Putting the pieces together:

The causes and consequences of *de novo* structural variation

Sjors Middelkamp

Layout by: Sjors Middelkamp
Cover design by: Ila van Kruijsbergen
Printed by: ProefschriftMaken | www.proefschriftmaken.nl
ISBN: 978-94-6380-446-2

Putting the pieces together: The causes and consequences of *de novo* structural variation

De stukjes vallen samen:
De oorzaken en gevolgen van *de novo* structurele variatie
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 17 september 2019 des middags te 2.30 uur

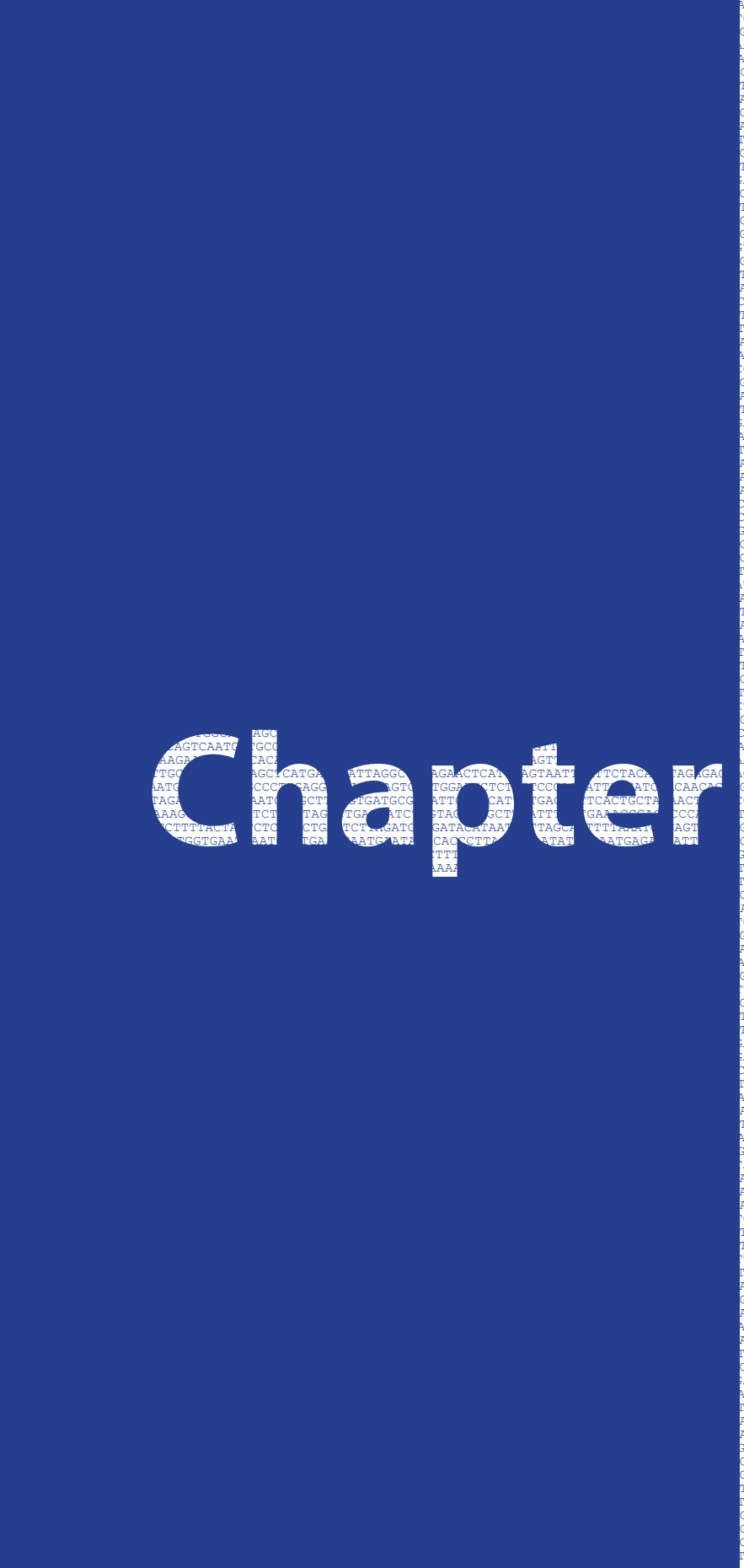
door

Sjors Harrie Antoon Middelkamp

geboren op 5 september 1989
te Hellendoorn

Promotor: Prof. dr. ir. E.P.J.G. Cuppen

Contents


Chapter 1

Introduction

1.1 Introduction

We all once started as a single cell: the zygote. From this single cell we developed into the ~37.200.000.000.000 cells we consist of today (Bianconi et al., 2013). This entire extraordinary development is programmed in DNA (deoxyribonucleic acid), the carrier of genetic information in all living organisms. DNA is the "code of life" consisting of long sequences of four core subunits called nucleotides or bases: adenine (A), thymine (T), guanine (G) and cytosine (C). These nucleotides form the typical double stranded helix structure which was first described in 1953 (Watson and Crick, 1953). The entire DNA code of one human, called a **genome**, contains two times 3.2 billion nucleotides that are stored on 23 pairs of **chromosomes**. Specific parts of genomes, the **genes,** contain the instructions to produce proteins, biomolecules built from amino acids that perform most of the biochemical reactions in cells (Figure 1.1). In total there are around 20,000 genes in the human genome (Frankish et al., 2019), coding for tens of thousands protein isoforms (Kim et al., 2014). Changes in the DNA sequence, called **mutations** or **variants**, can disturb normal protein activity and this can lead to diseases such as developmental disorders and cancer. Detection and interpretation of DNA variants have become important methods to determine the causes of diseases. DNA variants can be very small, consisting of a single base pair (**single nucleotide variant**), but they can also be very large, affecting millions of nucleotides. These large genomic rearrangements involving more than 50 nucleotides are called **structural variants**. In this thesis we studied the causes and consequences of *de novo* structural variants. We focus on germline structural variants that are present in each cell of the body and that can be passed on to offspring or occur *de novo* from one generation to the next.

