Chapter 1

Introduction

1.1 A short history on chromosomes and DNA

In 1865 the Augustinian friar and scientist Gregor Mendel was the first to give a systematic account of the heredity of traits following specific laws [59]. In the following decades it was discovered that during cell division a substance in the cell nucleus, dubbed *chromatin* (stainable substance) by German biologist Walther Flemming, was divided over the two halves of the cells during a process that Flemming called *mitosen*, or mitosis [83, 28, 16]. A few years later, in 1890, German histologist Richard Altmann noted the presence of granules in cells that he believed were elementary organisms enclosed within cells, features later renamed 'mitochondria' by German microbiologist Carl Benda [2, 6]. In 1888, German anatomist Wilhelm Waldeyer was the first to use the term chromosomen – chromosomes, meaning colored bodies – to describe the individual pieces of chromatin thread [100, 16]. In the last decade of the 19th century, the German biologist August Weismann proposed that the chromosomes were the bearers of hereditary material, which he called keimplasma, or germ plasm [104]. At the time he was unaware of Mendel's work. However, after its rediscovery at the turn of the century, the cytologists Walter Sutton, from the United States of America, and Theodor Boveri, from Germany, both showed that chromosomes follow Mendelian laws [91, 9, 10, 18].

The chromosome theory of heredity quickly became the leading theory in the field that became known as genetics, a term introduced by the English biologist William Bateson in 1905 [46]. Around the same time, he and his colleagues observed coupling between different traits in pea plants [99, 53], leading the British biologist Thomas Morgan, upon further *Drosophila* studies, to state that 'we find "associations of factors" that are located near together in the chromosomes' [60]. This led to the theory of linkage a few years later [61]. It was several more decades before the normal human chromosome number was correctly defined as 46 by Indonesian cytogeneticist Joe Hin Tjio in 1956 [94]. After that it took only a few more years, until 1959, for French scientists Lejeune, Gauthier and Turpin to connect Down syndrome to the presence of a small extra chromosome [49]. One year later, the Philadelphia-based researchers Hungerford and Nowell discovered a small abnormal chromosome present in people with human chronic myelogenous leukemia, demonstrating the use of cytogenic techniques in diagnosis of hematological diseases [67]. This chromosome was later named the 'Philadelphia chromosome' and shown to be the product of translocation between chromosomes 9 and 22 [76]. In the meantime, based on work by British physicist Maurice Wilkins and chemist Rosalind Franklin, American biologist James Watson and British physicist Francis Crick created the double-helix DNA model containing the

four nucleotides – Adenine, Cytosine, Guanine and Thymine – which are paired A=T and G≡C [103, 106, 30]. Several years later Crick and his team inferred – without being able to sequence – the triplet DNA-protein translation code [17, 108]. However, it was not until the following decade, when British Chemist Frederick Sanger invented DNA sequencing methods, that the DNA sequence itself could be read [80, 81]. In 1963, it was discovered that apart from the nucleus, mitochondria also contained DNA [63]. In 1983, Huntington's disease was the first human disease to be linked to a specific genomic marker [34]. In the following years more diseases were linked to genomic markers and genetic diagnostics expanded from analysis of chromosomes to inclusion of DNA analysis. After the invention of Polymerase Chain Reaction (PCR), DNA analysis became much easier [79, 78] and at the turn of the 21st century scientists were able to create a draft sequence of the human genome [48, 96].

The introduction of so-called next-generation sequencing in 2005 ushered in the start of yet another era [56]. Sequencing costs decreased rapidly to the point that a whole genome can now be sequenced for less than 1000 dollars [32], opening up new possibilities for human genome analysis and bringing the fields of cytogenetics and molecular genetics closer together[1].

1.2 Human genome variation

With improving genomic analysis techniques came increasing knowledge about the composition of the human genome. When comparing any two individuals, their six billion base pair human genome will show many differences, or DNA variants. On average, everyone has around three million DNA variants that differ from the major allele present in the population, of which 10.000-11.000 are non-synonymous variants that change the triplet code and result in an amino-acid change of a protein [25]. Most of those variations do not cause disease but, as will be discussed in section 1.6, some variants are associated with or can contribute to a congenital disorder or a predisposition for the development of a disease. Several types of DNA variants can be distinguished. The smallest are Single Nucleotide Variants (SNVs) and indels: insertions or deletions of one or more bases (Figure 1.1A-D). When a larger stretch of DNA is lost or duplicated, the variant is considered to be structural variation

[1]Paragraph 1.1 suggests a logical and continuous timeline between discoveries. However, many of those discoveries were heavily contested and others were made by several research groups independently around the same time. This means that the history told in this paragraph could just as well have contained other names. Their omission is not meant to discredit their scientific contribution, but this introduction is too short to give a more nuanced vision of the scientific progress in genetics.

(SV) and the term Copy Number Variation (CNV) is used (Figure 1.1E-H). The size threshold to distinguish a large indel from a small CNV is arbitrary, and different definitions are used in literature. While 1 kb was traditionally used as the lower threshold for CNVs, variants larger than 50 bp are now labelled as CNVs [71, 93, 105]. In formal notation, duplication is regarded as a tandem duplication, i.e. insertion of a duplicate sequence, following directly 3'of the original copy (figure 1.1F), leaving the formal term 'insertions' to signify all other nucleotide insertions [21] (Figure 1.1D and H). However, in practice, the term duplication is used in a broader sense for copy number gains that can also include translocational insertions [43, 36]. One subset of duplications is repeat expansions in which a repeated nucleotide sequence is extended. An example of this is the CAG repeat that is extended in Huntington disease [77]. Another type of SV are translocations, in which terminal parts of chromosomes are exchanged (figure 1.1I). In reciprocal translocations both derivative chromosomes are present without an apparent net loss or gain of chromosomal content, but in so-called unbalanced translocations, the translocation results in loss of part of one chromosome and gain of part of another chromosome [54]. In Robertsonian translocations, two acrocentric chromosomes are connected at the centromere [73] (Figure 1.1J). A further type of DNA variation are inversions in which a nucleotide sequence is replaced by its reverse complement sequence [90, 21] (Figure 1.1K). While reciprocal translocations and inversions are balanced events in principle, deletions or insertions are often present around the breakpoints in both types of variations [90, 87]. A further type of chromosomal variation are aneuploidies, in which whole chromosomes are lost or gained (and can be considered as whole chromosome CNVs), such as in Down syndrome (Figure 1.1L).

