Introduction
Pathogenic DNA variants can cause severe diseases that have a serious impact on both patients and their close relatives. Proper identification of pathogenic variants in a person’s DNA can reduce health issues arising from those diseases as this knowledge can steer the right treatment to improve a patient’s quality of life. Even if no treatment is available (or necessary), knowing which DNA variants one carries is important information as it confirms (or even determines) the clinical diagnosis. Knowing the odds of those variants affecting a possible future child provides information for decisions on whether or not to conceive children. Even during pregnancy, knowledge of which DNA abnormalities a fetus carries can be extremely helpful. Prospective parents can start making necessary preparations for future care, or make an informed decision on whether or not to continue the pregnancy.

The ultimate goal of this thesis is to develop and provide bioinformatics approaches to identify genetic variations in patients, in particular variations that affect the number of copies of a DNA sequence, so called copy number variations.

Genetic variations

Deoxyribonucleic acid (DNA) is a double-stranded molecule that encodes for the properties of a cell through specific sequences of base pairs. Any change in the sequence of DNA (a variant or mutation) may directly influence the way a cell functions, and can be passed on to next generations if they occur in the germ-line. Most variants are neutral, or have only limited implications. The occurrence of new variants is regarded as the driving force behind evolution. Changed genomes are tested against the environment, and those leading to fit organisms are most likely to reproduce.

Changes to DNA sequences can be classified in three main groups: single nucleotide variations (SNVs) [1], where only a single base pair is affected, copy number variations (CNVs) [2], where a longer stretch of a DNA sequence is either multiplied or deleted from the original sequence, and translocations and inversions, where genetic sequences are ordered differently [3, 4]. The work in this thesis focuses on the detection of CNVs.

Figure 1: Examples of sub-chromosomal copy number variations. Left: a deletion, the right side of the left arm of the chromosome is missing in the chromosome below with respect to the chromosome on the top. Right: the bottom chromosome now has an additional copy of a region, as shown by the marked areas on both chromosomes. [5]
Copy number variations

A copy number variation can vary in size from a few, up to several millions of base pairs. CNVs are deletions or amplifications, in which one or more copies of a genetic sequence are either removed or added, respectively (Figure 1). As only 2% of the DNA encodes for genes [6], and the DNA sequence is stored on two separate chromosomes, these changes do not always lead to phenotypic differences between human beings. However, losing a gene is likely to change development of cells in one way or another, possibly resulting in a disorder, although not all gene deletions are by definition pathogenic [7]. Alternatively, duplications may result in overproduction of specific proteins, which in turn can cause disorders as well [8]. Depending on the biological role of the genes involved, the age of onset of CNV associated diseases can vary from embryonal to late-onset [9, 10].

Aneuploidy

The term aneuploidy refers to a specific type of CNV where the number of chromosomes in a person differs from the expected number of chromosomes. For example, having an extra copy of chromosome 21, which results in Down’s syndrome [12, 13], is an aneuploidy (Figure 2). More specifically, having three copies of a chromosome is referred to as a trisomy, disomy refers to having two copies and lacking a copy is called a monosomy. An example of a monosomy is Turner syndrome, caused by monosomy of chromosome X [14].
Introduction

Figure 3: Illustration of chromosomal mosaicism, showing regions affected by missing a copy of chromosome X (colored salmon) while the other regions are unaffected (colored purple). \[21\]

The complete deletion or duplication of a chromosome is associated with early fetal death, or with distinguishable phenotypes. For example, when cases of trisomy 13 and 18 lead to live birth, the baby is unlikely to survive for more than a few days or years \[15, 16\]. Trisomy 21 does allow survival but comes with distinguishable facial features \[17\], intellectual disability \[18, 19\], and often with other defects such as congenital heart disease \[20\], requiring additional care during the full lifespan of the affected individual.

Somatic mutations

Mutations that are acquired after fertilization are called somatic mutations, as opposed to germline mutations which are passed on through reproduction. These mutations can include any of the previously mentioned types (SNV, CNV and translocation/inversion). The affected DNA is most often limited to a subset of cells in the body, causing mosaicism (Figure 3). Mosaicism is generally characterized by the fact that the genetic change is not necessarily harming the cell in which it occurs, or the carrier in general. For example, a mutation that would affect intellectual development but occurs only in the cells of a patient’s foot will not lead to severe intellectual disabilities. Additionally, if only a very small subset of cells in a tissue is affected, other (healthy) cells can sustain a tissue’s functionality, leading to milder phenotypes than non-mosaic cases of the same mutation.