1.2 Deciphering DNA codes by DNA sequencing

1.2.1 The early days of DNA sequencing

Our knowledge about genomes and their roles in development and disease has increased tremendously in the last decades. This extraordinary progress has been driven by revolutionary developments in **sequencing** technologies used to determine the sequence of DNA molecules (Heather and Chain, 2016; Shendure et al., 2017). The inventions of the chain terminator and the chemical cleavage methods in 1977 laid the foundations for the developments of most sequencing techniques in the following decades (Maxam and Gilbert, 1977; Sanger et al., 1977). DNA molecules are composed of two intertwined chains (strands) of complementary nucleotides. Nucleotides in one strand can form hydrogen bonds with specific nucleotides in the other strand according to base pairing rules; A normally pairs with T and C forms bonds with G. DNA molecules can be copied, or replicated, by proteins called DNA polymerases. For DNA sequencing, the two DNA strands are first separated ("denatured") into two single

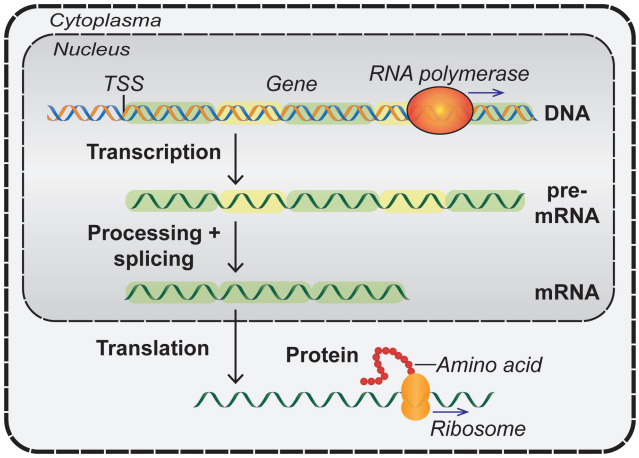


Figure 1.1 | Schematic overview of transcription and translation of the DNA code into amino acid sequences. *Human double stranded DNA is stored on 46 chromosomes, which are located in the nuclei of cells. The DNA sequences of genes can be transcribed to single-stranded pre-messenger RNA (pre-mRNA) molecules by RNA polymerases. RNA polymerases can be recruited to promoter regions upstream of the gene and initiate transcription from the transcription start site (TSS). Genes contain protein-coding sequences (exons, in green) but also non-coding sequences (introns, in yellow). These intronic sequences are removed from the mRNA by a process called splicing. The processed mRNA molecules are transported to the cytoplasm where they can be translated into an amino acid sequence by ribosomes. The amino acid sequence can be further processed into protein. This entire process is frequently referred to as the "central dogma of molecular biology".*

strands. These single DNA strands serve as templates for synthesis of complementary daughter DNA sequences by DNA polymerases. **Sanger sequencing**, as well as most other DNA sequencing methods, makes use of DNA polymerases to elongate DNA fragments with a parental strand as template (Figure 1.2) (Sanger et al., 1977). One Sanger sequencing reaction can be used to determine the sequence of around 300 to 1000 nucleotides and such a sequence is called a **read**. Overlap between the ends of the short sequences can be used to stitch (assemble) the short sequences together into longer stretches of continuous sequences, an approach that is used in "shotgun" sequencing of many small fragments (Weber and Myers, 1997). Especially in the early days, Sanger sequencing involved many time-consuming steps and therefore it was sometimes doubted whether it could be feasible to sequence an entire human genome consisting of 6.4 billion nucleotides. For example, automated Sanger sequencers could only read 1000 nucleotides per day in 1987 (Shendure et al., 2017). However, technical improvements such as the developments of fluorescently labelled terminator nucleotides, improved polymerases, improved (capillary) electrophoresis, automation and computational methods greatly improved the efficiency of Sanger sequencing, making it possible to determine the sequences of increasingly longer DNA molecules (Heather and Chain, 2016; Shendure et al., 2017). Finally, in 2001, two large consortia

of researchers (the public **Human Genome Project** (HGP) and the private Celera) succeeded in determining the nearly complete sequences of two human genomes (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). The revised version of the **reference genome** finished by the HGP in 2004 would form the basis for much of the biomedical research in the following decade (International Human Genome Sequencing Consortium, 2004). The tremendous undertaking of sequencing the first human genome, which had cost US$3 billion in total, started a new era in biology 50 years after the discovery of the structure of DNA: **the genomic era** (Guttmacher and Collins, 2003).

