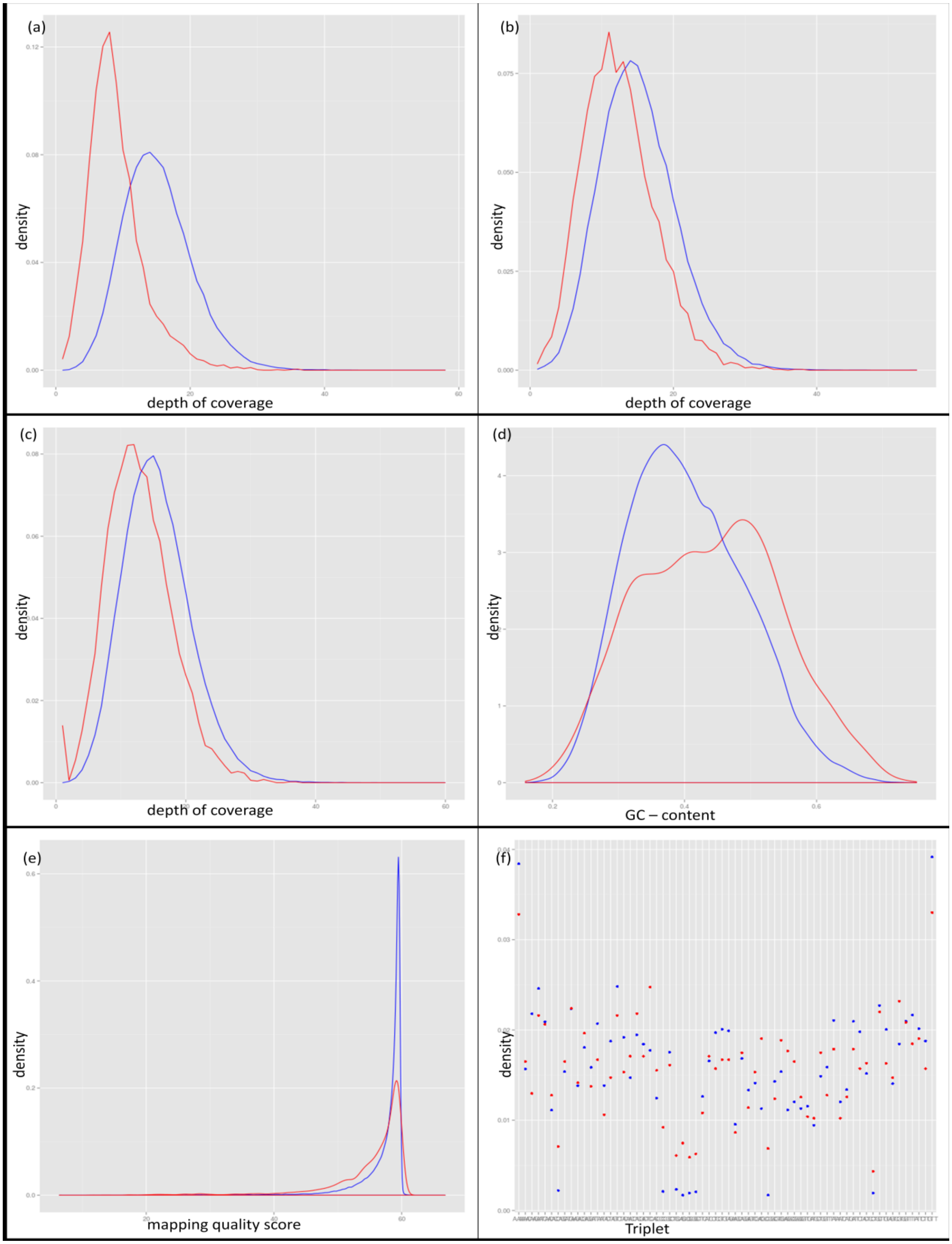

Figure 11 : plots of the densities of each of the considered features (i.e.: normalized histograms). Color blue marks the class of true positives (TPs) and color red marks the class of false negatives (FNs). (a) : child's depth of coverage. (b) : father's depth of coverage. (c) : mother's depth of coverage. (d) : GC-content. (e) : mapping quality. (f) : Triplet

We re-trained the logistic regression model on the training set described above and obtained an overall performance of 81% on the test set, with a performance of 83% on the FNs and a performance of 80% on the TPs. The model also showed good stability across 40 bootstrapping runs, so we selected it for implementation within the GATK.

The empirically found optimal logistic regression model uses only the three individuals' depths of coverage as covariates (father_dp, mother_dp and child_dp) and the GC-content. The mapping quality (MQ) and Triplet features were found to be insufficiently informative overall (Figure 11e and Figure 11 f) and were dropped. Furthermore, the GC-content, which is a feature of the reference sequence and not of the sequenced DNA, was found to have the largest regression coefficient.

Finally, we performed a test of the trained logistic regression model on real data. Namely, we computed the four selected features' values for a set of $\sim 17{,}000$ real de-*novo* mutations from the GoNL dataset and built a new test set. The observed performance on this set was 81%, thus concordant with the simulated DNMs.

3. Results and Discussion

3.1 PhaseByTransmission

To evaluate the performance and the improvement in discovery power of PBT on the X chromosome, we ran and compared the ploidy-aware version of PBT against the original version of PBT which assumes a ploidy of 2 at every site. We denote the original PBT as PBTO (pbt-original) and the ploidy-aware version as PBTX for the purposes of our discussion. The original version is run as a benchmark for evaluating the ploidy-aware version.

Because the original PBT assumes diploidy in all samples (including males), inheritance model is not correct for chromosome X, which results in an incorrect transmission prior for trio combinations that do not contain a DNM. This results in wrongly computed and represented genotype likelihoods (PLs), which are used to compute the joint trio-genotype likelihood. As the haploid genotypes contain only one allele, they can only be homozygous (either on the reference or the variant allele), so heterozygous likelihoods are meaningless and simply an artefact of the initial genotype calling performed on this chromosome. This notwithstanding, we expect PBTO to also detect part of some class of the true DNMs. Specifically, PBTO should have some power in correctly calling genotype combinations where the parents are both HOM_REF and the child is HOM_VAR (in case it is a boy) or HET (in case it is a girl):

$$HOM\_REF \ - \ HOM\_REF - HOM\_VAR \rightarrow male\ child, 1\ DNM$$

$$HOM\_REF \ - \ HOM\_REF - HET \rightarrow female\ child, 1\ DNM$$

Although the inheritance pattern is wrong, it correctly identifies one DNM for each of these specific cases and the modelling errors would only manifest in an underestimation of the transmission posterior (TP).

3.1.1 Initial Runs

As input for all our runs we have a VCF file containing all variation found on the X chromosome. The file was created by running the UnifiedGenotyper on the BAM files of all individuals in the GoNL dataset.

We expect an overall higher level of noise for the X chromosome, compared to autosomal chromosomes, mainly because all males are haploid. As we showed in Methods, this results in a lower average coverage for males. For females, the X chromosome can be treated as any autosomal chromosome. For males, we first expect a lower average coverage, thus reducing the statistical robustness of detecting variants and implicitly DNMs.

PBTO is run on the initial VCF file. In order to run PBTX, the initial, diploid called VCF file is parsed using HaploidWriter and the genotypes for all male individuals found are made haploid (w.r.t. genotype and phred scaled likelihoods). Two sets of parameters were tried for each PBT version respectively. First run uses a DNM prior of 10^{-4}. We use a DNM prior four orders of magnitude bigger than the empirically estimated one, of $\sim 10^{-8}$, so as to have high sensitivity in detecting true DNMs at the expense of more false positive hits. Many of the false positive hits can be afterwards filtered according to known error modes. The second run uses the same mutation prior plus the allele frequency prior computed with the allele dosage method. Allele frequency priors are computed, at each position, from the parents' called genotypes (and respective PLs). Because GATK did not yet fully support haploid genotypes/positions, the allele frequency used in PBTX is estimated based only on the mothers' genotypes, whereas the allele frequency used by PBTO considers all parents.

3.1.2 Processing Results

Initial results are summarized in Table 4. We first note a very big discrepancy in raw numbers of DNM calls. Eliminating all the pseudo-autosomal hits, for the reasons described above, we obtain an initial number of ~ 16500 DNM candidates identified by PBTO and ~ 51000 identified by PBTX. The larger proportion of both these numbers were expected to be false positive, because of the permissive prior used.

The significantly higher number of DNM hits that PBTX finds (i.e.: $\sim 3\ times\ more$) suggest a higher detection power, but also contains significantly more noise. The additional noise originates by an implicit filtering that PBTO performs, while PBTX does not. Namely, by treating the X-chromosome in an autosomal manner, PBTO considers 3 phred scaled likelihood values (PLs corresponding to the 3 possible genotypes of a biallelic site), for all trio individuals. This allows the method to "correct" the initially called genotypes in the case of very poor quality data. For example, for a trio genotype combination containing one DNM where one individual's PLs are $0{,}20{,}200$ (initially called HOM_REF, with a 10^{-2} probability of being HET, etc.), using a DNM prior of 10^{-4}, PBTO finds it more likely that the individual was missgenotyped and is in fact a HET, thus eliminating the DNM. This implicit filtering is severly impaired for PBTX because males are haploid and, typically, PL values are higher; i.e.: diploid PLs $0{,}20{,}200$ become haploid PLs $0{,}230$.

The second data cleaning criteria we apply is discarding hits that did not pass VQSR. As described in methods, the Variant Quality Score Recalibration evaluates the probability that a site where some variant is called is truly variant, or whether it is more likely to be the result of sequencing/alignment errors. The DNM hits outside the pseudo-autosomal regions that also pass VQSR are the basis of comparison for evaluating the improvement that PBTX brings.