1.3 Conventional techniques for variant detection

Over the years many different techniques have been developed to detect and chart the DNA constitution. The earliest was karyotyping, the technique used by Tjio and Levan, in which metaphase spreads are made that enable analysis of chromosomes using a microscope [94]. The development of chromosome staining techniques, such as Q,- C-, G- and R-banding, increased the resolution to a maximum of 5 Mb and enabled detection of smaller aberrations as well as more-specific determination of known variations [35, 51]. In situ hybridization techniques using radio- or fluorescent-labelled probes enabled detection of the presence and localization of specific parts of chromosomes [31, 5, 50]. It is particularly the latter, Fluorescence In Situ Hybridization (FISH), that paved the way for subchromosomal structural analysis, making

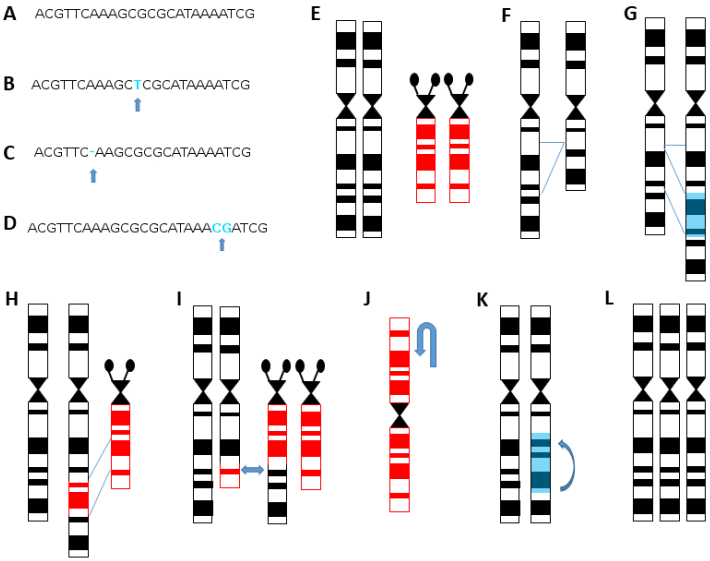

Figure 1.1: Human genome variation types: A) genomic base sequence, B) Single nucleotide variant, C) Indel: one base deletion, D) Indel: two base insertion, E) Two sets of chromosomes, F) CNV: Deletion, G) CNV: duplication, H) CNV: insertion, I) Reciprocal translocation, J) Robertsonian translocation, K) Inversion, L) Aneuploidy: trisomy

it possible to detect microdeletions of several hundreds of kilobases (kb) [19]. Further developments of this technique, such as fiber-FISH, increased the resolution to 50 kb using mechanically stretched chromosomes [70]. While these molecular techniques greatly advanced cytogenetics, analysis of solid tumors remained difficult because often no high-quality metaphases can be produced. Comparative Genomic Hybridization (CGH), an adaptation of FISH procedures, in which all patient DNA is fluorescently labelled and hybridized together with differently labelled reference DNA to high quality metaphases of a normal cell line enabled evaluation of aneuploidies, unbalanced translocations and CNVs [42]. In other words, all types of variations resulting in loss or gain of chromosomal material could be detected genome-wide to a maximum resolution of 10 Mb for deletions and 2 Mb for amplifications, without the need of patient metaphase spreads [47]. The same principle was used in array-CGH but, instead of metaphase spreads, a series of probes were used

as the hybridization target, making it possible to detect CNVs smaller than 1 kb depending on the number and placing of the probes [68, 69]. Using knowledge gained by earlier sequencing projects, it became possible to target specific SNPs, enabling the array to be used not only for CNV detection, but also as a genotyping tool [102]. A targeted technique to further enhance the resolution for CNV detection is Multiplex-Ligation Probe Amplification (MLPA), in which several targeted stretches of DNA are amplified in one experiment, after which a relative comparison is done within a series of samples. Depending on the included targets, deletions or duplications of single exons can be detected [84].

Where cytogenetics and molecular cytogenetics focused on the detection of structural variations, including copy-neutral variations and aneuploidies (figure 1.1E-L), molecular genetics focused on the detection of the nucleotide sequence, searching for SNVs, indels and repeat expansions (figure 1.1A-D). Often, Sanger sequencing was the method of choice here. However, only a short stretch of DNA of a single sample can be analyzed in a single experiment using this technique.

Variants are not always expected to be present in all cells from all tissues, as is the case with genetic mosaicism, including mitochondrial heteroplasmy, as well as in cancer. In karyotyping or FISH, a separate analysis is performed for each cell. By analyzing a large number of metaphases or nuclei, mosaicisms can be detected or excluded with high probability in the tissue studied [39, 4, 107]. Several DNA-based methods are also able to assess the presence of low fractions of a certain type of DNA in a larger pool. Real-time quantitative PCR measures fluorescence after each PCR cycle, then, through comparison with samples having a known concentration, fractions of targeted DNA stretches can be calculated for a sample [37]. Quantitative fluorescent (QF-)PCR measures the DNA concentration after a fixed number of PCR cycles [98]. A more recent addition is digital droplet PCR (ddPCR), where DNA fragments are encapsulated in oil droplets. For each droplet it is determined if a specific DNA sequence is present or not. Because tens of thousands droplets can be assessed in a single experiment, this technique has a high sensitivity for low-abundance variations [38]. It is no coincidence that so many techniques have been developed for DNA analysis, as each technique has distinct strengths and weaknesses. In karyotyping at low resolution, chromosome specific analysis can be done for the whole genome of a single cell. FISH increases resolution, but only gives information about targeted regions, while array gives high resolution whole genome information, but can't distinguish alleles and thus misses copy neutral structural variations. MLPA and Sanger sequencing have even higher resolution – the latter up to a single base pair – but, in a single experiment, are limited to analysis of a small part of

the genome. Therefore, using these conventional techniques to find all types of variations present in a single sample requires many different experiments.

1.4 Next-generation sequencing

In the mid-2000s, massive parallel sequencing was developed. With the introduction of this method, there was an immediate 50,000 fold drop in sequencing costs, resulting in the label 'next-generation sequencing' (NGS) [32]. NGS can be used for DNA as well as RNA sequencing. While the term NGS might suggest a single technique, it is in fact an umbrella-term encompassing many different technologies that sequence many DNA or RNA fragments in parallel and infer a read of the nucleotide sequence of each fragment. The first NGS platform available was developed by 454 Life Sciences using a pyrosequencing strategy [57]. Solexa then introduced NGS using reversible dye terminator chemistry [7] and Ion Torrent a non-optical system based on pH changes on nucleotide incorporation [75]. With these technologies being acquired by Roche, Illumina and Life Technologies, three strong contenders entered the short-read sequencing market. Other platforms focus on sequencing long single DNA molecules, such as Pacific Biosystems [26] and Oxford NanoPore [14], making use of real-time measurements of fluorescent signals and changes in current, respectively. Other contenders have since entered and left the NGS market, all using different chemistry and measurement tools. Because of this, technical bias is different from one technique to the other, although some genomic regions still remain a challenge for all platforms.

Although the exact methods used differ between different NGS techniques, their general approach is similar, as shown in figure 1.2, although the strong and weak points vary between the platforms. For NGS DNA analysis, various input materials can be used. Some contain fragmented DNA, such as blood plasma or formalin-fixed paraffin-embedded (FFPE) material (figure 1.2A), while others containing high quality DNA, for example white blood cells, bone marrow or cultured cells (figure 1.2B). The first step in all DNA NGS procedures is to isolate DNA from the material. In the materials where the DNA is already fragmented, short DNA fragments are isolated (figure 1.2C).