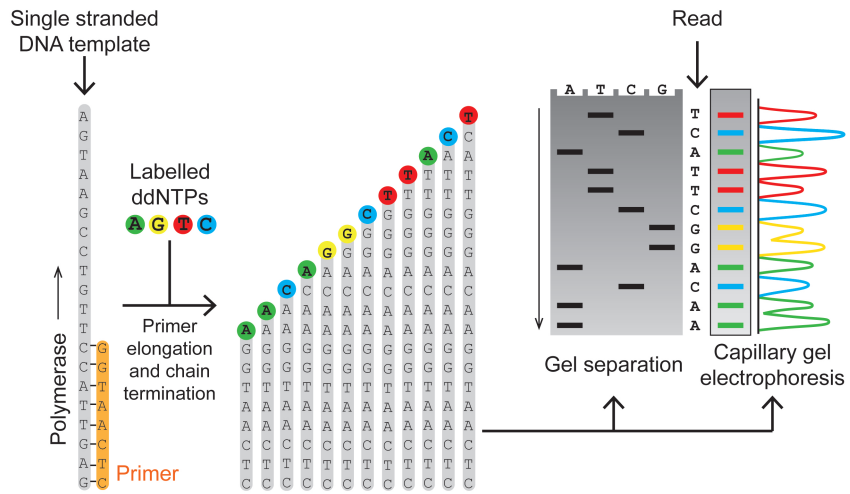


Figure 1.2 | Overview of the Sanger DNA sequencing method. *DNA molecules are first fragmented and denatured into single strands. Primers that can bind to sequences specific for the DNA fragment that needs to be sequenced are added to the mix. DNA polymerase can subsequently elongate the DNA sequence from the primer by adding nucleotides (dNTPs) complementary to the template sequence. However, in Sanger sequencing chain-terminating ddNTPs are also added to the reaction (Sanger et al., 1977). If such a ddNTP is incorporated, the DNA strand cannot longer be elongated by the polymerase, resulting in a DNA fragment of a specific length ending with a specific labelled ddNTP. Initially ddNTPs were radioactively labelled and four separate reactions had to be performed (one for each nucleotide). Later, ddNTPs labelled with different fluorophores were developed and these made it possible to perform the sequencing in one reaction (Prober et al., 1987; Smith et al., 1986). The labelled DNA fragments of different sizes are subsequently separated by size using (capillary) gel electrophoresis. Automated Sanger sequencing machines can perform up to 384 sequencing reactions in parallel, generating sequencing reads of 300-1000 nucleotides. Sanger sequencing is still routinely used nowadays to sequence short DNA fragments.*

1.2.2 The next generation sequencing revolution

The availability of the reference genome opened many new possibilities to study the contents, such as genes, and structure of the human genome. DNA sequencing itself however remained an expensive method (around US$1 per read of 600-700 basepairs in 2004 (Shendure et al., 2017)), making it still infeasible to sequence and compare many genomes. Sanger sequencing had been virtually the only sequencing technique

that had been performed for 25 years, but completely new techniques were in development in the early 2000's. The first revolutionary **next-generation sequencing** (NGS) platforms were published in 2005 (Margulies et al., 2005; Shendure et al., 2005). NGS methods can be used to sequence millions to billions of DNA molecules in one reaction (Figure 1.3) (Goodwin et al., 2016). This massively parallel sequencing is a major advantage over Sanger sequencing in which only one DNA molecule can be sequenced per reaction. The availability of the reference genome made it possible to determine the genomic position of sequencing reads by comparing the sequences of the reads to the reference genome, a process that is called mapping (Li and Durbin, 2009). It was not necessary anymore to build genome sequences from scratch (which is called *de novo* assembly (Chaisson et al., 2015)) and this resequencing makes data processing much faster. Initially there was fierce competition between several sequencing platforms that were based on different technologies (Goodwin et al., 2016). This competition drove the extraordinary developments of the NGS platforms and lead to rapidly increasing sequencing output and declining sequencing costs. Sequencing of a genome still cost around US$10 million in 2007 and only 5 years later these costs were reduced to just US$6000 (Wetterstrand, 2019). Illumina, a company that acquired sequencing technology developed by Solexa (Bentley et al., 2008), became the dominant player in the sequencing market and the vast majority of sequencing is performed on Illumina sequencers nowadays. The spectacular drop in sequencing costs (currently sequencing a whole genome costs around €1000) made it possible to sequence genomes at a large scale. These days sequencing itself is no longer the biggest challenge in genome research: the main challenge is to interpret the vast amounts of generated data.

1.3 Genetic variation in humans

1.3.1 Types of variation in human genomes

DNA variants can disturb the normal development of cells, tissues and organisms. However, genetic variation between human genomes is very common and not every genetic variant will lead to disease. Common variants detected in multiple healthy individuals are unlikely to cause severe congenital disorders (although they may form risk factors for more common disorders such as cancer), but rare variants only detected in patients with similar congenital disorders are likely to contribute to disease. It is therefore important to sequence many genomes to determine the frequencies of variants in the population and allow for the identification of potentially pathogenic variants and dissection of their potential role in disease. NGS technologies and the reduced sequencing costs made it possible to routinely sequence many genomes. In 2008, the 1000 Genomes consortium was launched and it sequenced more than 100 human genomes by 2010 (1000 Genomes Project Consortium et al., 2010),