However, cancer is a form of mosaicism as well, as it is a subset of cells with a different
genetic sequence. Even within a tumor there are often several different variations of the genetic sequence, corresponding to the tumor's development over time and showing the accumulation of new mutations during cell divisions.

Clearly somatic mutations, and mosaicism, can have severe effects on a patient's health and therefore are important for correct diagnosis. Their effects can be less straightforward to predict as the affected tissue is an important factor that may not be easily determined. Additionally, these mutations can be obtained during the lifespan of an individual, such as in the development of a tumor.

**Methods to detect CNVs**

**Karyotyping**

In classic karyotyping methods, a dye (often Giemsa) is applied to a sample of chromosomes. The dye stains AT rich regions visibly darker than GC rich regions. As the distribution of AT and GC content is not uniform over and within chromosomes, this staining gives them a recognizable pattern of lighter and darker bands when put under a microscope (Figure 4). Using this visible banding pattern, large variations can be spotted manually when the pattern deviates from known healthy chromosome banding patterns. Trained cytogeneticists can identify changes down to approximately 10 Mb, including translocations.
Introduction

Figure 5: An exemplar FISH image, showing chromosomes in blue and two region specific sequences colored red and green, respectively. The unusual co-localization of these sequences near the bottom left marks a genomic rearrangement (the spot with red and green dots directly next to each other).

Fluorescence in situ hybridization

Langer-Safer et al. introduced a reliable technique to find large copy number aberrations in 1982, called fluorescence in situ hybridization (FISH) [24]. In FISH, fluorescent probes are designed to hybridize to highly complementary strands of DNA. By targeting a specific region on a chromosome, the amount of copies (and locations) of that specific region can be determined across all chromosomes by observing the resulting signals using fluorescence microscopy (Figure 5). FISH is often applied to confirm suspicions of a known microdeletion or translocation. Its ability to determine (small) balanced translocations is an important advantage of this method over most of the other methods discussed here. Limitations of this method lie mostly in the targeted nature of the approach, as one has to use a specific sequence targeting the suspected aberration. If the targeted sequence is not part of the aberration, there is no way to determine the actual aberration from the same experiment.

Microarray

The microarray technology was first introduced in 1981 [25] but has become popular since 1995 with the introduction of the complementary DNA microarray (cDNA microarray) [26]. The microarray is a reliable and affordable approach for determining CNVs. Probes designed to match specific sequences of DNA across the whole genome are collectively synthesized at probe-specific areas on the microarray. When testing, DNA fragments are treated with a dye that responds to light and the sample is loaded onto the microarray. DNA fragments with a sequence that match a probe can attach to that probe, while DNA fragments that do not match any probes are washed off. Next, every batch is targeted by a laser while recording the light intensity emitted by the targeted batch, providing an indication of the relative amount of DNA fragments that matched the probes (Figure 6).
As this intensity is heavily influenced by uncontrollable environmental factors at the time of testing, as well as by the amount of DNA loaded on the array, it is of vital importance to normalize the data. In the two-channel approach, a second sample is hybridized along with the sample of interest. This second sample is treated with a different colored dye and should have no notable CNVs. Often a mixture of several samples is used to average out possible CNVs. By comparing light intensities between the two samples, probes targeting a genomic region that is affected by a CNV can be determined [28].

More recently, with the increased reliability of microarray technology, the one-channel approach has become the standard. Instead of loading a second sample on a chip and measuring intensities for both samples, a single sample is loaded and analyzed. However, detection of aberrations relies on comparing the measurements to a reference set, and batch effects need to be accounted for [29]. In large scale studies, this approach saves many experiments, and combining data from multiple studies can be easier than in the two-channel approach.

Alternative technologies have been introduced, such as oligonucleotide arrays and SNP arrays [30]. The SNP arrays can be employed to detect CNVs using the ratios of DNA fragments found with alternative alleles in heterozygous SNPs (Figure 7), and can reach higher resolutions [31]. Another benefit of SNP arrays is that they will reveal regions of homozygosity (ROHs).