Source materials containing higher quality DNA can give rise to longer DNA-fragments (figure 1.2D) or even very long DNA fragments, if DNA breakage is prevented during isolation (figure 1.2E). The short-read sequencing methods work best when using relatively short DNA fragments. For these techniques, DNA needs to be fragmented if the input fragments are too long

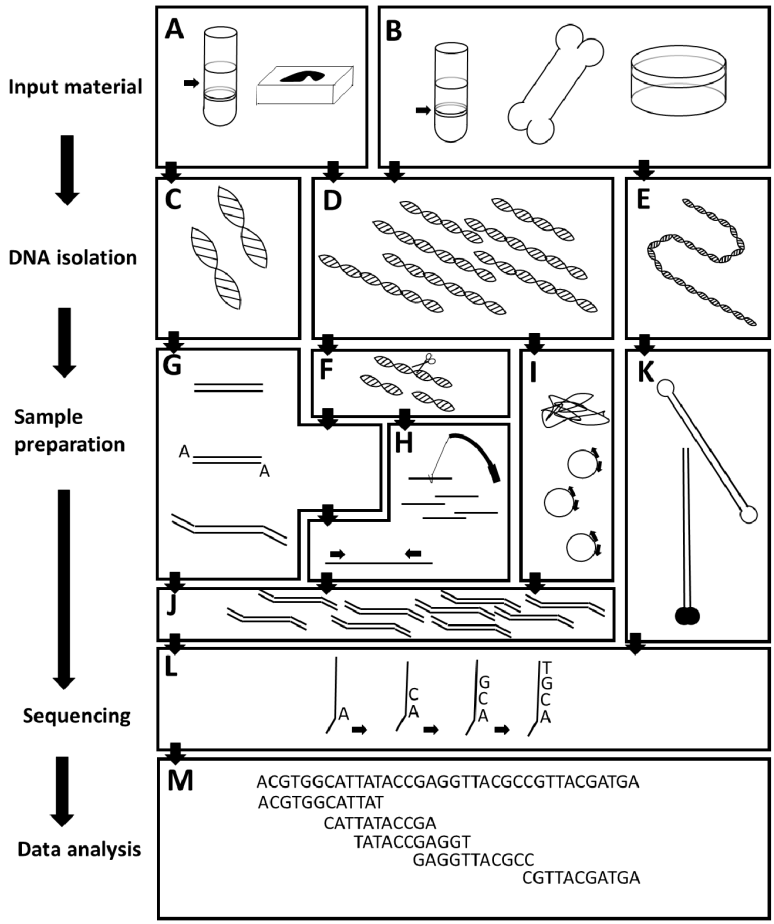

Figure 1.2: DNA Next-generation sequencing workflows. A) Sources of fragmented DNA, such as blood plasma or FFPE material, B) sources of high quality DNA, such as white blood cells, bone marrow cells or cultured cells, C) isolated fragmented DNA, D) isolated high-quality DNA, E) isolated long fragments of DNA, F) DNA fragmentation, G) sample preparation (end-repair, dA-tailing and adapter ligation), H) enrichment via capturing or amplicon sequencing), I) alternative sample preparation, such as Targeted Locus Amplification or ATAC-seq, J) PCR, K) long-read sequencing sample preparation, L) sequencing of the DNA, M) data analysis to transform sequenced DNA into sequence reads and subsequently into sample-specific genomic sequences.

(figure 1.2F). The most basic short-read strategy is whole genome sequencing (WGS). Sample preparation consists of adding so-called 'adapters' to DNA-fragments, thus making the fragments suitable for sequencing (figure 1.2G). If only a part of the genome needs to be sequenced, the DNA can be enriched for the sequences of interest (figure 1.2H). Various methods can be used to reach this goal, such as DNA capturing, in which short RNA or DNA sequences complementary to the region of interest called 'baits' are used to fish out specific parts of the genome. A second method is amplicon sequencing. Here, similar to Sanger sequencing, two primers are used that bind to their complementary sequence and copy the genomic sequence in between the primers. Such enrichment techniques are used in, for instance, whole exome sequencing (WES) and gene panels that target specific genes of interest. In addition to these 'standard' sample preparation methods, alternative sample preparations can be performed that have a different perspective on the genome (figure 1.2I), for instance using proximity ligation [89], targeted locus amplification [20], or chromatin-immunoprecipitation or by enzymatic digestion [41].

The final step in the sample preparation is PCR amplification to produce sufficient fragments of the DNA of interest to be sequenced (figure 1.2J). Alternatively, long-read sample preparation methods can be used (figure 1.2K). The bases of the DNA fragments are subsequently read by the sequencer (figure 1.2L). Data analysis is then carried to determine the nucleotide sequence of the DNA fragments, and the genomic sequence of the sample can be inferred through further processing, for instance through alignment of sequenced reads to a reference genome (figure 1.2M). Once the genomic sequence is inferred as far as possible, the presence or absence of variants can be de determined and interpreted in the context of a scientific or diagnostic question. An important step in variant calling and interpretation is to distinguish true positive and negative results from false ones. Knowing where variants can be missed, or where artefacts are more likely to occur, can be important for making a correct interpretation. Moreover, if the cause of artefacts is known, analysis procedures can be adapted to counteract sources of bias and create a more optimal balance between sensitivity and specificity.

1.5 Technical bias and error rates

Where conventional techniques have proven their worth in genetic diagnostics, NGS procedures and analysis still need to be optimized, and refining the methods to improve their sensitivity and specificity remains a challenge. The aim of the different NGS techniques is to measure the exact nucleotide

sequences of the DNA fragments. However, technical bias and sequencing errors create noise, resulting in some of the nucleotides in a sequence read being called incorrectly. This error rate is much higher for long-read technologies than for short-read sequencing. Depending on the chemistry and platform used, error rates range from 0.1% to 15% [32]. These error rates are presented as base quality scores and, when several reads are combined to infer a genotype, as a genotype quality score [65]. However, it has been shown that discordance rates between short-read samples that have been analyzed twice are higher than would be expected using the genotype quality scores [101, 82], which suggests that error rates are higher than the sequence data lead us to believe. Whereas some of the sequencing errors are random, each type of sequencer, as well as each experimental design, has its own systematic biases that occur at specific sequence patterns, inverted repeats or homopolymers [62, 82]. Because some of the errors are made during PCR amplification, base quality scores are not always sufficient to determine the chance that a specific base is called correctly for a sequenced DNA fragment. This is especially important when the aberration of interest is expected to occur in only a subset of the analyzed DNA fragments, as is the case for germline and somatic mosaic variants and for non-invasive prenatal testing (NIPT), where fetal DNA is analyzed in the presence of maternal DNA, because fewer sequence reads will be present to support a genotype call. An important contributor to the creation of bias during PCR is the GC percentage in the DNA fragment. If a high (>65%) or low (<12%) percentage of guanine or cytosine bases are present, the DNA fragments are barely amplified during PCR, with the amplification efficiency gradually increasing with GC percentages closer to 50% [1]. With each PCR cycle needed in the experiment, the GC bias will grow, although this bias can also occur during PCR-steps that are part of the sequencing procedure itself [74]. The severity of this bias can differ between samples and experiments. An extra effect of using many PCR cycles in sample preparation is that the number of reads originating from the same DNA fragment, called duplicate reads, will grow. This can lead to a risk of overestimating the effective coverage and sensitivity as well as the chance of amplifying errors occurring during extension in early PCR cycles, thus reducing specificity.