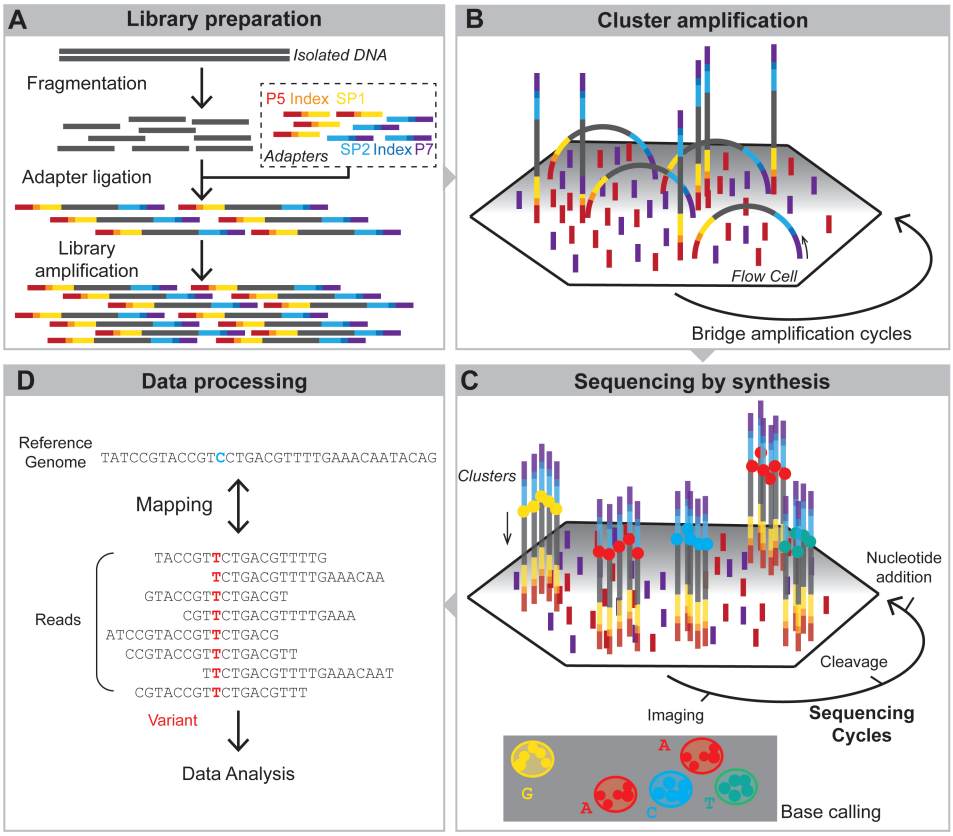


Figure 1.3 | Basics of massively parallel DNA sequencing on Illumina sequencers. ***(A)*** *DNA molecules need to be processed (prepared) prior to sequencing. First, DNA molecules are fragmented into smaller fragments of a specific size (depending on the application). Subsequently sequencing adapters (containing adapter and sequencing primer (SP) binding sites and barcodes/indices specific for each sample) are pasted (ligated) to the ends of the short fragments. The processed DNA molecules together form a sequencing library. The library is frequently amplified before sequencing and specific sequences (such as exonic sequences) can be captured from the library if necessary.* ***(B)*** *Prepared libraries are loaded on a flow cell which contain attached short DNA sequences that can bind (hybridize) to the DNA molecules in the library. Each hybridized DNA molecule is amplified in a process called bridge amplification. These copied molecules form millions of clusters of DNAs with identical sequence on the flow cell.* ***(C)*** *Sequencing-by-synthesis is performed after the clusters have been generated. One type of "reversible terminator" fluorescent nucleotides is incorporated in each cluster by polymerases. These nucleotides prevent further elongation of the template and the fluorescent signal emitted from each cluster is measured by microscopy imaging. Subsequently the terminating fluorophore is removed and the next base in template molecules can be determined in a new sequencing cycle. The images are translated to basepairs (base calling) and millions of sequencing reads are generated by performing multiple sequencing cycles.* ***(D)*** *The genomic positions of the reads are determined by comparing the read sequence to the reference genome (alignment or mapping). The mapped reads can be used for downstream applications, such as the detection of genetic variants. This figure is based on the images in (Goodwin et al., 2016) and "An introduction to Next-Generation Sequencing Technology" from Illumina (https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf).*

more than a 1000 by 2012 (1000 Genomes Project Consortium et al., 2012) and more than 2500 genomes by 2015 (1000 Genomes Project Consortium et al., 2015). Differences between genomes range from single nucleotide variants (SNVs) to very large structural variants (SVs) which can affect millions of basepairs. In addition, it is possible that cells contain abnormal numbers (more or less than 46 in most human cells) of whole chromosomes, which is called **aneuploidy** (Nagaoka et al., 2012). A typical human genome has around 3.5 million single nucleotide differences compared to the reference genome (1000 Genomes Project Consortium et al., 2015). In addition, a human genome has on average more than 20,000 structural variants that are larger than 50 basepairs (Chaisson et al., 2019; Nelson et al., 2019). There are several classes of structural variants of widely varying sizes including deletions, duplications, insertions, inversions and translocations (Figure 1.4) (Feuk et al., 2006). Deletions and duplications lead to respectively losses or gains of DNA which are also called **copy number variants** (CNVs) (Zarrei et al., 2015). Insertions occur when a fragment of DNA is moved to a different position in the genome. Translocations are exchanges between parts of different chromosomes. Genomic sequences can also be inverted. Translocations and inversions usually do not lead to a change in the amount of DNA and are therefore also called **balanced rearrangements** (Redin et al., 2017). Due to their large size structural variants involve more bases (roughly 11 megabases (Mb) per individual) than single nucleotide variants (Nelson et al., 2019; Sudmant et al., 2015). Deletions and insertions are the most common SV classes in the human genome (Chaisson et al., 2019; Nelson et al., 2019).