Due to noise in the light intensity signal, and depending on the number of probes on the array, the resolution of this approach is usually limited to CNVs larger than 15 kilobase. While sensitive enough for diagnosis of many aberrations, whole genes can be disrupted by a CNV far smaller than the 15 kb resolution, making the microarray unfeasible to detect certain short CNVs.
Introduction

Figure 7: Example of CNV detection using a SNP array. A vertical distribution other than three bands, as depicted in the right half of chromosome 1, indicates a different distribution than expected for heterozygous diploidy. Lacking the middle band indicates loss of a (heterozygous) copy of that region, i.e. chromosome 2, while additional bands indicate additional copies of a region, i.e. chromosome 15. This particular sample has multiple regions with a loss of heterozygosity, of which several are marked with arrows.

Next generation sequencing

One of the most notable recent improvements in measuring genome differences is the development of next generation sequencing (NGS) [33]. Briefly, NGS ‘reads’ the DNA sequences of fragmented DNA. The reading process is called sequencing. While various sequencing techniques have been developed, such as 454 (Roche) and SMRT (PacBio) sequencing, Illumina sequencing is currently the most commonly used technology [34]. This technology uses flow cells, small plates made of glass with multiple hollow tubes (lanes) in them. The surfaces of these lanes are filled with probes that allow specific sequences of DNA molecules to attach, which can then be sequenced by means of synthesis.

To sequence a sample, a library of sequence molecules needs to be created. This is achieved by ligating all DNA fragments to adapters, which allow these fragments to stick to the surface of the lanes (Figure 8a).

After samples are put into the lane of a flow cell, attached DNA fragments are amplified into clusters (Figure 8b). This results in many copies of exactly the same fragment. Sequencing works by repeatedly letting a mixture of dyed bases attach to all DNA fragments in a lane. Every base in this mixture carries a dye with a color unique to that type of base, and blocks other bases from attaching to the same fragment (Figure 8c). By exciting a cluster with a laser, a camera can detect the clusters current color, thus determining what type of base was attached to the fragment in that cluster. After reading the current color for all DNA fragments in a lane, a fluid is pushed through the lane to
Figure 8: Schematic overview of Illumina sequencing. Showing (a) the addition of adapters to DNA molecules during library preparation, (b) cluster generation on the flow cell, and (c) the sequencing itself, followed by (d) bioinformatic data analysis. [35]
remove the dyes, allowing new bases to attach in the next cycle.

The amount of times this process is repeated (cycles) determines how many bases are read from all DNA fragments in the sample. The sequences of base pairs obtained are referred to as reads. Using less cycles would result in obtaining shorter reads, independent of the actual length of the DNA fragments in a sample.

Sequencing longer reads increases the amount of errors made per base pair due to imperfections of the method described above, as not all base pairs in a cluster are properly cleared of their dye in every cycle. Rather than sequencing infinitely to determine the very end of a fragment, Illuma resorted to paired-end sequencing. Here, DNA fragments are read for a predefined number of cycles. Then all attached bases are cleared. After that, sequencing restarts, but now from the other end of every fragment. As the DNA fragments are attached to the surface of the lanes, their positions do not change during sequencing. This allows identification of which reads on the second half of paired-end sequencing match to which reads found in the first half of sequencing.

Whole genome and exome sequencing

NGS provides the technology to sequence fragmented DNA. When no selection is performed beforehand, NGS will sequence fragments across the whole genome. This is known as whole genome sequencing (WGS). When sequencing more fragments, at some point, the DNA sequence is sequenced multiple times. This is called coverage; the amount of reads covering an arbitrary position on the genome. Since the average coverage for every position is known, one can determine copy number variations from the number of reads mapped to a position.

In some cases, one is not interested in information from the whole genome, but only in genetically coding regions. Whole exome sequencing (WES) [36] applies a preprocessing step that uses beads to extract DNA fragments with specific sequences before sequencing. These beads are attached to fragments complementary to the sequences in known coding genes. When mixed with the DNA sample, matching DNA fragments will hybridize to the fragments on the beads. By extracting the beads and fragments attached to them, and removing the DNA fragments that did not match the targeted regions (well enough) from the sample, only the targeted DNA fragments remain. After detaching the beaded fragments, only fragments from the regions of interest are sequenced.