For WGS, fewer PCR cycles are usually needed in the sample preparation, leading to a relatively even coverage between different genomic regions. However, targeted techniques such as WES and targeted NGS (tNGS) that rely on selective amplification of genomic regions of interest require PCR during sample preparation. In general, the rule applies that the lower the amount of input material or the smaller the targeted region, the more PCR cycles are needed, up to more than 30 cycles for some procedures. At 30 PCR cycles,

over a billion copies of the same original DNA fragment are generated. In contrast, after 10 PCR cycles, just over one thousand copies are present. When randomly sheared DNA fragments are amplified and sequenced, duplicate reads can, to a certain extent, be identified based on the fact that they have an (almost) identical sequence. However, in amplicon-based sequencing, which uses primers to amplify a region of interest, it is expected that different original DNA fragments give rise to reads with the same sequence. This makes it more difficult to distinguish those reads from each other, unless separate molecular identifiers are used.

But, even when all technical bias is corrected for, not all parts of the genome are accessible, especially in short-read sequencing. Many parts of the genome are not unique, for instance genes that have pseudogenes [55]. When a DNA-fragment originating from such a region is sequenced, there is no way to determine from the sequenced read itself if it is informative for the region of interest or for the other region that has the same sequence.

1.6 DNA variant detection in genome diagnostics

In current genome diagnostics many of the DNA variant detection methods described in sections 1.3 and 1.4 are used. The types of variations that are searched for, as shown in figure 1.1, are different for different diagnostic questions. Moreover, the variants being examined can be present in only some of the cells – and therefore only part of the DNA analyzed – as discussed earlier. In the paragraphs below I discuss three important types of variants that need specific analysis and interpretation approaches: germline variants, somatic variants and variants found in prenatal testing.

1.6.1 Germline variants

Germline variants are present at the formation of the zygote and, in principle, are present in all cells, including the germline [33]. For genetic analysis, white blood cells or fibroblasts provide a source of high-quality DNA. Germline variants can be transmitted from parent to child and can therefore result in multiple affected relatives within a family. Depending on the nature of the variants, a disease phenotype may develop during childhood, or adulthood, or even not at all. For Mendelian diseases the inheritance pattern for variants in autosomal chromosomes (i.e. chromosomes 1-22) can be autosomal dominant (AD) or autosomal recessive (AR). In AD inheritance, a variant in only one of the alleles can result in the disease phenotype. In AR inheritance, both parents transmit a pathogenic variant. Variants present in sex-chromosomes or mitochondria have different inheritance patterns. Because men carry one

copy of each sex chromosome, a sex-chromosome-related recessive trait will result in a phenotype when a single copy of the causal variant is present. Mitochondria are always transmitted from mother to child, leading to phenotypes caused by mitochondrial variants only being inherited through the maternal line.

One example of an AD hereditary disease is Lynch syndrome, one of the most common cancer predisposition syndromes. In Lynch syndrome, SNVs, indels, intragenic deletions or duplications cause a deficiency in the mismatch repair system that significantly increases the risk of developing cancer compared to the general population, although, as in other cancer-predisposing syndromes, not all carriers of pathogenic variants develop cancer [95, 92]. It is estimated that around 1 in 300 people carry a pathogenic variant in one of the genes associated with Lynch syndrome [11]. One of the most common AR disorders is cystic fibrosis, which leads to dysfunctional chloride channels that cause thickened mucus and affects around 1 in 3500 individuals in Europe [109, 85]. Children with cystic fibrosis often inherit a non-functional allele of the *CFTR* gene from both of their parents, who themselves don't present with the disease phenotype because they have a functional copy of the gene. An example of a common recessive X-linked trait is red-green colorblindness, which affects 1 in 12 males and 1 in 200 females in populations with Northern-European ancestry [72]. The prevalence of mitochondrial diseases is highly dependent on the population and is associated with, among other conditions, neurological diseases and ataxia [12].

It is also possible that variants appear *de novo* during de formation of the gametes. *De novo* means that a variant is found in an individual even though neither of the parents carry this variant. Such a variant can arise through mistakes in copying DNA for SNVs and indels, through errors in crossing over for SVs, or through non-disjunction for aneuploidies. Examples of syndromes caused by SVs are Down syndrome (trisomy 21), Klinefelter syndrome (XXY), Turner syndrome (X0), Di-George syndrome (del 22q11) and the 1q21.1 microduplication syndrome.

1.6.2 Somatic variants

When a DNA variant is not present in the zygote but rather originates from a later cell division, it is called a somatic variant. If such a variant originates during embryonic development, it will be present in many cells; if it occurs later in life, it may be present in a small number of cells [29]. Some of the syndromes mentioned in the previous paragraph, Down syndrome and Turner syndrome for instance, can have their origin not only in germ cells, but also be the result of somatic mosaics. Mosaics may not lead to a clinical abnormal

phenotype, depending on the distribution of the somatic variants over cells and tissues. Low level mosaics in parents that include their germ cells may be difficult to distinguish from *de novo* cases discussed in the previous section. Mosaics may also arise through a germline variation with a rescued cell-line in which the variation is eliminated [22, 40].

Some disorders such as segmental neurofibromatosis [88] or McCune-Albright syndrome [24], in which parts of the body are affected while other parts are unaffected, are caused by mosaics. In cancer, somatic variants are the main cause of tumorigenesis. A tumor can develop when a gene variant causes uncontrolled cell division, as is the case with the Philadelphia chromosome [44], or fails to lead the cell into appropriate cell-death, as is the case with variations affecting the *MYC* gene [23]. A cell that develops such a variant can then grow into a clonal population, which can later on develop into further subclones, together constituting the tumor cells [66, 64, 58]. In advanced disease stages, some variants can be present in a high percentage of cells. However, in earlier stages, after treatment or when a new variant has arisen in a subclone, it can be the case that only a small percentage of the cells analyzed carry the variant. In addition, tumor samples sent in for analysis typically contain both tumor cells and normal cells (e.g. lymphocytes or stromal cells), which adds to the mosaic nature of gene variants in these samples.

Somatic variants in tumor or hematological cells can consist of all the variant types described in section 1.2. However, while large structural variants, including aneuploidies, are rare events when looking at germline variants, they are more prevalent in cancer cells, where complex aberrant karyotypes are also seen. The main challenge for somatic variant detection in tumors is the possible presence of a wide variety of DNA variants and, sometimes balanced, chromosomal aberrations in a low percentage of the cells or DNA to be analyzed. In addition, the material containing the variations, such as bone marrow or tumor material, is harder to come by and often of lower quality than that used for germline variation detection.

1.6.3 Prenatal testing

Genetic variants can also be detected prenatally. Conventionally, such tests are offered to pregnant women at an elevated risk of carrying a child with a chromosomal abnormalities, most notably Down syndrome, Patau syndrome (trisomy 13) and Edwards syndrome (trisomy 18), and for hereditary disease-causing-gene variants previously identified in one or both of the parents. Conventional invasive prenatal tests are performed using cells from the fetus or from extra-fetal tissue that shares genetic origin with the fetus: amniotic fluid

cells (fetal and extra-fetal origin) or chorionic villi cells (extra-fetal, placental). The main problem with the frequently used types of invasive procedures – amniocentesis and chorionic villi biopsy – is a risk of a procedure-related miscarriage of 0.3% and 0.5%, respectively [8]. Fortunately, the mother's blood can also be used as a source of short fragments of extra-fetal DNA [52]. This so-called cell-free fetal DNA (cffDNA) circulates through the blood stream of a pregnant woman, next to a greater fraction cell-free DNA (cfDNA) originating from her own cells. On average only around 12% of the cfDNA is cffDNA, though it can be much lower [3]. The cfDNA, including the cffDNA, can be isolated from blood plasma to enable non-invasive prenatal testing (NIPT). Because no invasive procedures are needed in NIPT, there is no risk of inducing a miscarriage. For this reason, NIPT has quickly become a mainstream genetic test. In the Netherlands NIPT has been offered to women with a high risk of carrying a child with a trisomy 13, 18 or 21 since 2014 and to all pregnant women since 2017 [15]. However, because a mosaic of cffDNA and maternal cfDNA is present, similar technical challenges have to be overcome to those faced in somatic variant testing.