1.3.2 Complex genomic rearrangements

SVs can also form complex combinations of variable severity. The most complex SVs affect large genomic regions usually on multiple chromosomes. The mechanisms leading to such complex SVs are grouped under the term **chromoanagenesis** ("chromosome rebirth") (Holland and Cleveland, 2012). An important discriminating factor between different classes of chromoanagenesis is the involvement of DNA replication, which can lead to gains of DNA. DNA fragments can be multiplied many times if a crisis occurs during DNA replication, leading to "**chromoanasynthesis**" (chromosome reconstitution) rearrangements (Liu et al., 2011). Parts of chromosomes can also be shattered and subsequently reassembled leading to complex rearrangements in a process called "**chromothripsis**" (Stephens et al., 2011). In contrast to chromoanasynthesis, DNA replication is not majorly involved during chromothripsis and this process therefore mainly leads to deletions and balanced rearrangements such as translocations and inversions (Figure 1.4D). Chromothripsis rearrangements have been both found in cancer genomes and in genomes of patients with congenital disorders (Zepeda-Mendoza and Morton, 2019). There are differences in SV characteristics and patterns between "somatic" (occurring in only some of the cells of

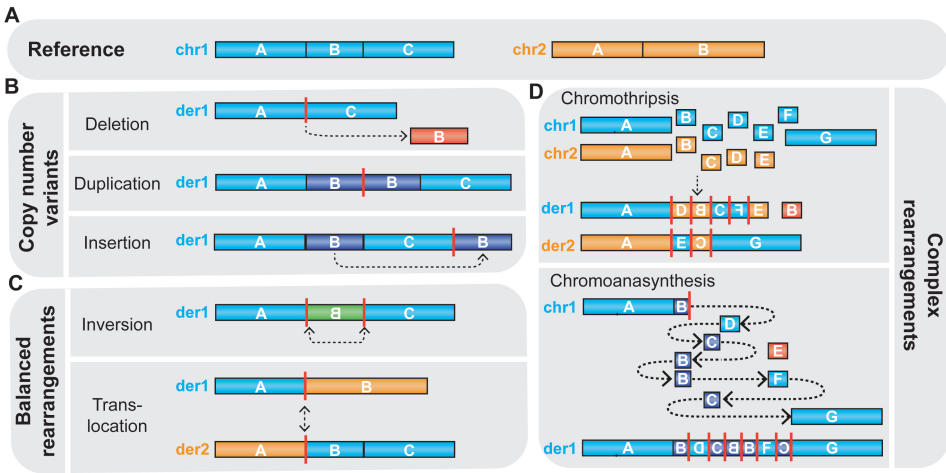


Figure 1.4 | Schematic examples of structural variants. (A) *Schematic representation of two chromosomes.* ***(B)*** *Copy number variants are SVs that lead to gains (duplications or insertions) or losses (deletions) of DNA sequences. Insertions are duplicated sequences that are placed somewhere else in the genome.* ***(C)*** *Balanced rearrangements include inversions and translocations that do not lead to major gains or losses of DNA. Balanced SVs, especially translocations, are relatively rare compared to copy number variants.* ***(D)*** *Complex genomic rearrangements including multiple SVs can be caused by several mechanisms. Chromothripsis is the shattering of parts of one or more chromosomes. Subsequent repair of the breaks can lead to complex genomic rearrangements involving multiple breakpoint junctions frequently on multiple chromosomes. Some fragments may be lost during repair, leading to deletions (fragment B). In contrast to chromothripsis, chromoanasynthesis involves copy number gains caused by errors during replication. Fork-stalling and template switching (FoSTes) is a mechanism that can cause chromoanasynthesis rearrangements (Hastings et al., 2009).*

the body, not contributing to the genomes of offspring) and "germline" chromothripsis (occurring in gametes of the parents and/or the early embryo and therefore affecting every cell of an organism) and it is not clear whether these rearrangements are caused by the same mechanism (Kloosterman and Cuppen, 2013). Cells have to replicate their chromosomes before they can divide into two daughter cells. Mitosis is the process in which the replicated chromosomes are separated from each other and distributed between the daughter cells. Chromosome segregation errors can occur during mitosis and it has been shown that the consequences of such mitotic errors can ultimately lead to chromothripsis. Sometimes a chromosome can lag behind when the other chromosomes are being pulled (segregated) to the daughter cells. Such a lagging chromosome can become damaged (Janssen et al., 2011) and/or may be excluded from a newly formed nucleus and form its own small micronucleus. Chromosomes in micronuclei are vulnerable for massive DNA damage and may undergo shattering (Crasta et al., 2012; Hatch et al., 2013; Ly et al., 2016, 2019; Zhang et al., 2015). Whether this also occurs in gametes or early embryos is not known, but it is known that micronuclei are common in early mammalian embryos (Chavez et al., 2012; Daughtry et al., 2019; Vázquez-Diez et al., 2016). Complex germline genomic rearrangements

are relatively rare, but they can have devastating effects on embryonic development (Pellestor et al., 2011; Zepeda-Mendoza and Morton, 2019).

1.4 Causes of *de novo* structural variation

1.4.1 Inheritance of genetic variants

Structural variants are common in human genomes, but not all of them have an effect on the development of the individual. Many genetic variants do have no or only a mild effect on the observable characteristics (called **phenotype**) of an individual, such as height and eye color. However, SVs affecting important developmental genes may disturb normal development and therefore lead to congenital disease phenotypes. An individual inherits one haploid genome (one copy of each chromosome) from their father and one haploid genome from their mother, which together form a diploid genome (two copies of each chromosome) during fertilization. Because an individual has two copies of each chromosome (except for chrX and chrY in males), each gene is also present twice in the genome. In many cases, if one copy of a gene (also called an **allele**) is affected by a genetic variant, the other copy can compensate. Healthy parents can be carriers of **heterozygous variants**, meaning that they have one normal and one affected allele. It is possible that the offspring inherits two affected alleles if both parents are carrier of similar heterozygous variants. In this case, the offspring is homozygous for the variant and there can be insufficient compensation for the loss of the gene(s). These **homozygous variants** can cause disease if the genes are important in embryonic development.