CNV detection from NGS data

While sequencing provides millions of reads per sample, these reads do not provide any useful information on their own. Making sense out of obtained data usually starts by assigning genomic positions to the reads by mapping the reads to reference genome. Tools like the Burrows-Wheeler Aligner (BWA) [37] try to determine the most similar position on a reference genome for any read obtained by sequencing, but also provide information on what parts do not fit perfectly. Such imperfect fits are possible whenever a small
change in DNA between the sample and the reference occurs, such as a single nucleotide variation (SNV) where one base pair is replaced by another.

Even when most of the millions of reads can be mapped back to their most likely origins on the genome, telling a healthy sample from an aberrated one is not a straightforward task. The reads are spread out over approximately 3.3 billion possible positions on the human genome, making simply browsing through all positions not an option. Additionally, not every mismatch between a read and the reference genome necessarily indicates a genetic aberration. Lots of technical mistakes happen during sequencing. Thus, tools to analyze a sample should apply some logic to only report aberrations if there is enough evidence.

Most tools for CNV detection make use of a comparison between the test sample and several reference samples [38]. Some implementations use straightforward comparisons where significance of read depth variations are determined by a z-score, or similar methods. When using a z-score, the mean read depth and the corresponding standard deviation of a chromosomal region are determined over the set of reference samples. These values are then compared to the actual read depth of the same region in a test sample, and the region is classified as aberrant if its significance value meets a predetermined threshold.

More sophisticated methods employ additional processing, taking other regions on the same chromosome into account. For example, hidden Markov models [39] can determine the most likely sequence of CNVs and unaffected regions on a chromosome, while circular binary segmentation [40, 41] can identify aberrated segments by finding edges where the difference in average read depth between the inner and outer regions is as large as possible.

While providing better results, more extensive methods to optimize results may come with additional compute time, which moreover translates to increase cost. For this reason, especially considering the huge amounts of data obtained by NGS, computational pipelines need to be designed that balance between accuracy and compute time.

Cell free DNA (cfDNA)

A sample needs to be obtained before any DNA is available for analysis. Usually, DNA is extracted from cells and is subsequently analyzed for diagnosis. However, removal of tissue is not always preferable or safe. Especially when the exact location of aberrated tissue is unknown, unreachable, or requires risky procedures, taking a biopsy is not an ideal option. For these situations, applications based on cell free DNA (cfDNA) may provide a solution.

When cells die, their contents do not simply disappear. Instead, the molecules they contain are broken down and either recycled or removed from the body. DNA molecules that were previously locked away in the nucleus of a cell are now released and float around freely in the blood stream, known as cell free DNA (cfDNA) [42, 43]. As the blood drags these cell-free molecules along, fragments that are no longer useful to the body can be removed by the kidneys. Sampling a few milliliters of an individual’s blood provides a mixture with millions of short cfDNA fragments. This mixture can be purified using a
Cell free fetal DNA (cffDNA)

Two decades ago cfDNA in pregnant women was found to be partially (3%-6%) of fetal origin \[45\], while later research showed this cfDNA actually occurs in higher percentages (~10%) \[46\]. This cell free fetal DNA (cffDNA) was found to originate from the placenta rather than from the fetus itself (Figure 9), but the placenta usually has the same genetic code as actual fetal DNA. The cffDNA is observable as early as 5 to 7 weeks of gestation, fetal fraction increases with gestational age \[47, 48\]. Additionally, this cffDNA was found to consist of noticeably smaller molecules than maternal cfDNA \[49\], and to diminish quickly, becoming undetectable for routine NIPT procedures within 2 hours after giving birth \[50\].
Non-invasive prenatal testing

To determine genetic aberrations of a fetus during pregnancy one usually takes a biopsy of the placenta. These tissue samplings, either amniocentesis or chorionic villus sampling (CVS), come with a risk (1:1000 - 2:1000) of causing miscarriage [52]. Analysis of cffDNA allows the detection of fetal genetic aberrations without the need for invasive sampling. Clearly, removing the need for painful and possibly dangerous procedures in prenatal healthcare is a much-desired improvement.