1.7 Aims of this thesis

As we have seen throughout the introduction, many different DNA variants can be present in a single sample. However, technical bias, size of the variation, copy-neutrality of variations, mixed cell-populations or DNA samples and the biological origin of analyzed DNA fragments can all create noise in the analysis process. The task of the clinical genetics laboratory is to look through this noise to detect and interpret the presence or absence of relevant variants. When using conventional techniques, many independent tests are needed to overcome different types of noise or to change resolution, sensitivity, number of variants analyzed and the ability to detect balanced variants or not. NGS has the potential to replace all these tests. However, not all types of variants are easy to detect. By using efficient sample preparation and analysis algorithms that can distinguish artefacts from variants, NGS is able to challenge conventional techniques and may become the method of choice for all diagnostic questions related to the detection of DNA variants. The studies in this thesis aim to improve NGS DNA analysis for detection of germline SNVs, indels and CNVs, somatic translocations and trisomy detection through NIPT, as well as interpretation of analysis outcomes. In this thesis, I introduce new tools, methods and algorithms for NGS DNA analysis and interpretation and, in some cases, use them in a practical application (figure 1.3).

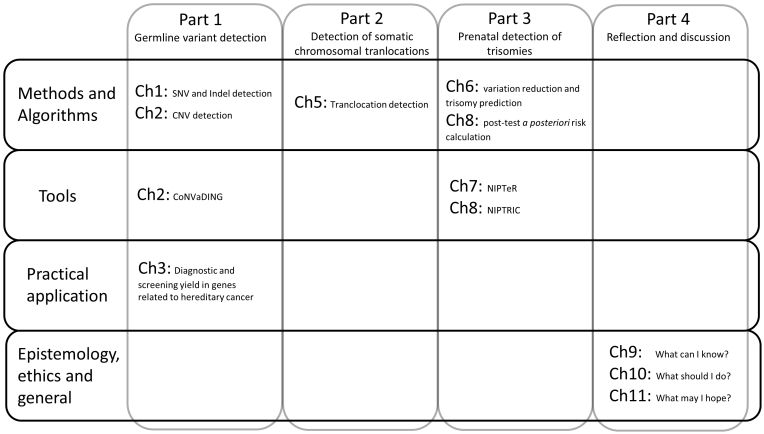

Figure 1.3: Overview of the topics addressed in the thesis chapters.

1.7.1 Part 1: Germline variant detection (chapters 2, 3 and 4)

The most prevalent germline variants – SNVs, indels and small CNVs – were conventionally analyzed mainly using Sanger sequencing for SNV and indel detection and MLPA to detect CNVs. However, only a short stretch of DNA can be analyzed in each measurement using these techniques, limiting the number of genes that can by analyzed in a single experiment. In chapter 2 we set out to implement tNGS as a stand-alone diagnostic test to enable analysis of a large set of genes in a single test and replace Sanger sequencing in clinical diagnostics. For this we developed, validated and established quality criteria for a tNGS genepanel to detect SNVs and indels with high sensitivity and specificity in 48 genes involved in cardiomyopathies, ultimately demonstrating that tNGS is a technique suitable for diagnostic use. In chapter 3 we further expand the application of tNGS and enable simultaneous detection of CNVs up to the single exon level, next to SNVs and indels. Because it is likely in tNGS that CNV breakpoints are located outside targeted regions, CNVs can only be inferred through analysis of read depth. However, laboratory-induced variability of read depth is larger than biological variability. To look through the experimental noise and detect (single-exon) CNV in tNGS data, we introduce new algorithms with strict quality control that we implement in the open-source tool CoNVaDING (**Co**py **N**umber **Va**riation **D**etection **I**n **N**ext-generation sequencing **G**ene panels). In chapter 4 we set out to use

the tools and methods developed in the previous chapters in the context of hereditary cancer, for which we have analyzed 85 genes in 2,090 patients and 1,326 individuals from the general Dutch population. The first goal here was to determine the diagnostic yield, focusing on genes with a relation to the cancer type warranting referral. The second goal was to determine the findings if, in addition to these genes, we search for pathogenic or likely pathogenic variants in genes without such a relation (secondary findings), and how often variants leading to a cancer predisposition occur in the general Dutch population.

1.7.2 Part 2: Detection of somatic chromosomal translocations (chapter 5)

The second part of this thesis consists of a single chapter that focuses on somatic translocation detection. In current hematological malignancy diagnostics SVs, including translocations, are detected using various conventional techniques. Using karyotyping, large rearrangements are detected on a single-cell basis. However, this technique is unable to detect some so-called cryptic translocations. FISH and RT-PCR are needed to detect those, but these techniques can only target one SV or fusion-gene at a time. In chapter 5 we aim to develop an NGS-based technique to target 18 genes and detect translocations involving one of those genes commonly involved in acute leukemia, regardless of their translocation partner, to be suitable for use as a first-line screening tool in diagnostics. For this we make use of Targeted Locus Amplification (TLA) [20] to create a multiplex TLA acute leukemia gene panel. In addition to the genes themselves, our panel captures DNA physically close to the targeted genes, which enables the capture and detection of chromosomal translocation partners even if they are not in the targeted panel. We develop analysis and interpretation strategies and demonstrate for several targeted genes that the panel detects translocations involving those genes at 10% aberrant cells. We conclude that multiplex TLA is a promising technique that it needs further optimization before it can replace conventional methods.

1.7.3 Part 3: Prenatal detection of trisomies (chapters 6, 7 and 8)

Part three of this thesis is dedicated to NIPT. Where conventional methods for prenatal trisomy detection, such as karyotyping, FISH, QF-PCR or array, rely on invasive procedures, NIPT can be performed using ultralow-coverage NGS data. Using a basic sample preparation with as few PCR cycles as possible, the short cfDNA fragments are made available for sequencing. Several algorithms

are described in the literature to analyze such ultralow-coverage NGS data to predict the presence of a trisomy [13, 27, 86]. These strategies rely on the comparison of the sample of interest to a group of non-trisomy control samples to determine if significantly more sequence reads are present that originate from DNA fragments of the potential trisomic chromosome. Because cffDNA is mixed with maternal DNA, a trisomy will only cause a small increase in the fraction of reads of the chromosome involved. Therefore it is important to make the variability in chromosomal fractions as small as possible between samples. In chapter 6 we introduce novel algorithms to analyze ultralow-coverage NGS data and obtain a higher sensitivity for trisomy detection than found using earlier described calculations. In addition, we create a quality metric that can be used to detect if the available reference samples are suitable for comparison with the sample analyzed. In chapter 7 we describe *NIPTeR*, an open-source R package that makes the algorithms developed in chapter 6 available along with the algorithms described in the literature for analysis of NIPT data. Two women receiving a similar test result from NIPT do not necessarily have a similar risk of carrying a child with a trisomy. In chapter 8 we focus on the clinical interpretation of the NIPT result, taking into account not only biological and technical characteristics of the test, but also the population to which the woman being tested belongs. Including these pre-test conditions in the interpretation might result in different risk profiles for women from different risk-groups who have the same raw test result. We created algorithms to calculate such a personalized post-test risk for a specific fetal trisomy and made these available in NIPTRIC, an online calculator.