Running both PBTO and PBTX with the allele frequency prior produces initial call sets that are larger by 30% and 10% respectively. Considering the property that the AF prior encodes(i.e.: the Hardy-Weinberg equilibrium), as well as the properties of the sequencing technology, we observe that using this prior increases calling sensitivity after data cleaning is performed as well.

	no AF prior		with AF prior	
PBT version	PBTO	PBTX	PBTO	PBTX
all	17418	102938	22448	112675
non pseudo-autosomal	16449	51076	20639	56599
VQSR - PASS	14450	40024	17498	42878

Table 4: Raw results of running the two PBT versions with and without the allele frequency prior. Results after each data cleaning step are shown

3.1.3 The Allele Frequency prior

As the data shows, using the AF prior increases sensitivity of the resulting DNM hits at the expense of an easy to eliminate noise. The reason lies in the nature of the sequencing process and is backed up by previous PBT DNM calling endeavours. Namely, we have better likelihood of finding DNMs at sites that were previously monomorphic (or very close to monomorphic), as opposed to sites known to be polymorphic, which is in turn explained by the binomial distribution of reads from each of the two chromatids at some position. If a site is known to harbour a polymorphism in the population, there is a probability, proportional to the allele frequencies at that site, that an individual has one or two of either alleles. Considering the average coverage of 12, of GoNL data, and the genome size, it is expected that, by chance, there will be individuals with HET positions where one of the alleles is severely under-represented or simply not captured at all, thus resulting in incorrect genotyping. The rate of false positives is thus increased by not having sufficient power to detect the child's "mutant" allele in one of the parents. In order for the probability of this happening to become insignificant (i.e.: drop below 0.05) we would require a coverage of at least 36 overlapping read per locus. The AF prior thus increases discrimination power by increasing the posterior probability of trio-genotype combinations where parents are HOM_REF at monomorphic sites, as opposed to trio-genotype combinations where parents are HOM_REF at polymorphic sites. As expected, we observed that PBTX assigns higher transmission posterior values for mutations detected in monomorphic sites when using the AF prior (Figure 12).

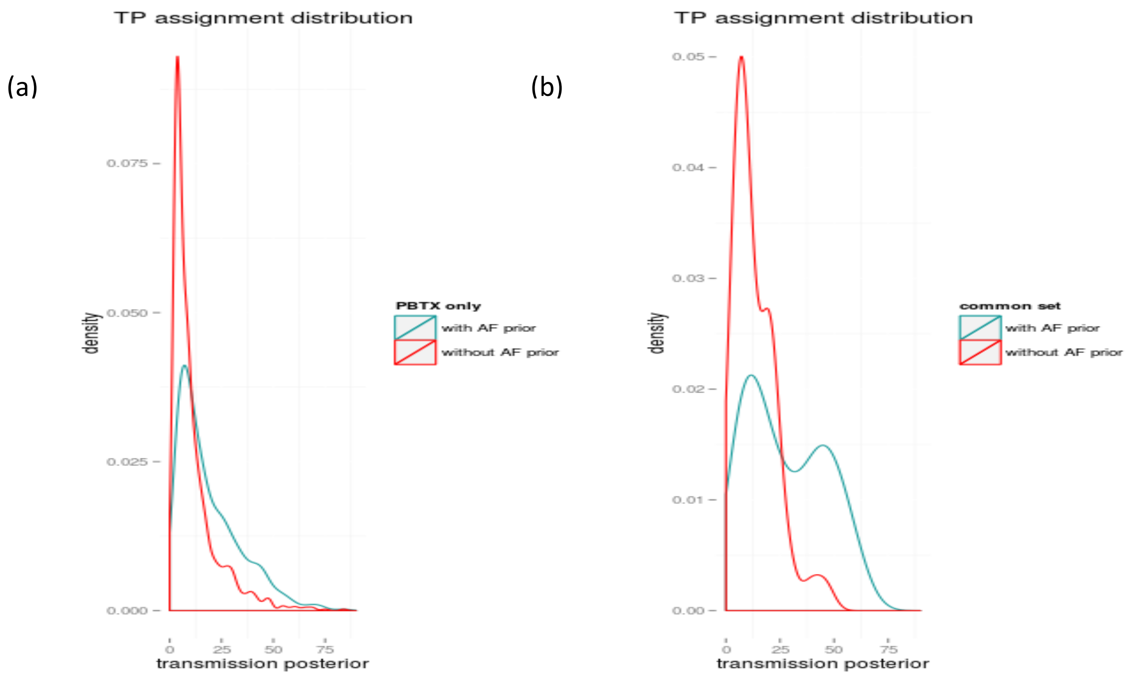

Figure 12: The PBT assigned transmission posterior distribution with and without using the AF prior. The sets used for ploting are the sets after the initial data cleaning. (a): TP distribution over the set of hits found obly by PBTX. (b): TP distribution over the set of hits found by both PBTX and PBTO.

3.1.4 Comparison of Results

Further in our comparison we looked at DNMs that were called by both methods and sets of DNMs that were called exclusively by PBTO and PBTX respectively. We will refer to the results obtained from the two runs using the AF prior for clarity and because the performance is better in all stages, but a direct comparison can be made from Table 5.

Using the AF prior in running both PBTO and PBTX we obtain a set of 459 mutations, throughout all families, that were called by both versions, with the remainder of each set being called only by either method separately. We filter these sets, by systematic error modes found so as to obtain better quality candidate sets. The first filter we apply to these sets is that of restricting possible DNMs to combinations where both parents were called as HOM_REF. We further filter out homozygous genotypes (all parents and boys offspring) that contain any amount of evidence of the other allele in their allele depth (AD) field. Lastly, we filter out positions that have an alternative allele count (AC) larger than one across the whole population used for calling, the one corresponding to the mutant allele present in the child of the respective DNM combination.

Because we anticipated different recognition power between male and female offspring, due mainly to coverage-driven quality differences, we split the mutations by sex. We indeed notice a consistently higher number of mutations found in girl offspring, by contrast to boy offspring. The sequential results after applying each filter are illustrated in table5.

The homozygous reference parents filter enforces the AF-prior reasoning described above. We note that the set of $Common$ hits is only sensibly adjusted while the set $onlyX$ of mutations found exclusively by PBTX is reduced by 96% to a remainder of 2093 candidates and the set $onlyOld$ of mutations found exclusively by PBTO is left with only 2 candidates. This shows that PBTX remains at least as sensitive as PBTO. The 2,093 candidates found only by PBTX need to be further assessed for quality in order to assess whether PBTX is more sensitive or less specific than PBTO.

Evidence of the other allele in a homozygous genotype, even if not sufficient to make UG call the position as HET, is typically an indication that the other allele is in fact present, but the binomial reads amplification did not amplify the two alleles equally at this position or, in the case of haploid genotypes, very bad quality data in terms of alignment. Lastly, we are more confident in mutations found at previously monomorphic sites, because of the sequencing technology pipeline and the coverage-driven power limitation, namely the same error mode that one of the two present alleles may not be captured. Sites with higher than 1 alternative allele count, are also indicative of regions with systematic mapping problems.

We consider the resulting sets high confidence DNM candidates and we note that using PBTX roughly doubles the numbers of such mutations, compared to using the initial, PBTO version. We note that, despite having substantially larger detection power both for male-offspring mutations as well as well as for female-offspring, a significant difference in number of found mutations remains, likely due to covariates not captured by PBT.

We also note that after we filter for what is most likely to be noise, there are no mutations found by PBTO that are not found by PBTX.

	no AF prior						with AF prior					
subset	Common		onlyX		onlyOld		Common		onlyX		onlyOld	
initial	89		39935		14361		459		42419		17039	
Hom_Ref parents	44		1140		0		428		2093		2	
offspring sex	m	f	m	f	m	f	m	f	m	f	m	f
initial	21	23	207	933	0	0	134	294	708	1385	0	2
"clean" ADs	12	14	58	308	0	0	54	231	152	404	0	1
AC = 1	3	9	19	169	0	0	23	204	40	222	0	0

Table 5 : The sets of DNM hits after each consecutive filtering step.

3.1.5 Explaining Results

The improved performance PBTX shows, w.r.t. PBTO can be found in features of the calls produced by the two versions. The use of a haploid model makes PBTX assign higher transmission posteriors to the mutations found by both methods, as depicted in Figure 13a. The same observation stands, although the distribution suggests a smaller difference, if we consider all the mutations that each method finds (Figure 13b).

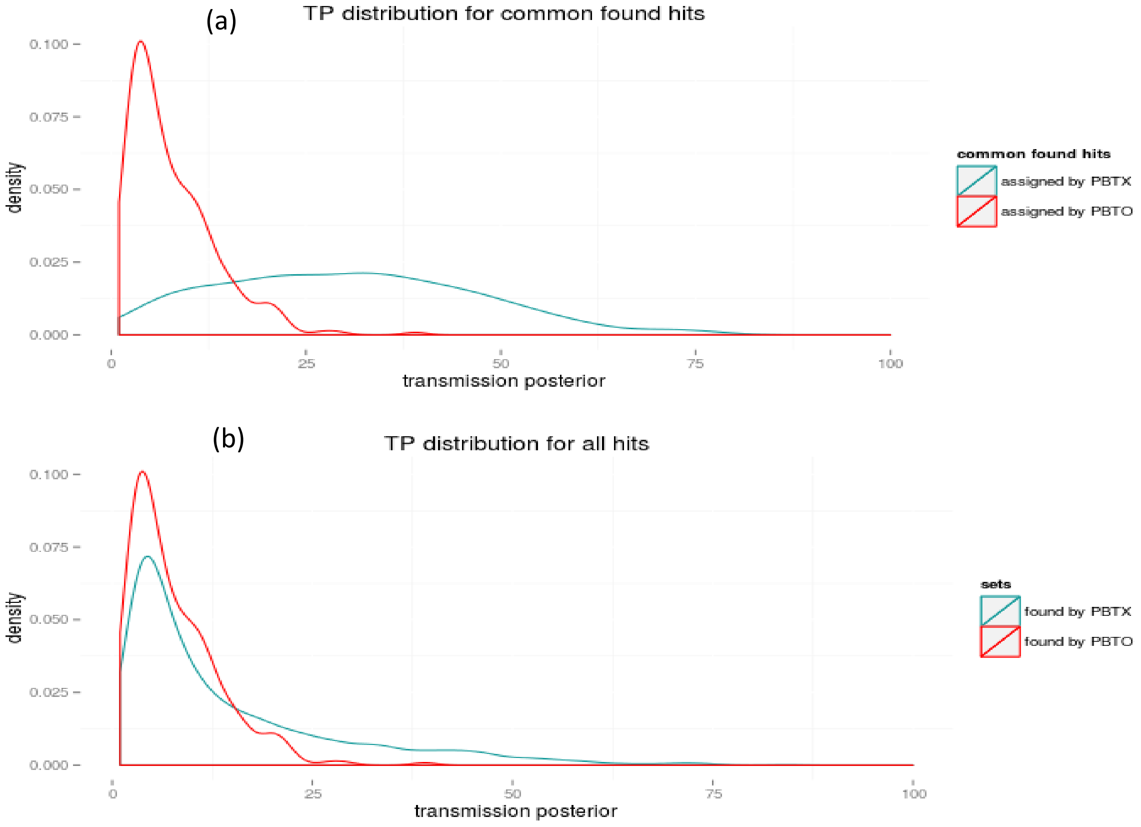

Figure 13 : Transmission posterior density plots. (a): TPs assigned by either method to the set of commonly found hits. (b): TPs assigned by either method on the set of all mutations found by each respective method.