Lo et al. were the first to show that the whole fetal genome can be determined from cffDNA with the same read distribution as maternal DNA [53]. This finding in 2004 has laid the basis for fetal CNV detection in maternal plasma, which made non-invasive prenatal testing (NIPT) possible [54]. The cffDNA obtained from maternal plasma does, however, contain a lot of maternal cfDNA (80%-95% [46]). This complicates the detection of fetal genetic aberrations as they are easily overshadowed by (small) technical variations in maternal DNA. Early approaches that exploited cffDNA using microarray technology [55, 56] appeared promising for NIPT. However, NGS is increasingly used to replace array technologies, and recent developments made NGS implementations sensitive enough for low fractions of fetal DNA (>4%) [57].

Detecting fetal CNVs using cffDNA NGS data

Using next generation sequencing on maternal cfDNA, Chiu et al. [58] developed a z-score method to detect pregnancies of children with trisomy 21 (Down syndrome). In this test, all reads obtained through sequencing of cfDNA in maternal plasma are mapped to a reference genome, both for the test sample as well as for the reference samples. The distribution of the read frequencies on chromosome 21 is determined over the reference samples, and compared to the test sample using a z-score approach. If significantly different (absolute z-score >3), chromosome 21 is classified as aberrated. The method was reliable at affordable sequencing depth for diagnostic purposes [54, 59], but extending the same approach to detect trisomy 13 and 18 proved less reliable [60]. Methods to find sub-chromosomal aberrations have been developed as well [61], although their reliability was limited when samples were obtained at affordable sequencing depths. When no cost limitations to sequencing are taken into account, even small CNVs (as small as 300 kb) can be identified correctly [62].

Clearly, CNV detection is based on a statistical analysis. Hence, more data will give more statistical power for correct diagnosis. With less data per sample, variations in sequencing depth increase, decreasing the signal-to-noise ratio in a sample. For this reason, resequencing reference samples for every batch of test samples is required in the method introduced by Chiu [58]. To improve sensitivity, and possibly reduce the necessity of resequencing healthy samples, alternative approaches try to reduce the effects of read depth variations. Either by comparison among more similarly behaving chromosomes [63], or approaches to reduce biases such as GC-content [64].
Introduction

Figure 10: Schematic overview of how DNA is compressed using several levels of winding and folding. Important for this thesis is the “Beads-on-a-String” frame, where the double helix structure winds around a histone complex to form a nucleosome.

Detecting the fraction of fetal cfDNA in maternal blood

Without knowing the fraction of fetal DNA, one is not sure whether one is able to detect any aberration in the fetus. For example, a pregnancy where the mother has considerably more body mass than an average woman can result in significantly less abundant cffDNA in maternal plasma, severely complicating the detection of aberrations in the fetal genome. Additionally, some trisomies affect the placental growth, limiting its size and therefore the amount of cffDNA in maternal plasma. Consequently, it is important to know the fetal fraction of cell free DNA within the maternal plasma for having a reliable diagnostic test for fetal genetic aberrations.

In case of a male fetus, fetal fraction determination can be performed based on Y-chromosomal sequences [65, 66]. While this provides a straightforward solution for half of the pregnancies, another solution is needed to ensure proper testing for female fetuses. Interestingly, a difference in methylation profiles between maternal DNA and fetal DNA can be observed, and specific lab tests use this difference to get an indication of the fetal fraction in a sample [67]. The requirement of an extra lab flow, however, comes with additional costs per sample as well as increased chances of sample swaps and false indications (the determination of the fetal fraction in a sample is analyzed separately from the detection of CNVs in that same sample). Clearly, a bioinformatics approach to obtain the fetal fraction from the sequencing data directly is a superior solution for NIPT purposes. If enough sequencing depth is available, the distribution of single nucleotide polymorphisms (SNPs) can provide an indication of the fetal fraction [68], but the amount of data required is not feasible for routine diagnostics.