1.7.4 Part 4: Reflection and discussion (Chapters 9, 10 and 11)

Inspired by the three questions posed by Immanuël Kant in his *Kritik der reinen Vernunft* published in 1781/1787: "what can we know?", "what should I do?" and "what may I hope?" [45][p. 728], in part four of this thesis I reflect on and discuss the methods, tools and algorithms described in this thesis. In chapter 9 I look back on the chapters from an epistemological point of view. In genetic diagnostics we infer the genetic or genomic constitution of a person through a measurement outcome. I elaborate on the concept of noise that I define as 'everything that, from a certain perspective, blocks the path between reality and measurement outcome'. Throughout this thesis we are battling four types of such noise: biological noise, laboratory-induced noise, sequencing noise and data analysis noise. The variants of interest are hidden behind this noise, but through innovative perspectives we are better able to look through the noise and correctly interpret measurement outcomes.

In chapter 10 I make an ethical reflection on the technologies introduced in this thesis. I use the theories of Peter-Paul Verbeek who states that artefacts are morally charged and mediate human action [97][p 21]. I try to uncover intended and unintended moral consequences of the availability of the methods, tools and algorithms presented in this thesis.

In chapter 11 I address the last question of Kant: 'what may I hope?' and put the work presented in this thesis in broader perspective in the general discussion and to give future perspectives on developments in NGS DNA analysis.

Bibliography

[1] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011.

[2] R. Altmann. *Elementärorganismen und ihre beziehungen zu den zellen*. Metzger & Wittig, Leipzig, 1890 zweite auglage 1894.

[3] Ghalia Ashoor, Leona Poon, Argyro Syngelaki, Beatrice Mosimann, and Kypros H. Nicolaides. Fetal fraction in maternal plasma cell-free dna at 11–13 weeks' gestation: Effect of maternal and fetal factors. *Fetal Diagnosis and Therapy*, 31(4):237–243, 2012.

[4] Sikkema-Raddatz B, S Castedo, and GJ te Meerman. Probability tables for exclusion of mosaicism in prenatal diagnosis. *Prenat. Diagn.*, 17(2):860–866, 2009.

[5] JGJ Bauman, J Wiegant, P Borst, and P van Duijn. A new method for fluorescence microscopical localization of specific dna sequences by in situ hybridization of fluorochrome-labelled rna. *Exp. Cell Res.*, 128(2):485–490, 1980.

[6] C. Benda. Ueber die spermatogenese der vertebraten und höheren evertebraten. ii. theil. die histiogenese der spermien. *Arch. Anal. Physiol.*, pages 393–398, 1898.

[7] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, and et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[8] L. Beulen, B. H. W. Faas, I. Feenstra, J. M. G. van Vugt, and M. N. Bekker. Clinical utility of non-invasive prenatal testing in pregnancies with ultrasound anomalies. *Ultrasound in Obstetrics & Gynecology*, 49(6):721–728, Jun 2017.

[9] T Boveri. *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns*. Gustav Fischer, Jena, 1904.

[10] T Boveri. Die blastomerenkerne von ascaris megalocephala und die theorie der chromosomenindividualität. *Arch Zellforsch*, 3:181–268, 1909.

[11] Cancer.net. Lynch syndrome: www.cancer.net/cancer-types/lynch-syndrome, 2005-2018.

[12] Patrick F. Chinnery and Aurora Gomez-Duran. Oldies but goldies mtdna population variants and neurodegenerative diseases. *Frontiers in Neuroscience*, 12, Oct 2018.

[13] R. W. K. Chiu, K. C. A. Chan, Y. Gao, V. Y. M. Lau, W. Zheng, T. Y. Leung, C. H. F. Foo, B. Xie, N. B. Y. Tsui, F. M. F. Lun, and et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of dna in maternal plasma. *Proceedings of the National Academy of Sciences*, 105(51):20458–20463, Dec 2008.

[14] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nature Nanotechnology*, 4(4):265–270, Feb 2009.

[15] NIPT Consortium. Meerovernipt: Online: http://www.meerovernipt.nl/content/de-studies-trident-1-en-trident-2 (visited on march 9th 2018), 2018.

[16] T Cremer and C Cremer. Rise, fall and resurrection of chromosome territories: A historical perspective. part i. the rise of chromosome territories. *Eur. J. Histochem*, 50(3):161–176, 2006.

[17] FHC Crick, L Barnett, S Brenner, and RJ Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, Dec 1961.

[18] EW Crow and Crow JF. 100 years ago: Walter sutton and the chromosome theory of heredity. *Genetics*, 160:1–4, 2002.

[19] Chenghua Cui, Wei Shu, and Peining Li. Fluorescence in situ hybridization: Cell-based genetic diagnostic and research applications. *Frontiers in Cell and Developmental Biology*, 4, Sep 2016.

[20] Paula J P de Vree, Elzo de Wit, Mehmet Yilmaz, Monique van de Heijning, Petra Klous, Marjon J A M Verstegen, Yi Wan, Hans Teunissen, Peter H L Krijger, Geert Geeven, and et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature e Biotechnology*, 32(10):1019–1025, Aug 2014.

[21] JT den Dunnen, R Dalgleish, DR Maglott, RK Hart, MS Greenblatt, and et al. Hgvs recommendations for the description of sequence variants: 2016 update. *Human Mutation*, 37(6):564–569, Mar 2016.

[22] L. Devlin and P.J. Morrison. Accuracy of the clinical diagnosis of down syndrome. *Ulster Med. J.*, 73:4–12, 2004.

[23] D. Dominguez-Sola and J. Gautier. Myc and the control of dna replication. *Cold Spring Harbor Perspectives in Medicine*, 4(6):a014423–a014423, Jun 2014.

[24] Claudia E Dumitrescu and Michael T Collins. Mccune-albright syndrome. *Orphanet Journal of Rare Diseases*, 3(1), May 2008.

[25] RM Durbin, DL Altshuler, RM Durbin, GR Abecasis, DR Bentley, and et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[26] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, and et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[27] H. Christina Fan and Stephen R. Quake. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS ONE*, 5(5):e10439, May 2010.

[28] W Flemming. *Zellsubstanz, Kern und Zelltheilung*. F.C.W. Vogel, Leipzig, 1882.

[29] Steven A. Frank. Somatic mosaicism and disease. *Current Biology*, 24(12):R577–R581, Jun 2014.

[30] RE Franklin and RG Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, Apr 1953.