Coverage of each individual in the trio influences directly the initial genotype quality thus detection power. Figure 14, Figure 15 and Figure 16 show the depth of coverage distribution by individual, within the highly likely candidate sets. We plotted separately for boy and girl offspring to get insight into whether there are systematic reasons for the big discrepancy in number of mutations. Furthermore, the plot of (mutant) allele count density gives an intuition of the behaviour of the AF prior. We further distinguished between mutations found by both methods, and mutations found only by PBTX (there were no high confidence mutations found only by PBTO).

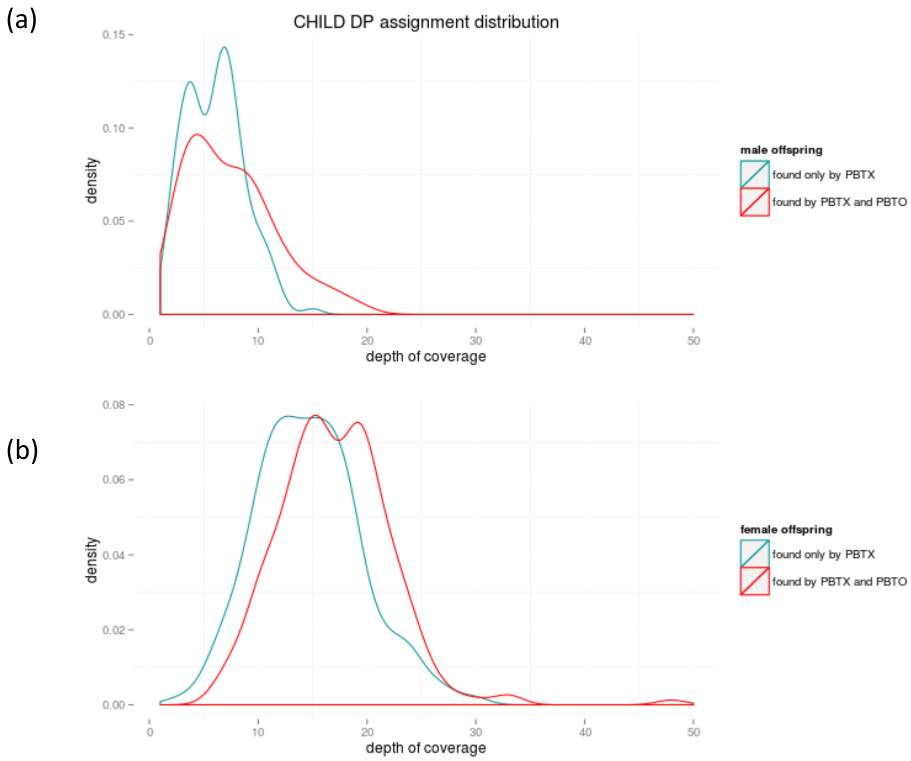

Figure 14 : Offspring depth of coverage density, within the high confidence sets. (a): for male offspring. (b): for female offspring.

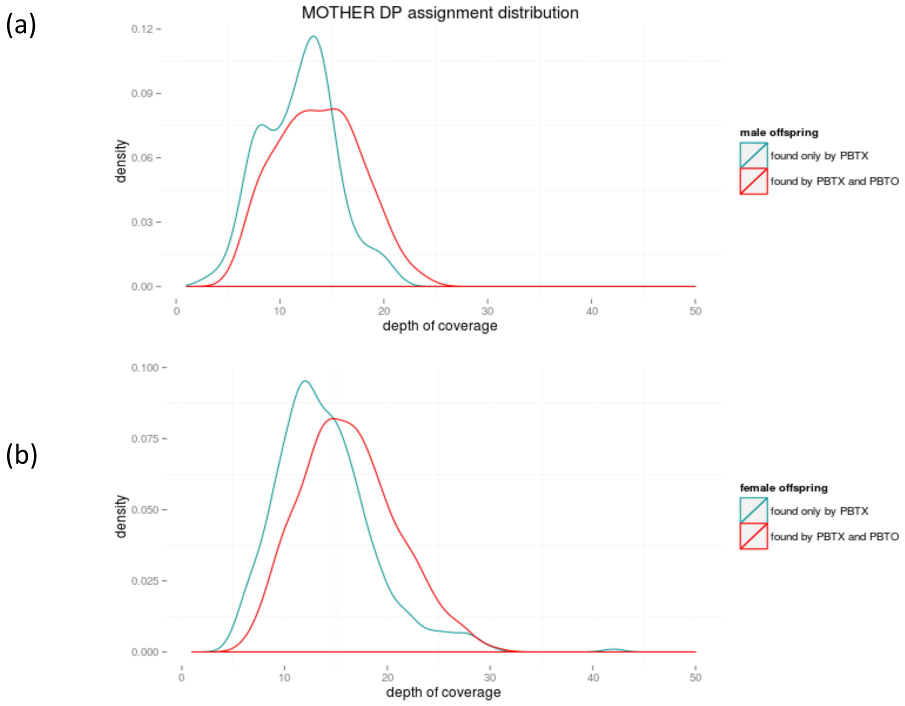

Figure 15 : Mother depth of coverage density, within the high confidence sets. (a): for male offspring. (b): for female offspring.

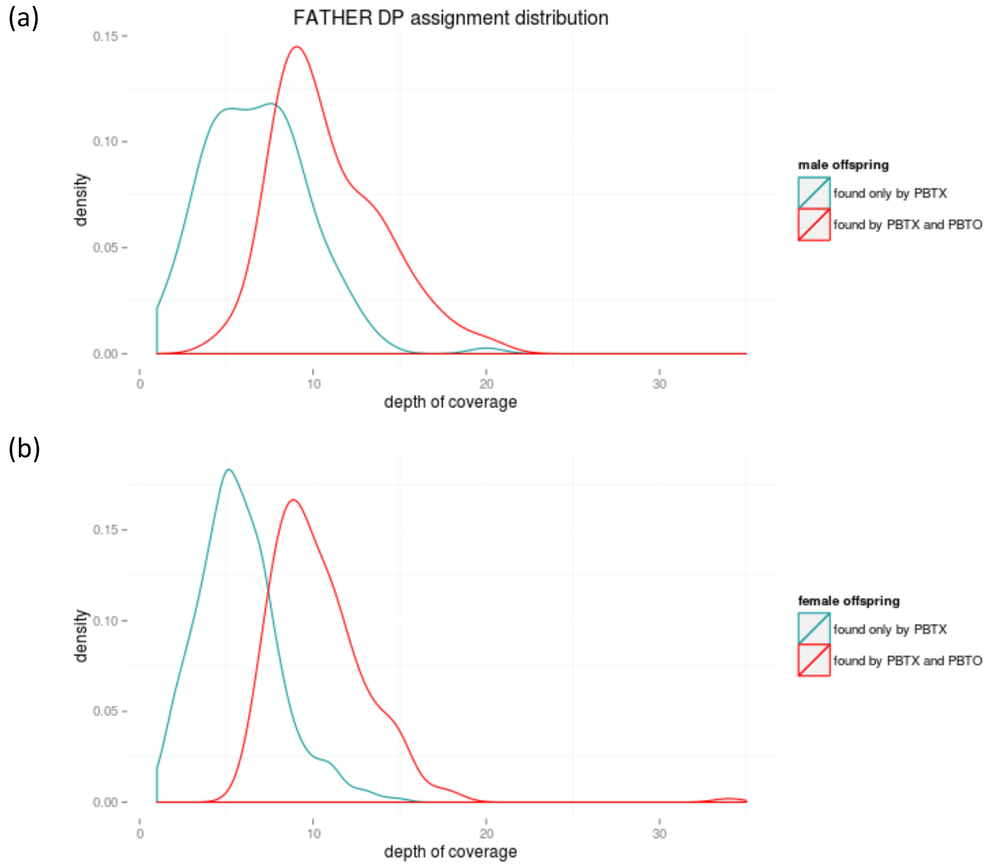

Figure 16 : Father depth of coverage density, within the high confidence sets. (a): for male offspring. (b): for female offspring.

We first note that all the feature distributions for the set of common found DNMs, by PBTX and PBTO, which are directly correlated to detection performance, have a higher mean, and in some cases smaller variance (i.e.: father coverage for boy offspring DNMs) than the mutations found only by PBTX. This is to be expected, as the PBTO decision boundary allows us to capture a set of well captured, i.e.: by sequencing, mutations. By properly adjusting the model in PBTX, we allow for better discrimination across a larger spread of the sequencing data output spectrum.

The main difference observable between boy and girl candidate mutations is the lower coverage of the offspring. This is expected, as the males are haploid, but it does create lower quality genotype calls for males. A correlated source of lower data quality/detection power in boys w.r.t. girls can be observed by plotting the allele count distribution of the candidate hits (i.e.: on the sets obtained $before$ applying the AC=1 filter) as depicted in Figure 17.