1.4.2 *De novo* SVs are caused by erroneous DNA double-strand break repair in the germline

In addition to inheriting variants already present in the parents, it is also possible that embryos gain new variants that are not present in their parents. These ***de novo* variants** can arise in the parental oocytes, sperm cells, precursors of these cells or in the early embryo. DNA is a very stable molecule, but DNA molecules can be damaged by endogenous (such as collapsing replication forks or topoisomerases) and exogenous sources (such as radiation) (Mehta and Haber, 2014). Some damage can lead to a break of both DNA strands. These **double-strand breaks** (DSBs) can have severe consequences for the integrity of the genome and therefore cells have developed many mechanisms to repair such breaks. It is estimated that there are roughly ten DSBs in a cell each day (Lieber, 2010). However, repair of these breaks is not always perfect and this can lead to formation of new structural variants. Non-homologous end joining (NHEJ) and homologous recombination (HR) are the two major DSB repair pathways with several subclasses in mammals. There are many differences between these pathways including differences in the properties of the breaks that can be

repaired, the enzymes that are used, timings in the cell cycle, cell types in which they are dominant and the speed and precision of repair (Her and Bunting, 2018). The repair mechanism that created an SV can later frequently be deduced by determining the genomic sequence around the breakpoint junction, because the repair pathways require different sequence characteristics and frequently create a specific molecular scar during repair.

Most spontaneous DSBs occur during DNA replication (Syeda et al., 2014). After DNA replication, an intact copy (a homolog on a sister chromatid) of the broken DNA molecule is present that can be used as a template for precise DSB repair by **homologous recombination** (Mehta and Haber, 2014; Syeda et al., 2014). Although homologous recombination is usually an error-free mechanism to repair DSB, errors can be made if a wrong DNA sequence is recognized as homolog and used as template for repair. This process, which is called non-allelic homologous repair (NAHR), can occur between large repetitive regions (low copy repeats) in the genome (Liu et al., 2012). Because these regions are only present at specific locations in the genome, the SVs caused by NAHR are similar and recurrent between individuals (Carvalho and Lupski, 2016). These recurrent *de novo* SVs at specific locations in the genome can lead to over 40 different **recurrent genomic disorders** (Vissers and Stankiewicz, 2012; Watson et al., 2014).

The presence of homologous sequences is essential for homologous recombination, but these sequences are only present after DNA replication. Most DSBs are repaired by more flexible **non-homologous end joining** mechanisms which do not require extensive homology (Her and Bunting, 2018). Instead, components of NHEJ pathways first stabilize the two ends of the broken DNA, subsequently process the ends if necessary and finally ligate the processed ends together (Chang et al., 2017). This frequently leads to losses or insertions of multiple nucleotides at the breakpoints. Additionally, the wrong broken DNA ends may be ligated together and SVs can be formed if multiple breakpoints are present. Although not strictly necessary, NHEJ frequently makes use of short, microhomology sequences of 1-4bp that are similar at both ends of the breaks. Two other, less well understood repair pathways that make use of different enzymes, alternative (or microhomology mediated) end-joining (a-EJ) and single strand annealing (SSA), require longer homologous sequences of respectively 2-20bp and >20bp (Chang et al., 2017). The amount of homology detected around a breakpoint junction can therefore give insight in the repair mechanism involved in repairing a break-junction and the creation of an SV.

1.4.3 *De novo* SVs can be formed at different stages in the germline

Genomes of children contain on average between 50 and 100 *de novo* single nucleotide variants that are not present in the genomes of their parents (Acuna-Hidalgo et al., 2016). *De novo* SVs are much rarer and it is estimated that roughly one in five children

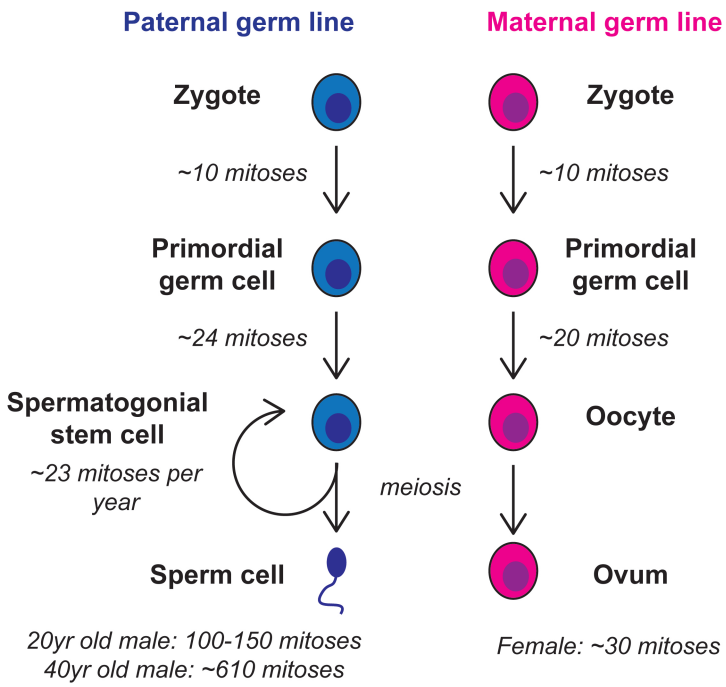


Figure 1.5 | Schematic overview of the mitoses in the male and female germlines. *The pool of gametes in male germline is continuously replenished after the start of puberty by mitoses of spermatogonial stem cells. The required mitoses and DNA replications make the paternal germline susceptible to DNA damage and mutations, which is one of the major reasons for relatively high amount of de novo variation on paternally inherited chromosomes. Additionally, sperm cells are vulnerable for DNA damage during maturation and transport after completing meiosis, which can also lead to the formation of de novo SVs. This figure is based on information from (Rahbari et al., 2015)*