[31] JG Gall and Pardue ML. Formation and detection of rna-dna hybrid molecules in cytological preparations. *Proc Natl Acad Sci U S A*, 63(2):378–83, 1969.

[32] S Goodwin, JD McPherson, and WR McCombie. Coming of age: ten years of next-generation sequencing technologies. 17(6):333–351, Jun 2016.

[33] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, and et al. *An Introduction to Genetic Analysis, 7th edition*. W.H.Freeman, New York, 2000.

[34] JF Gusella, NS Wexler, PM Conneally, SL Naylor, MA Anderson, and et al. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 306(5940):234–238, Nov 1983.

[35] JL Hamerton and PA Jacobs. Paris conference (1971): Standardization in human cytogenetics. *Cytogenetics*, 11:313–362, 1972.

[36] NM Hanemaaijer, B Sikkema-Raddatz, G van der Vries, T Dijkhuizen, R Hordijk, , and et al. Practical guidelines for interpreting copy number gains detected by high-resolution array in routine diagnostics. *European Journal of Human Genetics*, 20(2):161–165, Sep 2012.

[37] CA Heid, Stevens J, Livak KJ, and Williams PM. Real time quantitative pcr. *PCR Methods Appl.*, 6(10):986–994, 1996.

[38] Benjamin J. Hindson, Kevin D. Ness, Donald A. Masquelier, Phillip Belgrader, Nicholas J. Heredia, Anthony J. Makarewicz, Isaac J. Bright, Michael Y. Lucero, Amy L. Hiddessen, Tina C. Legler, and et al. High-throughput droplet digital pcr system for absolute quantitation of dna copy number. *Analytical Chemistry*, 83(22):8604–8610, Nov 2011.

[39] E.B. Hook. "exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use. *Am. J. Hum. Genet.*, 29:94–97, 1977.

[40] Ernest B. Hook and Dorothy Warburton. Turner syndrome revisited: review of new data supports the hypothesis that all viable 45,x cases are cryptic mosaics with a rescue cell line, implying an origin by mitotic loss. *Human Genetics*, 133(4):417–424, Jan 2014.

[41] Shan Jiang and Ali Mortazavi. Integrating chip-seq with other functional genomics data. *Briefings in Functional Genomics*, 17(2):104–115, Mar 2018.

[42] A Kallioniemi, O-P Kallioniemi, D Sudar, D Rutovitz, JW Gray, and et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, 1992.

[43] S-HL. Kang, C Shaw, Z Ou, PA Eng, ML Cooper, and et al. Insertional translocation detected using fish confirmation of array-comparative genomic hybridization (acgh) results. *American Journal of Medical Genetics Part A*, 152A(5):1111–1126, May 2010.

[44] Zhi-Jie Kang, Yu-Fei Liu, Ling-Zhi Xu, Zi-Jie Long, Dan Huang, Ya Yang, Bing Liu, Jiu-Xing Feng, Yu-Jia Pan, Jin-Song Yan, and et al. The philadelphia chromosome in leukemogenesis. *Chinese Journal of Cancer*, 35(1), May 2016.

[45] I. Kant. *Kritik der Reinen Vernunft*. Felix Meiner Verlag, Hamburg, 1781/1787 Herausgabe 1956.

[46] M Keynes and TM Cox. William bateson, the rediscoverer of mendel. *Journal of the Royal Society of Medicine*, 101(3):104–104, Mar 2008.

[47] James LA. Comparative genomic hybridization as a tool in tumour cytogenetics. *The Journal of Pathology*, 187(4):385–395, 1999.

[48] ES Lander, LM Linton, B Birren, C Nusbaum, MC Zody, and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[49] J Lejeune, M Gauthier, and R Turpin. Les chromosomes humains en culture de tissus. *C. R. Acad. Sci*, 248:602–903, 1959.

[50] JM Levsky. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, 116(14):2833–2838, Jul 2003.

[51] S. Liu, L. Song, D. S. Cram, L. Xiong, K. Wang, R. Wu, J. Liu, K. Deng, B. Jia, M. Zhong, and et al. Traditional karyotypingvscopy number variation sequencing for detection of chromosomal abnormalities associated with spontaneous miscarriage. *Ultrasound in Obstetrics & Gynecology*, 46(4):472–477, Oct 2015.

[52] Y.M.D. Lo, N. Corbetta, Chamberlain P.F., Rai V., Sargent I.L., Redman C.W.G., and Wainscoat J.S. Early report.presence of fetal dna in maternal plasma and serum. *Lancet*, 350:485–487, 1997.

[53] I Lobo and K Shaw. Discovery and types of genetic linkage. *Nature Education*, 1(1):139, 2008.

[54] S Maithripala, U Durland, J Havelock, S Kashyap, J Hitkari, and et al. Prevalence and treatment choices for couples with recurrent pregnancy loss due to structural chromosomal anomalies. *Journal of Obstetrics and Gynaecology Canada*, Dec 2017.

[55] Diana Mandelker, Ryan J. Schmidt, Arunkanth Ankala, Kristin McDonald Gibson, Mark Bowser, Himanshu Sharma, Elizabeth Duffy, Madhuri Hegde, Avni Santani, Matthew Lebo, and et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genetics in Medicine*, 18(12):1282–1289, May 2016.

[56] ER Mardis. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303, Jun 2013.

[57] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, and et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Jul 2005.

[58] Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4):613–628, Feb 2017.

[59] G Mendel. Versuche über pflanzen-hybriden. *Verh. Naturforsch. Ver. Brünn*, 4:3–47, 1866.

[60] TH Morgan. Random segregation versus coupling in mendelian inheritance. *Science*, 34(873):384–384, Sep 1911.

[61] TH Morgan, AH Sturtevant, Muller HJ, and Bridges CB. The mechanism of mendelian heredity. *Henry Holt, New York*, 1915.

[62] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, and et al. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, May 2011.

[63] M.M.K. Nass and S. Nass. Intramitochondrial fibers with dna characteristics: I. fixation and electronic staining reactions. *J. Cell Biol.*, 19:593–611, 1963.

[64] Christopher T Naugler. Population genetics of cancer cell clones: possible implications of cancer stem cells. *Theoretical Biology and Medical Modelling*, 7(1), Nov 2010.

[65] Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, Jun 2011.

[66] P.C. Nowell. Clonal evolution of tumor cell populations. *Science*, 194:23–38, 1976.

[67] PC Nowell and DA Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Science*, 142:1497, 1960.

[68] D Pinkel, R Segraves, D Sudar, S Clark, I Poole, and et al. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, 20(2):207–211, Oct 1998.

[69] Boone PM, Bacino CA, Shaw CA, Eng PA, and Hixson PM et al. Detection of clinically relevant exonic copy-number changes by array cgh. *Human Mutation*, 31(12):1326–1342, Nov 2010.

[70] AK Raap, RJ Florijn, LAJ Blonden, J Wiegant, JW Vaandrager, and et al. Fiber fish as a dna mapping tool. *Methods*, 9(1):67–73, 1996.

[71] R Redon, S Ishikawa, KR Fitch, L Feuk, GH Perry, and et al. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, Nov 2006.

[72] Genetics Home Reference. Color vision deficiency. online: https:ghr.nlm.nih.gov/condition/color-vision-deficiency, 2019.