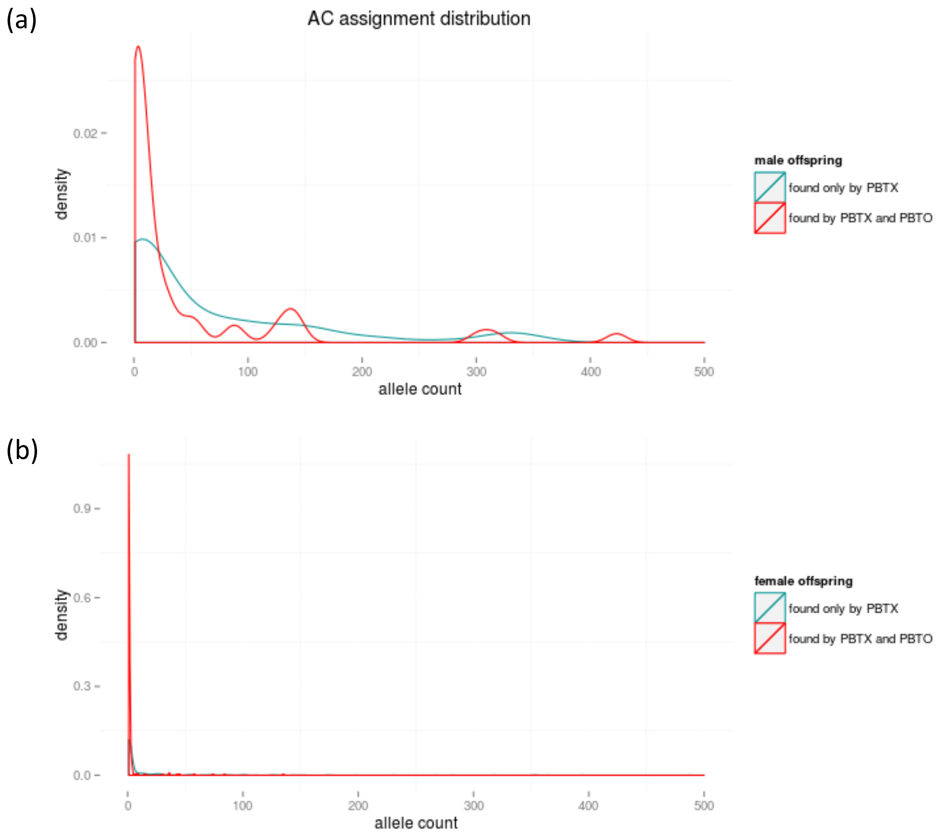

Figure 17 : Allele count density plots for the high confidence sets of mutation, before applying the AC filter. (a): for male offspring. (b): for female offspring

We note that this filter reduces the set of mutations found in boys $3.3 - fold$ and the set of mutations found in girls $1.49 - fold$ only. We argue that lower coverage in males (fathers and male-children) corroborated with alignment mismatches and/or hard to align regions of the genome, can impair the correct assessment of the alternative AC at some site. Namely, low coverage increases the significance of, erroneous, evidence of the alternative allele resulting in some false HOM_ALT haploid genotype calls. This in turn increases the alternative allele AC and reduces power for PBT. The difference in DNM calling quality between boys and girls can also be observed, in Figure18.

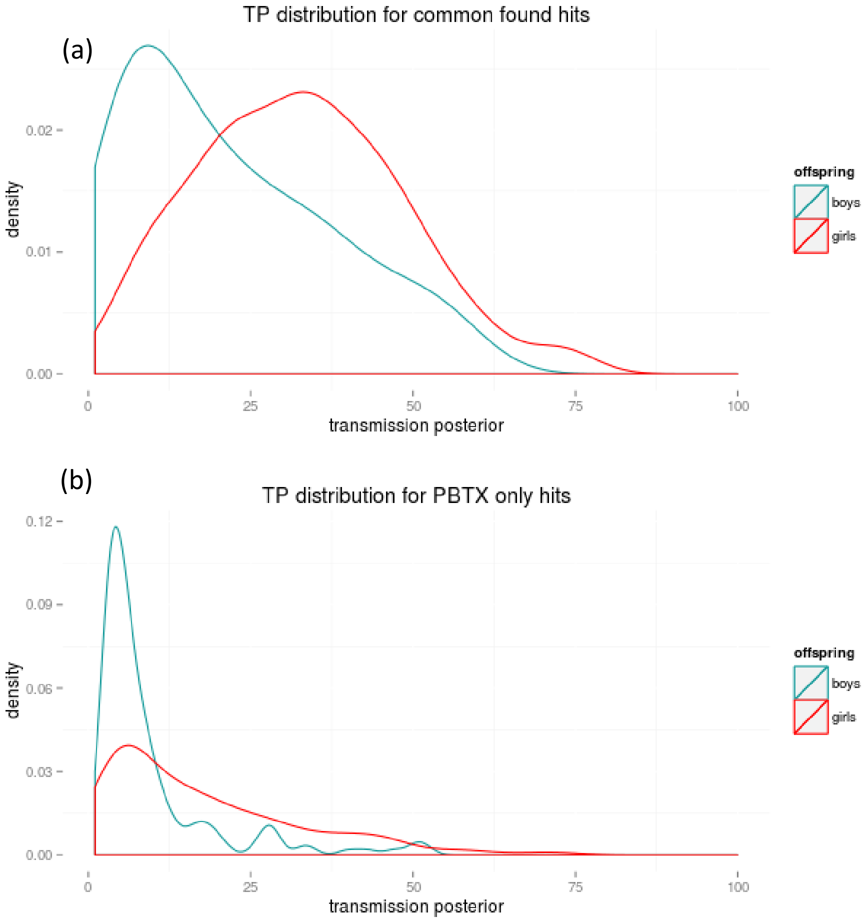

Figure 18: (a) Distribution of the transmission posteriors, as assigned by PBTX to the mutations found by both PBTO and PBTX. (b) Distribution of the assigned transmission posteriors, to the mutations found by PBTX only.

3.1.6 Validation sets

In order to confirm found mutations as well as to evaluate PBT performance under different data characteristics, a set of candidate mutations were selected for validation. Validation will be performed on an Illumina MiSeq machine at a much higher coverage (i.e.: at least 60 overlapping reads per locus) using paired-end reads of 250bp, so as to bypass all the low coverage error modes and decrease alignment problems.

To this end we selected validation sets from the high confidence sets produced by our analysis. Further restricting the output of the, final, AC filter (see Table 5) by the assigned transmission posterior we obtain the following high confidence validation sets:

- 23 DNMs in boy-offspring found by both methods
- 23 DNMs in boy-offspring found by PBTX only
- 164 DNMs in girl-offspring found by both methods
- 107 DNMs in girl-offspring found by PBTX only

These are sets for which we expect a high validation rate. For boys, hits with a TP above 10 were selected and for girls hits with a TP above 20 were selected. We chose a higher TP threshold for girl hits because the quality of the calls is higher in this case, as well as because we can afford being more stringent due to higher number of hits.

Furthermore, some sets homologous to the high confidence ones were generated, for which we have no prior expectation as to how many mutations will validate, but that will help us profile PBT behaviour better. The idea is to build lower confidence sets, that in feature space would be situated around the discrimination border of PBT. They are then built as follows: all initial data cleaning (non-pseudo-autosomal, VQSR) is applied as well as parents' HOM_REF filter. Hits not passing either one of these filters are too unlikely to be validated. The condition regarding evidence of the alternative allele in HOM_REF genotypes is relaxed to allow up to 3 reads containing the alternative allele at that site. This is done because little alternative allele evidence can also be due to simple misalignments. Furthermore, the AC condition is also relaxed to allow mutations at sites where less than 10 alternative alleles have been observed in the population. This accounts to miss-genotyping in other samples, at the position of interest(i.e.: artificially inflated alternative AC) or, excluding miss-genotyped individuals, it would be equivalent to evaluating the power of detecting DNMs that arise at sites where the minor allele frequency is up to 2.5%-5%, as opposed to the stringent filtering of just monomorphic sites. After excluding the hits selected as high confidence, we obtain the following lower confidence validation sets:

- 34 DNMs in boy-offspring found by both methods
- 101 DNMs in boy-offspring found by PBTX only
- 132 DNMs in girl-offspring found by both methods
- 329 DNMs in girl-offspring found by PBTX only

3.1.7 Complete Genomics Data

The parents from 20 random families out of the 250 GoNL families have been sequenced a second time, using Complete Genomics at a coverage of 45x. This enables high quality genotyping that can be compared against DNM calls. The Complete Genomics dataset contains genome-wide calls for these 20 parents. Given that only the parents of the respective families are sequenced at this quality, we cannot use it to validate all genotypes of a putative DNM (from one of these families) but we can find false positives at sites that are polymorphic in the CG data. By validating parents' genotypes found in PBTX hits in the GoNL dataset, against the parents' genotypes produced by the Complete Genomics data, we would only need to further validate the offspring genotype. Given the error modes presented and the adequate filtering for them respectively, we consider validation of parents' genotypes to be a strong indicator of a true underlying mutation.

A number of 2892 PBTX hits were found in all of the families for which Complete Genomics data is available. Considering only the "highly likely" sets of mutations produced by PBTX of 63 boy-offspring and 426 girl-offspring mutations respectively, 23 boy-offspring and 31 girl-offspring mutation sites were found also in Complete Genomics data. For all these 54 mutations, the Complete Genomics data was in concordance with PBTX output, namely, all the parents were HOM_REF in both datasets. Until validation can be performed on the initially selected sets, we consider these results promising

3.1.8 More sensitive initial calls

After filtering done by PBT, namely families with thins and families where only one of the parents is available are dropped, our running input dataset contains 231 families. Out of these, 102 families are of boy-offspring and 129 families have girl-offspring.

Considering previously estimated mutation rates, we expect an average number of 60 de-*novo* mutations in individual's genome. Under the simple assumption of an uniform distribution of mutations across the genome, and the length of our genome and of X Chromosome respectively, we expect an average number of 3 mutations per individual, on the X chromosome.

$$X - linked\ DNMs = \frac{(3 * 10^9 * 60)}{155 * 10^6} \sim 3.0$$

Given that the X chromosome is haploid in males, the number of expected DNMs in boy-offspring is then 1.5. Consequently, we expect a number of approximately 387 true mutations in all of the 129 girl-offspring and a number of 153 true mutations in all of the 101 boy-offspring. The set of 426 "highly likely" mutations found in girls suggests good sensitivity, whereas the 63 mutations found in boys indicate insufficient detection power. Somewhat smaller detection power is expected because of lower quality data in male individuals due to coverage.

We investigated whether additional power can be gained by calling males as haploid using UG rather than adjusting their genotypes and PL values using HaploidWriter (HW). To this end all X chromosome sites were called again, using the haploid model of UG and only the males as calling set. Wherever a variant was thus found in the male population, the rest of the dataset (i.e.: all the females in all 250 families) was genotyped at that position as well and the two outputs were merged into one VCF file. We ran PBTX on this VCF using the same 10^{-4} mutation prior as well as the AF prior and compared results with the previous PBTX run that used the same parameters. After identical filtering steps (see above), we obtain a set of 96 "highly likely" DNM hits in boy-offspring, that include all the 63 such hits found by the previous run of PBTX, thus roughly a 50% increase in sensitivity.

Further investigating the 33 promising mutation hits that were not previously detected by PBTX, we discover that these sites were not found at all as variants by the initial diploid model UG genotyper. The HaploidWriter is used to adjust haploid genotypes that were initially called as diploid and make them suitable for downstream analysis, such as PBTX. Using the haploid model to do the original genotyping proves, as expected, more powerful in detecting haploid variants in both lower alignment data quality and lower allele frequencies (in this case singletons).

The 96 DNM hits in boys are still at a mere 62% of the expected value of 153. We conclude that in order to further increase sensitivity, we require higher coverage. We note however, that the estimation of the expected value of mutation in boys, on the X chromosome, is an upper bound. The actual expected value is lower, if we take into account the fact that most germline mutations come from the parental germline.