is born with a *de novo* SV (Brandler et al., 2018; Collins et al., 2019; Kloosterman et al., 2015). Interestingly, most of the *de novo* SVs and SNVs are present on the chromosomes inherited from the father (Acuna-Hidalgo et al., 2016; Brandler et al., 2018; Hehir-Kwa et al., 2011; Kloosterman et al., 2015). A major difference between the male and female germline is the number of involved mitoses. The pool of male gametes is constantly replenished after the start of puberty by mitoses of spermatogonial stem cells (roughly ~23 divisions per year), whereas the number of female gametes becomes fixed early in development around birth (Figure 1.5). Errors can occur during each round of spermatogonial stem cell DNA replication, leading to increasing numbers of *de novo* variants (Rahbari et al., 2015). The number of mutations in spermatogonial stem cells increases with age and therefore genomes of offspring of older fathers usually contain more *de novo* single nucleotide variants (Acuna-Hidalgo et al., 2016). Such a paternal age effect has also been suggested for *de novo* SVs (Hehir-Kwa et al., 2011), but the number of *de novo* SVs studied in detail so far may have been too low to draw strong conclusions about such an age effect for SVs yet. Another reason for the elevated

number of *de novo* SVs on paternally inherited chromosomes is the vulnerability of sperm cells for double stranded breaks during post-meiotic maturation, condensation of the genome, transport to an oocyte and unpacking of the genome after fertilization (González-Marín et al., 2012; Sakkas and Alvarez, 2010). Sperm cells gradually lose the capacity to repair such breaks during their maturation and therefore these breaks can only be repaired in the zygote after fertilization.

De novo germline variants can also be induced after fertilization. Such variants may only end up in a portion of the cells of an individual, depending on the stage of development in which they arise (more cells are likely to be affected if a variant arises in one of the first cell divisions). The presence of multiple cell lineages with different genomes within an individual is called mosaicism. Although the first cell divisions are essential for embryonic development, these early cell division are often surprisingly error-prone. Many errors of chromosome segregation occur during the first cell division, which frequently leads to mosaicism in early embryos. This mosaicism is one of the causes of the high mortality of human embryos, which is thought to affect up to 70% of all human embryos (Jarvis, 2016; McCoy, 2017). The precise causes of the genomic instability in human embryos is not well understood, but it is likely that this instability can also lead to the formation of *de novo* SVs in the germline and thereby play a role in causing developmental disorders.

1.5 The role of structural variants in developmental disorders

1.5.1 The contribution of genetic variation to neurodevelopmental disorders

On one hand, the frequencies of *de novo* variants seem relatively low compared to the number of common variants in the genome. On the other hand, these *de novo* variants have not been subjected to stringent evolutionary selection and therefore they have a higher chance to affect genes essential for embryonic development. Many genetic disorders are caused by *de novo* variants (Veltman and Brunner, 2012). Disorders caused by a variant in one of the two alleles, which is usually the case with *de novo* variants, are called **dominant disorders**. In contrast, **recessive disorders** only occur when both alleles are affected. Sometimes two alleles are affected by different variants (for example a deletion on one allele and a pathogenic SNV on the other allele), which together disrupt the function of the gene. Such combined variants are called **compound heterozygous variants**. Thus, several modes of inheritance exist and an important goal of clinical genetics is to discover which model likely explains the phenotype of a patient.

Although the vast majority of genetic variants does not have an influence on a phenotype, some variants can have a major impact. The effects of such genetic variants depend on which genes are affected and to what extent the functions of

these genes are disrupted. Thousands of genes with specific functions are involved in different cells at different moments in embryonic development. Variants in these genes can have very different phenotypic outcomes and therefore there are many different genetic disorders. Most of these disorders are rare, meaning they occur in one individual per 2000 individuals at most. It is estimated that there are between 6000-8000 different rare diseases and most of these are caused by genetic defects (Hartley et al., 2018). Together these disorders, although individually rare, affect millions of people (the estimated number of patients is more than 30 million in Europe alone (EURORDIS, 2019)). It is estimated that around 2-5% of all children are born with a neurodevelopmental disorder (Wright et al., 2018). **Neurodevelopmental disorders** form a broad group of disorders disturbing the development of the central nervous system. This can lead to mental disorders such as intellectual disability, autism spectrum disorders, neuropsychiatric disorders and motor function disorders. The severity of disorders varies widely and frequently involves other congenital abnormalities such as skeletal phenotypes (including cleft palate, hand or foot abnormalities) as well (which are grouped under "multiple congenital abnormalities and/or intellectual disability or MCA/ID). These disorders usually have an enormous impact on the lives of the children and their families. It is important to determine a potential genetic cause of the disorder, because knowing the cause of the disease can help with treatment in some cases and may help to give a prognosis for disease progression and possible complications. A genetic diagnosis can also be important for family planning. If a pathogenic variant is homozygous in a patient, it is likely that both parents are heterozygous carriers of the variant and there is a 25% chance that a possible next child will also inherit both pathogenic variants. In such a case prenatal screening for the variant becomes an option if the parents want another child. In addition, this means that the variants are likely also present in more family members and it can be helpful to screen them for presence of the variant and warn them for the possible consequences. If a variant is *de novo* and only present on one allele of the patient, it is much less likely that a next child will get the variant. For these reasons one of the main goals of clinical genetics is to identify the disease-causing (pathogenic) genetic variants in patients.