[73] WRB Robertson. Chromosome studies. i. taxonomic relationships shown in the chromosomes of tettigidae and acrididae: V-shaped chromosomes and their significance in acrididae, locustidae, and gryllidae: Chromosomes and variation. *Journal of Morphology*, 27(2):179–331, Jun 1916.

[74] Simone Roeh, Peter Weber, Monika Rex-Haffner, Jan M. Deussing, Elisabeth B. Binder, and Mira Jakovcevski. Sequencing on the solid 5500xl system – in-depth characterization of the gc bias. *Nucleus*, 8(4):370–380, Jun 2017.

[75] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, and et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, Jul 2011.

[76] JD Rowley. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243:290–293, 1973.

[77] D.C. Rubinsztein, J. Leggo, R. Coles, E. Almqvist, V. Biancalana, J.J. Cassiman, K. Chotai, M. Connarty, D. Crauford, A. Curtis, D. Curtis, M.J. Davidson, A.M. Differ, C. Dode, A. Dodge, M. Frontali, N.G. Ranen, O.C. Stine, M. Sherr, M.H. Abbott, M.L. Franz, C.A. Graham, P.S. Harper, J.C. Hedreen, and M. R. Hayden. Phenotypic characterization of individuals with 30-40 cag repeats in the huntington disease (hd) gene reveals hd cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am. J. Hum. Genet.*, 59:16–22, 1996.

[78] R Saiki, D Gelfand, S Stoffel, S Scharf, R Higuchi, and et al. Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. 239(4839):487–491, Jan 1988.

[79] R Saiki, S Scharf, F Faloona, K Mullis, G Horn, and et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–1354, Dec 1985.

[80] F Sanger and AR Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–448, 1975.

[81] F Sanger, S Nicklen, and AR Coulson. Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.*, 74(12):5463–5467, December 1977.

[82] Melanie Schirmer, Umer Z. Ijaz, Rosalinda D'Amore, Neil Hall, William T. Sloan, and Christopher Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic Acids Research*, 43(6):e37–e37, Jan 2015.

[83] A Schneider. Untersuchungen über plathelminthen. *Jahresberichte der Oberhessischen Gesellschaft für Natur- und Heilkunde in Gießen*, 14:69–140, 1873.

[84] JP Schouten, McElgunn CJ, Waaijer R, Zwijnenburg D, and Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, 30(12):e57, 2002.

[85] Virginie Scotet, Ingrid Duguépéroux, Philippe Saliou, Gilles Rault, Michel Roussey, Marie-Pierre Audrézet, and Claude Férec. Evidence for decline in the incidence of cystic fibrosis: a 35-year observational study in brittany, france. *Orphanet Journal of Rare Diseases*, 7(1):14, 2012.

[86] A. J. Sehnert, B. Rhees, D. Comstock, E. de Feo, G. Heilek, J. Burke, and R. P. Rava. Optimal detection of fetal chromosomal abnormalities by massively parallel dna sequencing of cell-free fetal dna from maternal blood. *Clinical Chemistry*, 57(7):1042–1049, Apr 2011.

[87] M Simioni, F Artiguenave, V Meyer, IC Sgardioli, NL Viguetti-Campos, and et al. Genomic investigation of balanced chromosomal rearrangements in patients with abnormal phenotypes. *Molecular Syndromology*, 8(4):187–194, 2017.

[88] Michał Sobjanek, Magdalena Dobosz-Kawalko, Igor Michajlowski, Rafal Peksa, and Roman Nowicki. Segmental neurofibromatosis. *Advances in Dermatology and Allergology*, 6:410–412, 2014.

[89] Malte Spielmann, Darío G. Lupiáñez, and Stefan Mundlos. Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7):453–467, Apr 2018.

[90] PH Sudmant, T Rausch, EJ Gardner, RE Handsaker, A Abyzov, and et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, Sep 2015.

[91] WS Sutton. On the morphology of the chromosome group in brachystola magna. *Biological Bulletin*, 4:24–39, 1902.

[92] Bente A. Talseth-Palmer, Denis C. Bauer, Wenche Sjursen, Tiffany J. Evans, Mary McPhillips, Anthony Proietto, Geoffrey Otton, Allan D. Spigelman, and Rodney J. Scott. Targeted next-generation sequencing of 22 mismatch repair genes identifies lynch syndrome families. *Cancer Medicine*, 5(5):929–941, Jan 2016.

[93] Renjie Tan, Yadong Wang, Sarah E. Kleinstein, Yongzhuang Liu, Xiaolin Zhu, Hongzhe Guo, Qinghua Jiang, Andrew S. Allen, and Mingfu Zhu. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, 35(7):899–907, May 2014.

[94] JH Tjio and Levan A. The chromosome number of man. *Hereditas*, 42(1-2):1–5, 1956.

[95] Katarzyna Tutlewska, Jan Lubinski, and Grzegorz Kurzawski. Germline deletions in the epcam gene as a cause of lynch syndrome – literature review. *Hereditary Cancer in Clinical Practice*, 11(1), Aug 2013.

[96] J. C. Venter. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[97] P.-P. Verbeek. *Moralizing Technology – Understanding and designing the Morality of Things*. The University of Chigago Press, Chicago, 2011.

[98] F von Eggeling, M Freytag, R Fahsold, B Horsthemke, and U Claussen. Rapid detection of trisomy 21 by quantitative pcr. *Hum. Genet.*, 91:567–570, 1993.

[99] Bateson W, Saunders ER, and Punnett RC. Further experiments on inheritance in sweet peas and stocks: preliminary account. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 77(517):236–238, 1906.

[100] W Waldeyer. Über karyokinese und ihre beziehung zu den befruchtungsvorgängen. *Archiv für mikroskopische Anatomie*, 32:1–122, 1888.

[101] Jeffrey D. Wall, Ling Fung Tang, Brandon Zerbe, Mark N. Kvale, Pui-Yan Kwok, Catherine Schaefer, and Neil Risch. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, 24(11):1734–1739, Oct 2014.

[102] DG Wang. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, May 1998.

[103] JD Watson and FHC Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[104] A Weismann. *Das Keimplasma. Eine Theorie der Vererbung*. Gustav Fischer, Jena, 1892.

[105] Amy B. Wilfert, Arvis Sulovari, Tychele N. Turner, Bradley P. Coe, and Evan E. Eichler. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Medicine*, 9(1), Nov 2017.

[106] MHF Wilkins, AR Stokes, and HR Wilson. Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–740, Apr 1953.

[107] Daynna J. Wolff, Adam Bagg, Linda D. Cooley, Gordon W. Dewald, Betsy A. Hirsch, Peter B. Jacky, Kathleen W. Rao, and P. Nagesh Rao. Guidance for fluorescence in situ hybridization testing in hematologic disorders. *The Journal of Molecular Diagnostics*, 9(2):134–143, Apr 2007.

[108] C Yanofsky. Establishing the triplet nature of the genetic code. *Cell*, 128(5):815–818, Mar 2007.

[109] E. Yu and S. Sharma. *Cystic Fibrosis. [Updated 2018 Mar 20]. In: StatPearls*. Treasure Island FL: StatPearls Publishing, Available from: https://www.ncbi.nlm.nih.gov/books/NBK493206/, 2018.

List of Tables

List of Figures