Given that boys do not inherit any genetic code from the father's X chromosome, the expected number of X-linked mutations will be less than half of those in girls. The actual proportion of germline mutations inherited from the father is not yet defined within a reasonable confidence interval, so we will not attempt to correct for it, for the purpose of this project.

3.1.9 Father Age Effect

It was proven in literature, that a statistically significant association exists between the number of offspring mutations and father's age at conception, along with estimations of the magnitude of this effect. While attempting to reproduce the magnitude of this effect is beyond the purpose of this project, finding a significant association in our results adds robustness.

To this end, we fit a simple linear regression model, where the target variable is the number of mutations found in the female offspring families that were used by PBTX. We use father's age as describing variable and average coverage over the whole trio as a covariate, known to influence detection. Figure19 illustrates the curve defined by father's age coefficient of the resulting model.

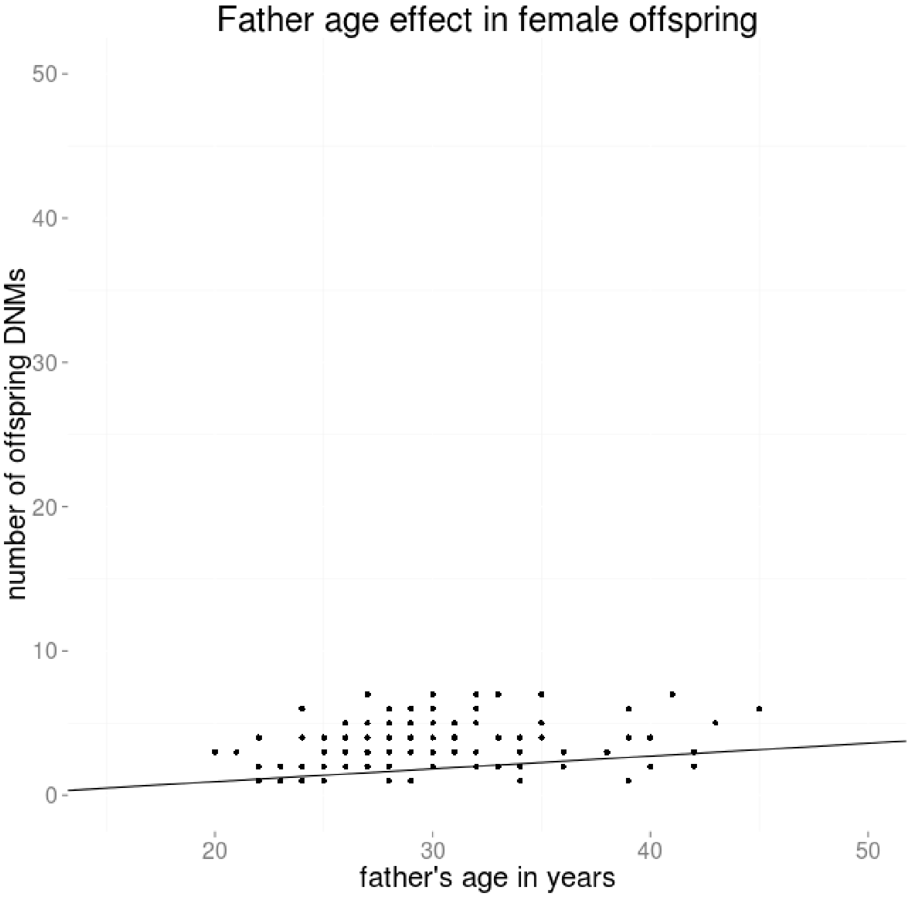

Figure 19 : Father's age effect on germline mutation rate, within the set of high confidence mutations. Each dot represents a female offspring trio. The line is the linear regression approximation when the average trio coverage parameter is fixed to 0.

While the resulting model explains quite little of the variation in number of mutations, the association with father's age was found to be significant, with a p value of 0.001 although it explains only 0.07% of the variation in number of mutations. Training the same model on the set of male offspring mutations, we find no statistically significant association. This is expected, as boys do not inherit anything from the father, on the X chromosome.

3.2 HaploidWriter

Running PhaseByTransmission when alignment data is not available requires the extra step of transforming initially called diploid genotypes to haploid. We evaluate the performance of our HaploidWriter, in order to assess the whether this step influences PBT performance and/or power. To this end we selected one random trio from the GoNL dataset, comprised of father, mother and a male offspring.

We used UG to call all positions on the X chromosome, for all three individuals, using the available alignment data. We ran HaploidWriter on the resulting vcf file to transform all diploid genotypes of the father and the child to the haploid version. The trio alignment data was again used to call all individuals using the haploid model of UG. We compared phred scaled likelihoods of the father and child's genotypes as produced by haploid UG and HaploidWriter separately, using haploid UG as truth values. Whether the two tools produced the same genotype, in terms of alleles, can be inferred directly from the PLs, but no such differences were found. We selected bi-allelic sites only, for simplicity and compared the PL value corresponding to the genotype that was not called (i.e.: the PL of the called genotype is 0). Over the aprox 237000 haploid genotypes found, we observed very high correlation (Figure 20) and an average difference in PL values of 6.5, with a variance of 28. Given the phred quality transformation, a difference of 6.5 corresponds to less than one order of magnitude. The variance corresponds to $\sim$3 orders of magnitude in probability space and could influence the outcome of analysis performed.

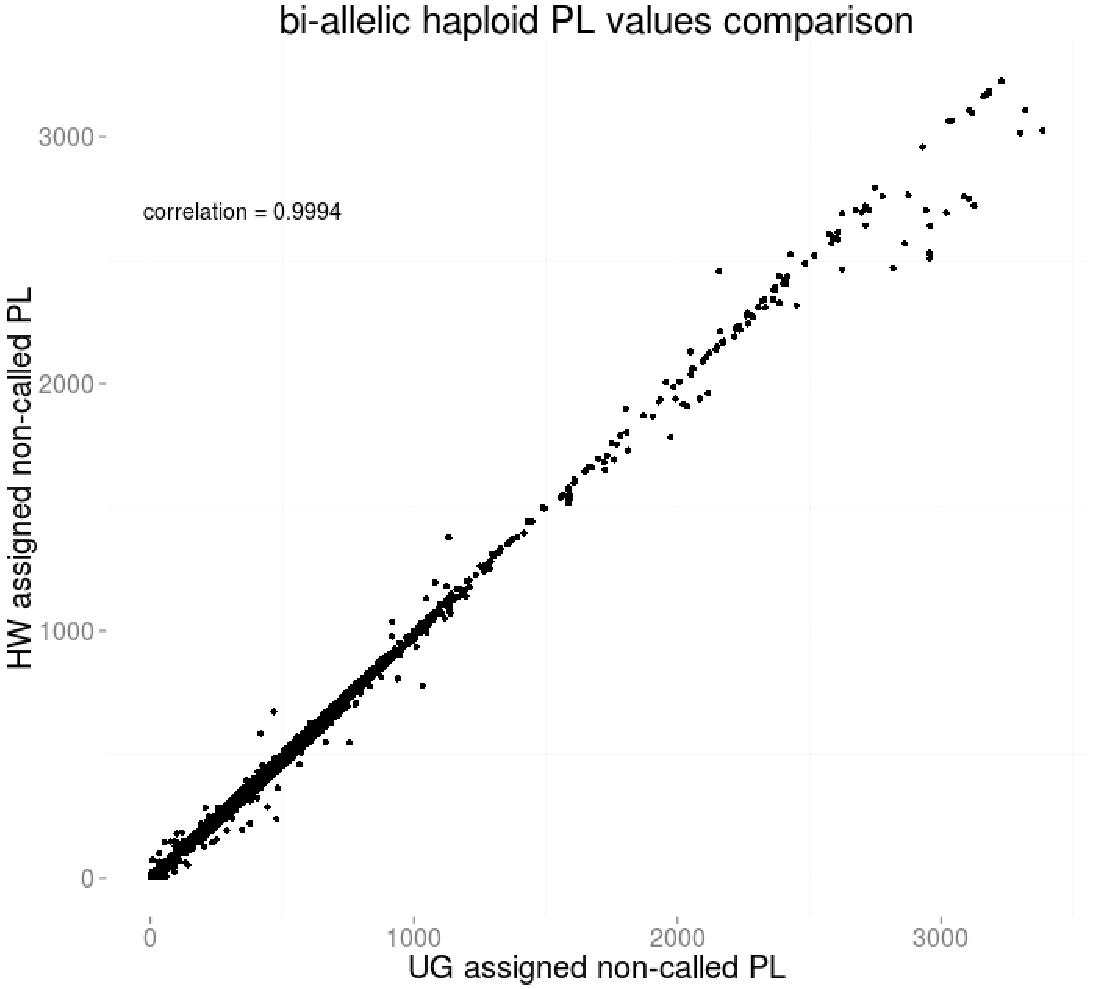

Figure 20 : Haploid assigned PL values versus haploid-UG assigned PL values for the whole X chromosome of all male individuals (father and son) in an arbitrary GoNL family.

By observing the distribution of genotype PL values, on the whole X chromosome of the same individuals, and the binned mean of differences between the two methods' assignments (Figure 21), we note that the vast majority of PL values assigned are below 1000(i.e.: probability of misgenotype

$= 10^{-100}$). Larger values are extreme data cases (extreme points in sequencing coverage), or artificially inflated due, for example to systematic misalignments, resulting in false high coverage and consequently falsely confident calls. These points are likely to be excluded from analyses through various filters (i.e.: VQSR). For the PL values below 1000, we observe that the quality of the HaploidWriter transformation decreases as the likelihood value increases. For likelihoods < 100 the average error is 2 (i.e.: $10^{-0.2}$ in probability space), thus insignificant. For the last bin, $900 < PL < 1000$, the average error increases to 20. We argue that, for all practical purposes, a deviation of the probability of error within two orders of magnitude from a true error probability in the order of 10^{-100} is also neglectable.