Fetal cfDNA is smaller than maternal cfDNA

Maternal cfDNA and cffDNA fragments have different sizes [49]. Based on this difference Lo et al. [69] developed a method which uses the length distribution as indicator for fetal fraction. Using paired-end sequencing, the length of sequenced molecules can be determined, and thus the fetal fraction. For many diagnostic labs, employing paired-
end sequencing requires changing their implemented NIPT workflow. Further, it would increase sequencing time per sample, reduce the sample turnover rate, and increase the cost per sample.

When looking into the actual size differences between maternal and fetal cfDNA, it appears that the threshold between the (smaller) fetal DNA fragments and (longer) maternal fragments lies close to the length a DNA sequence requires to wind around a histone complex (Figure 10) [70], indicating that observed size differences between maternal cfDNA and cffDNA are influenced by the presence of nucleosomes.

Contributions of this thesis

While the NIPT approach as introduced by Chiu et al. was affordable, several of its limitations sparked interest to improve the method they described, to enhance results and applicability of NIPT. Improvements included reducing the cost per sample, correct detection of aneuploidies other than trisomy 21, and detection of sub-chromosomal aberrations. This thesis describes a number of advances that realize these improvements, and are now applied for clinical purposes.

The contributions of this thesis are split into three main parts.

Part I focuses on the development of tools for detecting fetal CNVs from cfDNA in maternal blood. The main challenge is caused by the signal-to-noise ratio in the sequencing data obtained from maternal plasma. Because the fetal cfDNA is only a small fraction, even subtle noise hampers the detection of fetal CNVs. Chapter 1 introduces a within-sample comparison approach, enabling reliable aneuploidy and sub-chromosomal CNV detection using (partially fetal) cfDNA obtained from maternal plasma. In chapter 2, we discuss the introduction of this within-sample comparison method in a diagnostic setting, as well as its implications for the future of NIPT. Chapter 3 reflects a discussion on maternal cancer interfering with calling fetal CNVs. The implementation of novel detection methods comes with new, unexpected findings, and one of such findings is discussed in chapter 4. In this chapter, we show that carrying a common genetic property with no known phenotypes, like a rare fragile site, can influence NIPT results.

Part II focuses on determining the fraction of fetal cfDNA in a maternal plasma sample. While fetal cfDNA is assumed to be in maternal plasma, without indication of the actual fetal fraction in a sample the NIPT results cannot be considered fully reliable, as a lack of fetal cfDNA in a sample can cause a false negative result. Chapter 5 introduces a method that determines the fraction of cffDNA from maternal plasma which is based on a bias in fragment length between fetal and maternal cfDNA caused by the presence of nucleosomes. This results in a small detectable shift in the starting points of the cfDNA fragments between fetus and mother, from which the fetal fraction can be derived. This method is compared to other methods that determine the fetal fraction in chapter 6.
providing an overview of strengths and weaknesses of each of the methods. This comparison includes a method we developed called DEFRAG.

**Part III** extends the *within-sample comparison* approach to whole exome sequencing. As WES is heavily influenced by the distribution of target regions on the genome, straightforward CNV detection methods fail to provide reasonable accuracy. Consequently, results from state-of-the-art methods lack a sufficient agreement with proven array-based methods. In chapter 7, we introduce a WES-based CNV detection method (WISExome) using the *within-sample comparison* approach, and show that this method achieves an improved agreement with array results. This opens possibilities to get CNV information from the same patient data as is being used to acquire information about single nucleotide variations, which is of great benefit in routine diagnostic settings.

An important driver in this thesis is that our methodologies can be applied in a routine diagnostic setting to improve general health care. As a consequence, we did not opt for more advanced measurements such as sequencing with higher coverage, paired-end sequencing, or alternative sequencing such as methylation or long-read sequencing. Motivations for this decision are mainly cost per sample and ease of sample handling (minimizing changes for sample swaps, etc.). This has put more challenges on the bioinformatics analysis for which we present a number of solutions in this thesis. We do find it important that our proposed methodologies are widely available for clinical diagnostics. Therefore, all our methodologies and implementations are freely available to promote uptake by the academic community. Consequently, with this thesis we have made a substantial contribution to improving healthcare.
References


Introduction


[23] Figure: Pmx @ Wikimedia Commons. FISH image of bcr/abl positive rearranged metaphase. Aug. 5, 2005. URL: https://commons.wikimedia.org/wiki/File:Bcrablmet.jpg.