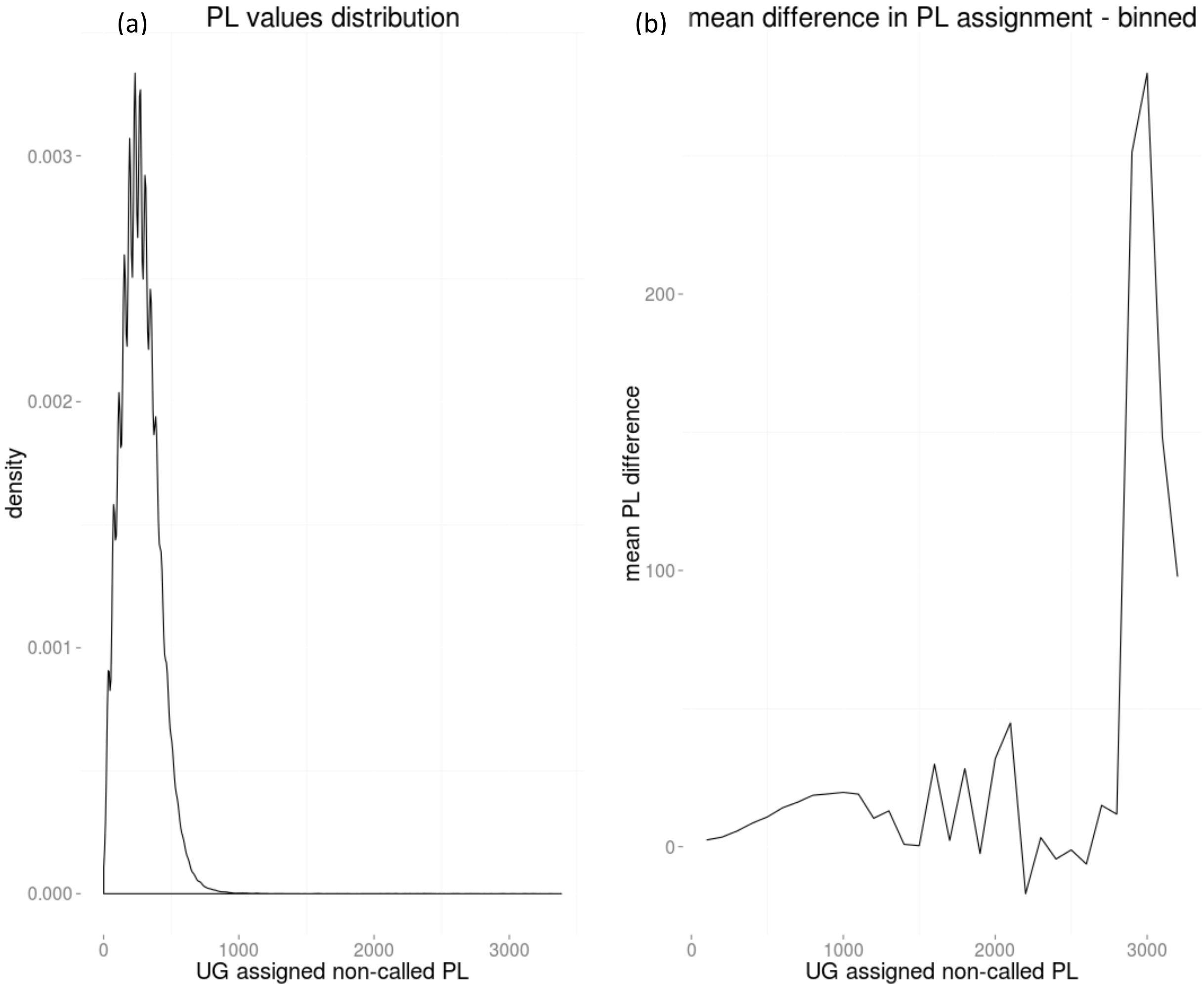

Figure 21 : (a): distribution of PL values as assigned by UG; considered "true" underlying PL distribution. (b): mean absolute difference between HW assigned PLs and UG assigned PLs. The UG assigned PL axis is binned using a 10 units per bin

We tested the influence of HaploidWriter on an analysis by looking at the final set of high confidence mutations found in boys (as described in previous section's $More\ sensitive\ initial\ calls$). We thus have 60 DNM hits which were called using both haploid UG and HaploidWriter separately. We considered all the male genotypes, of the offspring and of the fathers. Comparing the assigned PL values (Figure 22), we note an average absolute error of 23 units (i.e.: HaploidWriter w.r.t. UG). The difference in PBTX assigned transmission posterior for the respective DNM hits was by contrast only 0.45 (i.e.: error of $10^{-0.045}$). The average PL transformation error of 23 is higher than the expected < 16, as computed from the averages in the corresponding bins, mainly because of the very small number of genotypes used, 120, compared to the ~ 233000 that were found in the whole X comparison.

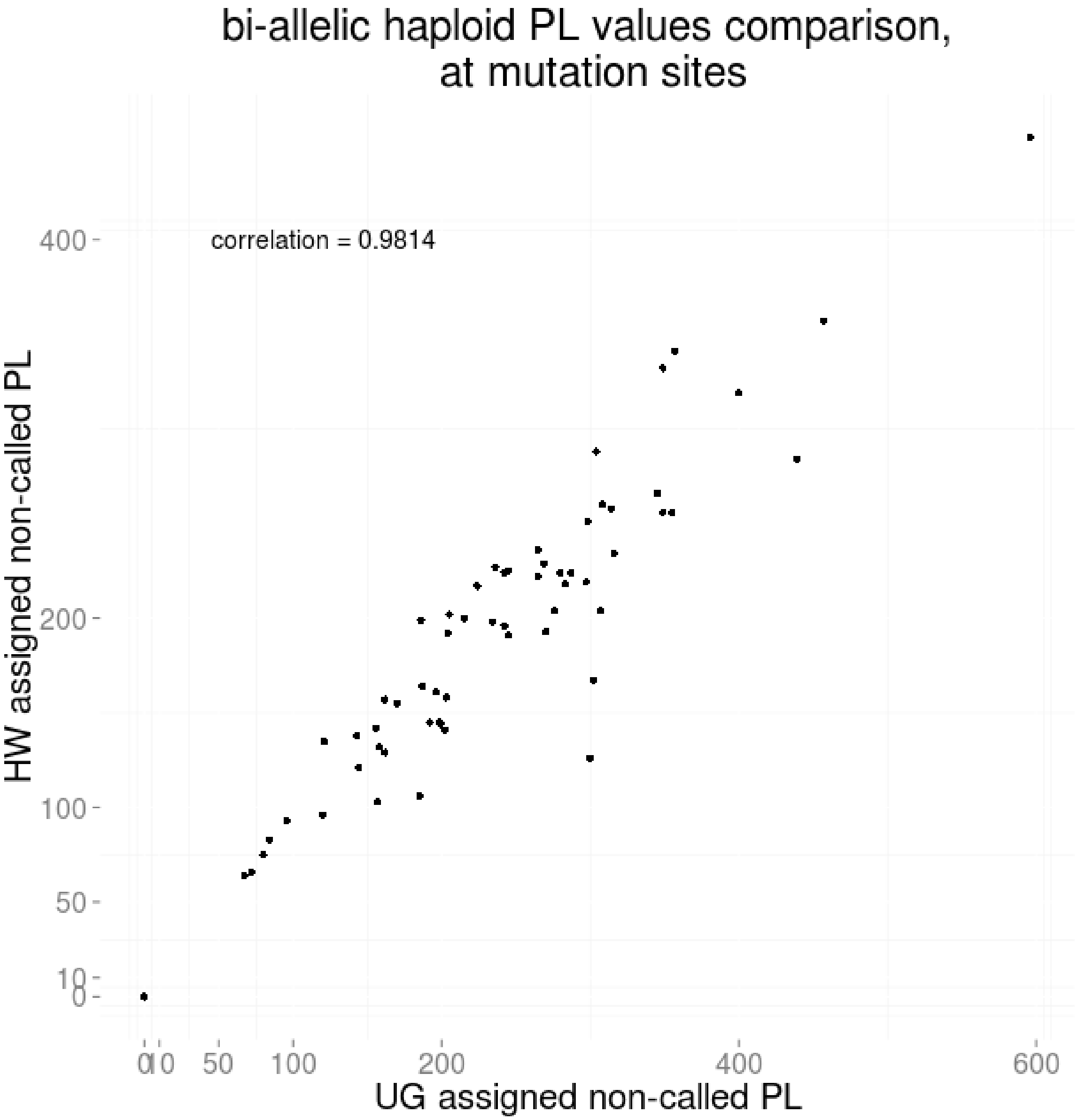

Figure 22 : HW assigned PL values versus UG assigned PL values from the high confidence set of 60 DNM hits found in male offspring. Both offspring and father PL assignments are plotted

3.3 De-NovoMutationPowerCaller

In order to test the usability of our developed tool, we ran the DNMPC, using the trained logistic regression model, on the sequencing data of a number of available trios from the GoNL dataset. Outputting the regression value for each genomic location is impractical considering the input size (10^9), whereas outputting the class label (i.e.: 0 or 1), by thresholding the regression value, does not offer sufficient sensitivity in results. We therefore employ a simple binning of the regression output, namely 10 uniformly distributed bins (i.e.: $[0,0.1)$, $[0.1,0.2) \ldots 0.9,1.0)$. This enables us to select genomic locations down to the desired level of confidence, w.r.t. DNM discovery power. Furthermore, we expect that DNM detection power varies relatively smoothly across the genome. Thus by storing intervals of adjacent positions that fall in the same bin, we expect a significant reduction in the size of the output.

We selected, for further analysis, the results on chromosome 1 of an arbitrary GoNL trio. Given that am individual chromosome is a relatively well defined structural unit of our DNA and that it theoretically respects all the statistical properties that are valid for the whole of our genome, we consider chromosome 1 (i.e.: the largest) to be a good, scaled, representation, for the purpose of this analysis, of the whole genome.

Chromosome 1 has a reference sequence of 248,956,422 base-pairs, i.e.: $\sim 10^8$. The DNMPC , using the described binning, produced a sequence of $\sim 28{,}000{,}000$ intervals of the different confidence levels corresponding to DNM discovery power. This reduces the representation of chr.1 by one order of magnitude. Using the bin values to make a prediction, we observe a class distribution similar to the one observed in the simulated data set with 78% true positives and 22% false negatives respectively. Analyzing the distribution of the produced intervals however, we observe an average interval length of 18 base-pairs, with a variance of 7. Considering the median interval length of 7 base-pairs and the maximum interval length of 1094 base-pairs, we infer that the mapping of the genetic code to DNMPC confidence bins produced is very jittered.

3.3.1 Adjusting the bins

The expected pattern of bins that the DNMPC would produce, consists of some large, high confidence intervals (i.e.: corresponding to non GC-rich regions), with a noisy region of intervals in between them. The overall very noisy intervals produced by the initial run indicate that the original, straightforward, binning should be improved. We therefore inspect the distribution of the data points into bins, as produced by our model on a representative test set (i.e.: that conserves the false negative to true positive ratio) from our simulated data points. Figure 23 shows the allocation performance within each bin (a) and the distribution of correctly assigned data points across the bins (b). We notice the large per-bin performance discrepancy (Figure 23a) between the lower 5 bins (corresponding to a false negative classification) and the upper 5 bins (corresponding to a true positive classification). This is due to the uneven class distribution; i.e.: the $\sim 20\%$ TPs that are misclassified reduce the performance of the lower bins much more than the $\sim 20\%$ FNs that are misclassified. From the information in these plots, the following bins were defined:

- $0, 0.1)$: evaluates to a FN; lowest DNM discovery power; highest performance bin in identifying genomic regions with high false negative rates
- $0.1, 0.3)$: evaluates to a FN; low DNM discovery power; the greater proportion of the FNs evaluate to this bin (i.e.: if we are interested in eliminating most of the genomic regions where we are likely to not discover DNMs, this bin should be considered) as seen in Figure 23b
- $0.3, 0.5)$: lowest performing bin that is evaluated as a false negative; although evaluates to a FN, many TPs evaluate to this bin
- $0.5, 0.7)$: evaluates to a TP; suggests good DNM detection power
- $0.7, 1.0)$: evaluates to a TP; highest confidence bin; performance $\sim 100\%$

Running the DNMPC again, on the chromosome 1 sequencing data of the same family, we obtain $\sim 9{,}000{,}000$ intervals, one third of the initial number of intervals. Furthermore, the average interval length is 102 base-pairs, with a variance of 30, and the largest interval is $> 50{,}000$ base pairs long, with more than 400 intervals of length $> 7{,}000$. This is much closer to our expectation of some long, high confidence, intervals with smaller, varying confidence intervals in between. Considering the very high bin performance of all the bins that evaluate to true positive (Figure 23a), the $0.5, 0.7)$ and $0.7, 1.0$ bins can be further merged together, thus obtaining longer high confidence intervals.

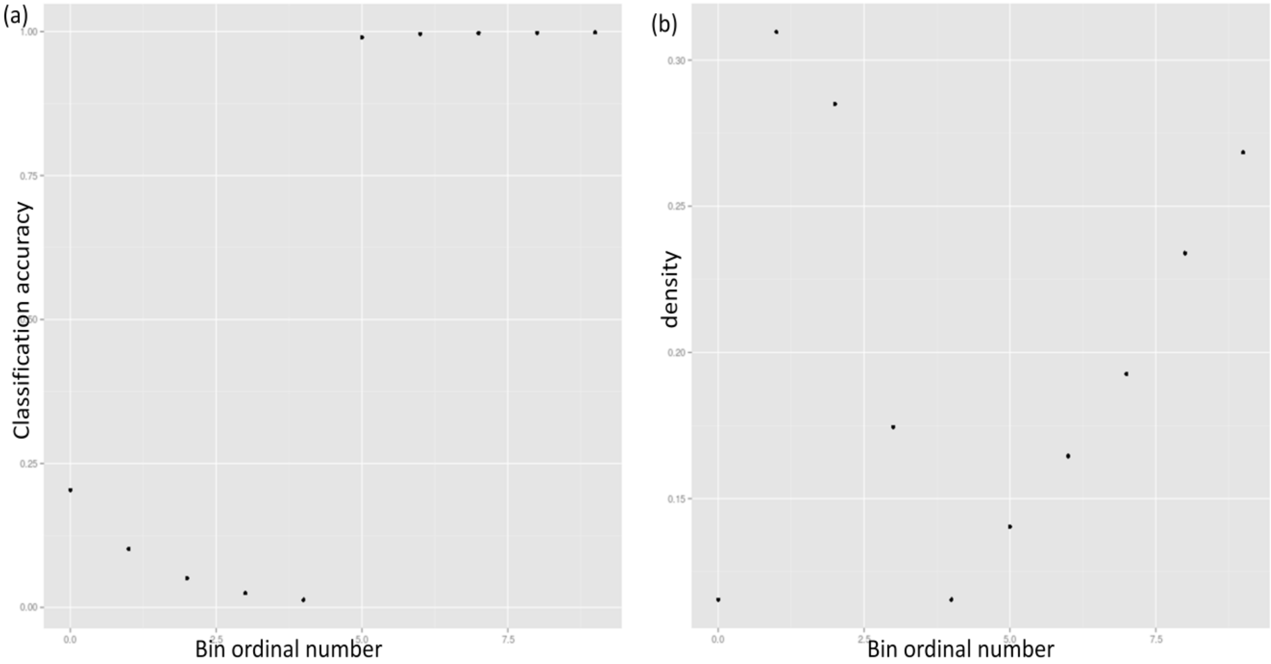

Figure 23: (a) Performance of each bin as initially defined (10 uniform bins in the $[0, 1)$ interval) Performance is computed as: #correctly classified points that evaluate to this bin/#points that evaluate to this bin. (b) Distribution of correctly classified points over bins. There are 2 effective distributions in the plot: one for bins 1 through 5 for false negatives and one for bins 6 through 10 for true positives. For each bin we computed: #correctly classified points that evaluate to the bin/ #points belonging to the class that the bin evaluates to

4. Conclusions

In this thesis we have successfully modified the existing autosomal PhaseByTransmission (PBT) model, to correctly identify de-*novo* mutations that arise on the X chromosome. The analysis is comprised of a lengthy, modular pipeline, where data uncertainty from the underlying technology (i.e.: sequencing), as well as from intermediate computational steps (i.e.: alignment) is accumulated and propagated. We therefore presented a comprehensive PBT model that correctly encapsulates all of the available information, from sequencing data to Mendelian inheritance patterns, population wide information (i.e.: allelic frequency) and prior domain knowledge about the de-*novo* events under consideration. Combined, they create the framework for statistically robust findings. Furthermore, PBT is a straightforward model that introduces minimum amount of computational overhead, which is vital in the context of very large amounts of data such as genomic data. PBT can be straightforwardly further extended with respect to scope (i.e.: multi-allelic sites) and/or detection power (i.e.: consider larger pedigrees). These extensions however, generate a polynomial (i.e.: quadratic) and/or exponential time complexity increase respectively, thus further analysis is necessary in order to make the best decision.

We further showed how the statistical model cannot fully account for all the variation or error modes present in the data and how post-filtering, using domain knowledge about common error modes, can be just as important in obtaining a final set of sensible and robust results. We tested the results against proven characteristics (i.e.: father age effect) and showed that they hold.

The choice of what point in the analysis pipeline one designs a tool for is crucial to efficiency and we illustrated that we can design a discovery tool such as PBT at the highest level (i.e.: VCF file level) while demonstrating good results. Furthermore, we showed how to make PBT accessible to use for other parties, by providing the additional tools needed (i.e.: HaploidWriter) that correctly transform the commonly available data into sensible PBT input. We furthermore evaluated the precision of our HaploidWriter transformation and showed its limitation/effect on final PBT sensitivity, by contrasting it with a more robust pipeline that can only be applied however, when lower level data is available (i.e.: alignment data).

Finally, we presented a tool, the De-*Novo*MutationPowerCaller, that attempts to profile PBT performance across different genomic regions, with different properties. We presented the setbacks of such an endeavour, due to the lack of proper, real, training data, and presented a robust way of generating synthetic data to train our model. We identified which features influence PBT de-*novo* discovery and built a simple and fast model that can profile the entire genome w.r.t. de-*novo* mutations discovery, by distinguishing between different confidence tranches. We further showed how to optimize these tranches so as to reduce dimensionality of the output, with minimal loss in precision.

References

[1] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P. J. R. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J.-B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, D. Hadley, M. Harris, P. Harrison, S. Brady, A. Hicks, E. Holloway, L. Hui, S. Hussain, C. Louis-Dit-Sully, J. Ma, A. MacGilvery, C. Mader, A. Maratukulam, T. C. Matise, K. B. McKusick, J. Morissette, A. Mungall, D. Muselet, H. C. Nusbaum, D. C. Page, A. Peck, S. Perkins, M. Piercy, F. Qin, J. Quackenbush, S. Ranby, T. Reif, S. Rozen, C. Sanders, X. She, J. Silva, D. K. Slonim, C. Soderlund, W.-L. Sun, P. Tabar, T. Thangarajah, N. Vega-Czarny, D. Vollrath, S. Voyticky, T. Wilmer, X. Wu, M. D. Adams, C. Auffray, N. a. R. Walter, R. Brandon, A. Dehejia, P. N. Goodfellow, R. Houlgatte, J. R. Hudson, S. E. Ide, K. R. Iorio, W. Y. Lee, N. Seki, T. Nagase, K. Ishikawa, N. Nomura, C. Phillips, M. H. Polymeropoulos, M. Sandusky, K. Schmitt, R. Berry, K. Swanson, R. Torres, J. C. Venter, J. M. Sikela, J. S. Beckmann, J. Weissenbach, R. M. Myers, D. R. Cox, M. R. James, D. Bentley, P. Deloukas, E. S. Lander, and T. J. Hudson, "A Gene Map of the Human Genome," *Science*, vol. 274, no. 5287, pp. 540–546, Oct. 1996.

[2] Z. D. Zhang, A. Frankish, T. Hunt, J. Harrow, and M. Gerstein, "Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates," *Genome Biol.*, vol. 11, no. 3, p. R26, 2010.

[3] D. Zheng, Z. Zhang, P. M. Harrison, J. Karro, N. Carriero, and M. Gerstein, "Integrated pseudogene annotation for human chromosome 22: evidence for transcription," *J. Mol. Biol.*, vol. 349, no. 1, pp. 27–45, May 2005.

[4] P. P. Dennis and R. F. Young, "Regulation of ribosomal protein synthesis in Escherichia coli B/r.," *J. Bacteriol.*, vol. 121, no. 3, pp. 994–999, Mar. 1975.

[5] P. Czeleń and P. Cysewski, "Structural and energetic properties of canonical and oxidized telomeric complexes studied by molecular dynamics simulations," *J. Mol. Model.*, vol. 19, no. 8, pp. 3339–3349, Aug. 2013.

[6] A. Stark and E. Seneta, "Wilhelm Weinberg's early contribution to segregation analysis," *Genetics*, vol. 195, no. 1, pp. 1–6, Sep. 2013.

[7] A. Stark and E. Seneta, "On S.N. Bernstein's derivation of Mendel's Law and 'rediscovery' of the Hardy-Weinberg distribution," *Genet. Mol. Biol.*, vol. 35, no. 2, pp. 388–394, Apr. 2012.

[8] O. Mayo, "A century of Hardy-Weinberg equilibrium," *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.*, vol. 11, no. 3, pp. 249–256, Jun. 2008.

[9] A. Eyre-Walker and P. D. Keightley, "The distribution of fitness effects of new mutations," *Nat. Rev. Genet.*, vol. 8, no. 8, pp. 610–618, Aug. 2007.

[10] J. F. Crow, "The origins, patterns and implications of human spontaneous mutation," *Nat. Rev. Genet.*, vol. 1, no. 1, pp. 40–47, Oct. 2000.

[11] A. S. Kondrashov, "Measuring spontaneous deleterious mutation process," *Genetica*, vol. 102–103, no. 1–6, pp. 183–197, 1998.

[12] A. Caballero, "Estimation of the upper limit of the mutation rate and mean heterozygous effect of deleterious mutations," *Genet. Res.*, vol. 88, no. 3, pp. 137–141, Dec. 2006.

[13] H. T. Wibisono and W. Pusparini, "Sumatran tiger (Panthera tigris sumatrae): a review of conservation status," *Integr. Zool.*, vol. 5, no. 4, pp. 313–323, Dec. 2010.

[14] Z. Zhang, W. Dai, Y. Wang, C. Lu, and H. Fan, "Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1)," *Arch. Virol.*, vol. 158, no. 1, pp. 145–154, Jan. 2013.

[15] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn, "Genetic signatures of strong recent positive selection at the lactase gene," *Am. J. Hum. Genet.*, vol. 74, no. 6, pp. 1111–1120, Jun. 2004.

[16] K. J. Guinan, R. T. Cunningham, A. Meenagh, A. Gonzalez, M. M. Dring, B. W. McGuinness, D. Middleton, and C. M. Gardiner, "Signatures of natural selection and coevolution between killer cell immunoglobulin-like receptors (KIR) and HLA class I genes," *Genes Immun.*, vol. 11, no. 6, pp. 467–478, Sep. 2010.

[17] E. C. Castelli, C. T. Mendes-Junior, L. C. Veiga-Castelli, M. Roger, P. Moreau, and E. A. Donadi, "A comprehensive study of polymorphic sites along the HLA-G gene: implication for gene regulation and evolution," *Mol. Biol. Evol.*, vol. 28, no. 11, pp. 3069–3086, Nov. 2011.

[18] M. Kimura and T. Ohta, "The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population," *Genetics*, vol. 61, no. 3, pp. 763–771, Mar. 1969.

[19] P. Rahman, A. Jones, J. Curtis, S. Bartlett, L. Peddle, B. A. Fernandez, and N. B. Freimer, "The Newfoundland population: a unique resource for genetic investigation of complex diseases," *Hum. Mol. Genet.*, vol. 12 Spec No 2, pp. R167–172, Oct. 2003.

[20] C. Moreau, J.-F. Lefebvre, M. Jomphe, C. Bhérer, A. Ruiz-Linares, H. Vézina, M.-H. Roy-Gagnon, and D. Labuda, "Native American admixture in the Quebec founder population," *PloS One*, vol. 8, no. 6, p. e65507, 2013.

[21] N. Jäger, M. Schlesner, D. T. W. Jones, S. Raffel, J.-P. Mallm, K. M. Junge, D. Weichenhan, T. Bauer, N. Ishaque, M. Kool, P. A. Northcott, A. Korshunov, R. M. Drews, J. Koster, R. Versteeg, J. Richter, M. Hummel, S. C. Mack, M. D. Taylor, H. Witt, B. Swartman, D. Schulte-Bockholt, M. Sultan, M.-L. Yaspo, H. Lehrach, B. Hutter, B. Brors, S. Wolf, C. Plass, R. Siebert, A. Trumpp, K. Rippe, I. Lehmann, P. Lichter, S. M. Pfister, and R. Eils, "Hypermutation of the inactive X chromosome is a frequent event in cancer," *Cell*, vol. 155, no. 3, pp. 567–581, Oct. 2013.

[22] J. Taylor, S. Tyekucheva, M. Zody, F. Chiaromonte, and K. D. Makova, "Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison," *Mol. Biol. Evol.*, vol. 23, no. 3, pp. 565–573, Mar. 2006.

[23] D. F. Conrad, J. E. M. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, P. Awadalla, and 1000 Genomes Project, "Variation in genome-wide mutation rates within and between human families," *Nat. Genet.*, vol. 43, no. 7, pp. 712–714, Jul. 2011.

[24] M. W. Nachman and S. L. Crowell, "Estimate of the Mutation Rate per Nucleotide in Humans," *Genetics*, vol. 156, no. 1, pp. 297–304, Sep. 2000.

[25] J. B. S. HALDANE, "The mutation rate of the gene for haemophilia, and its segregation ratios in males and females," *Ann. Eugen.*, vol. 13, no. 4, pp. 262–271, Jun. 1947.

[26] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease," *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 565–575, Aug. 2012.

[27] P. Awadalla, J. Gauthier, R. A. Myers, F. Casals, F. F. Hamdan, A. R. Griffing, M. Côté, E. Henrion, D. Spiegelman, J. Tarabeux, A. Piton, Y. Yang, A. Boyko, C. Bustamante, L. Xiong, J. L. Rapoport, A. M. Addington, J. L. E. DeLisi, M.-O. Krebs, R. Joober, B. Millet, E. Fombonne, L. Mottron, M. Zilversmit, J. Keebler, H. Daoud, C. Marineau, M.-H. Roy-Gagnon, M.-P. Dubé, A. Eyre-Walker, P. Drapeau, E. A. Stone, R. G. Lafrenière, and G. A. Rouleau, "Direct measure of the de novo mutation rate in autism and schizophrenia cohorts," *Am. J. Hum. Genet.*, vol. 87, no. 3, pp. 316–324, Sep. 2010.

[28] A. Hoischen, B. W. M. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, K. Devriendt, M. Z. Amorim, N. Revencu, A. Kidd, M. Barbosa, A. Turner, J. Smith, C. Oley, A. Henderson, I. M. Hayes, E. M. Thompson, H. G. Brunner, B. B. A. de Vries, and J. A. Veltman, "De novo mutations of SETBP1 cause Schinzel-Giedion syndrome," *Nat. Genet.*, vol. 42, no. 6, pp. 483–485, Jun. 2010.

[29] A. Hoischen, B. W. M. van Bon, B. Rodríguez-Santiago, C. Gilissen, L. E. L. M. Vissers, P. de Vries, I. Janssen, B. van Lier, R. Hastings, S. F. Smithson, R. Newbury-Ecob, S. Kjaergaard, J. Goodship, R. McGowan, D. Bartholdi, A. Rauch, M. Peippo, J. M. Cobben, D. Wieczorek, G. Gillessen-Kaesbach, J. A. Veltman, H. G. Brunner, and B. B. B. A. de Vries, “De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome,” *Nat. Genet.*, vol. 43, no. 8, pp. 729–731, Aug. 2011.

[30] M. P. Adam and L. Hudgins, “Kabuki syndrome: a review,” *Clin. Genet.*, vol. 67, no. 3, pp. 209–219, Mar. 2005.

[31] A. Sirmaci, M. Spiliopoulos, F. Brancati, E. Powell, D. Duman, A. Abrams, G. Bademci, E. Agolini, S. Guo, B. Konuk, A. Kavaz, S. Blanton, M. C. Digilio, B. Dallapiccola, J. Young, S. Zuchner, and M. Tekin, “Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia,” *Am. J. Hum. Genet.*, vol. 89, no. 2, pp. 289–294, Aug. 2011.

[32] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, “Common SNPs explain a large proportion of the heritability for human height,” *Nat. Genet.*, vol. 42, no. 7, pp. 565–569, Jul. 2010.

[33] H.-H. Ropers, “X-linked mental retardation: many genes for a complex disorder,” *Curr. Opin. Genet. Dev.*, vol. 16, no. 3, pp. 260–269, Jun. 2006.

[34] L. E. L. M. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, B. W. M. van Bon, A. Hoischen, B. B. A. de Vries, H. G. Brunner, and J. A. Veltman, “A de novo paradigm for mental retardation,” *Nat. Genet.*, vol. 42, no. 12, pp. 1109–1112, Dec. 2010.

[35] M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, and M. C. O’Donovan, “De novo mutations in schizophrenia implicate synaptic networks,” *Nature*, vol. advance online publication, Jan. 2014.

[36] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, “Large-scale and automated DNA sequence determination,” *Science*, vol. 254, no. 5028, pp. 59–67, Oct. 1991.

[37] H. Swerdlow, S. L. Wu, H. Harke, and N. J. Dovichi, “Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette,” *J. Chromatogr.*, vol. 516, no. 1, pp. 61–67, Sep. 1990.

[38] R. G. Blazej, P. Kumaresan, and R. A. Mathies, “Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 19, pp. 7240–7245, May 2006.

[39] J. Shendure and H. Ji, “Next-generation DNA sequencing,” *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1135–1145, Oct. 2008.

[40] M. L. Metzker, “Sequencing technologies - the next generation,” *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, Jan. 2010.

[41] M. J. Chaisson, D. Brinza, and P. A. Pevzner, “De novo fragment assembly with short mate-paired reads: Does the read length matter?,” *Genome Res.*, vol. 19, no. 2, pp. 336–346, Feb. 2009.

[42] B. Ewing and P. Green, “Base-calling of automated sequencer traces using phred. II. Error probabilities,” *Genome Res.*, vol. 8, no. 3, pp. 186–194, Mar. 1998.

[43] “GATK Guidebook 2.7-4.” .

[44] W.-P. Lee, M. Stromberg, A. Ward, C. Stewart, E. Garrison, and G. T. Marth, “MOSAIK: A hash-based algorithm for accurate next-generation sequencing read mapping,” *ArXiv E-Prints*, vol. 1309, p. 1149, Sep. 2013.

[45] H. Li, J. Ruan, and R. Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores,” *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, Nov. 2008.

[46] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang, “SOAP2: an improved ultrafast tool for short read alignment,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 15, pp. 1966–1967, Aug. 2009.

[47] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[48] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009.

[49] H. Li, “Mathematical Notes on SAMtools Algorithms,” 12-Oct-2010. [Online]. Available: http://lh3lh3.users.sourceforge.net/download/samtools.pdf. [Accessed: 24-Jan-2014].

[50] J. A. Graves, M. J. Wakefield, and R. Toder, “The origin and evolution of the pseudoautosomal regions of human sex chromosomes,” *Hum. Mol. Genet.*, vol. 7, no. 13, pp. 1991–1996, Dec. 1998.

[51] A. Helena Mangs and B. J. Morris, “The Human Pseudoautosomal Region (PAR): Origin, Function and Future,” *Curr. Genomics*, vol. 8, no. 2, pp. 129–136, Apr. 2007.