1.5.2 Traditional methods to detect structural variants in clinical genetics

Next generation sequencing drastically changed clinical genetics. Traditionally, geneticists could only screen for variants in a few genes using Sanger sequencing, which meant that genetic testing could only be done in a targeted and strongly hypothesis-based fashion for the most common genes known for a certain disorder. In addition, low resolution cytogenetic tests, techniques that determine the structure of chromosomes usually by microscopy-based visualization, could be performed to determine large megabase-sized chromosomal rearrangements and translocations. Mainly inherited variants occurring within families could be studied, but these variants

only cause a fraction of genetic disorders. In the early 2000's **microarray** techniques including ArrayCGH and SNP arrays became more commonly used to detect deletions and duplications (Speicher and Carter, 2005). These techniques do not determine the sequence of DNA, but they can identify gains and losses of pieces of DNA by comparing ratios between differently labelled reference and patient samples. Several different ArrayCGH and SNP array platforms of varying resolution and costs are available, but most routinely used arrays can detect CNVs larger than 10 kilobases (Kb) (Alkan et al., 2011; Pinto et al., 2011). The microarrays made unbiased, relatively high-resolution (compared to karyotyping) study of CNVs in the genomes of individual patients possible (Miller et al., 2010). Because array-based copy number profiling cannot be used detect single nucleotide variants and balanced SVs, array-based analyses are usually complemented with multiple other genetic tests if no pathogenic variant is detected. Before the rise of NGS techniques, the genetic cause of developmental disorders could typically only be determined in minority (~15-20%) of the cases (Vissers et al., 2015; Wright et al., 2018). This success rate was strongly boosted by routine introduction of NGS-based approaches.

1.5.3 Whole exome sequencing greatly improves diagnostic yield

Technically whole genome sequencing (WGS) can be used to determine the sequence of an entire genome including both balanced and unbalanced SVs and SNVs and thus could be used as a "one-test-fits-all" in clinical genetics. Because WGS can be used to screen the entire genome for variants, WGS can be used as an unbiased, hypothesis-free **genomics-first** approach (Stessman et al., 2014, 2016). However, despite the phenomenal cost reductions in recent years, WGS (including all involved activities such as bioinformatics) is still relatively expensive. In addition, handling the huge amount of produced information, of which a large part is not fully understood, is sometimes seen as a difficult challenge. Many variants of unknown significance are detected by WGS. Furthermore, variants may be discovered that are not related to the disorder of the patient, but that are associated with other, sometimes late-onset, diseases such as cancer. Such findings are called unsolicited or incidental findings and there is still much ethical debate whether such findings should be reported to the patient and their families, especially when there are no early treatment options (Wright et al., 2018). For all these practical and ethical reasons, many clinical genetics laboratories only sequence or analyse selected genes (gene panels) or only the protein-coding parts (exons) of the genome. In **exome sequencing**, first described in 2007, protein-coding sequences are targeted before sequencing by PCR or capture probes (Choi et al., 2009; Ng et al., 2009; Wood et al., 2007). Only ~1.2% of the genome (roughly two times 34 million basepairs) consists of exons and therefore relatively little sequencing is required to cover the entire exome. Most currently known pathogenic genetic variants are located in the exome, partly because pathogenic variant discovery has focused on

protein-coding regions (Zappala and Montgomery, 2016). The introduction of WES in genetic diagnostics has been a great success in the discovery of novel disease genes as well as in identifying pathogenic variants in patients and improving diagnostic yield (Vissers et al., 2015). The sequences of hundreds of thousands exomes have now been published (for example by the Exome Aggregation Consortium (ExAC)), providing an excellent view of both benign and potential pathogenic genetic variation within genes (Lek et al., 2016). Variants in around 1700 genes have now been associated with rare disease (DECIPHER, 2019; Wright et al., 2018), including ~700 genes involved in intellectual disability (Vissers et al., 2015). By combining WES with arrays around 50% of the patients with neurodevelopmental disorders can be diagnosed (Wright et al., 2018). In ~40% of cases a (likely) pathogenic SNV is found (McRae et al., 2017) and in around 10-15% a pathogenic SV is identified (Cooper et al., 2011; Hochstenbach et al., 2011; Kaminsky et al., 2011; Miller et al., 2010; Wright et al., 2018). This is an impressive leap forward in diagnostic yield compared to only few years ago, but it remains important to find a cause for the 50% of the patients that do not receive a diagnosis.

1.5.4 Detection of variants by whole genome sequencing

WGS has the potential to detect relevant variants missed by WES and array and may be used to further improve the diagnostic yield (Gilissen et al., 2014). Although WGS outperforms other methods in detecting pathogenic variants (Belkadi et al., 2015; Lelieveld et al., 2015; Meienberg et al., 2016; Stavropoulos et al., 2016; Trost et al., 2018), there are still some challenges. WGS is still performed mostly in a research setting and especially bioinformatic tools are still evolving rapidly. Single nucleotide variants are relatively easy to detect in WGS data and the **Genome Analysis Toolkit** (GATK) is standardly used to detect germline SNVs (McKenna et al., 2010). Detection of SVs is more challenging for several reasons. One issue is that many structural variants are located around repeated regions that are present in a large part of the genome. Because of the abundance of repeats in the genome, it can be difficult to determine the exact genomic position of a sequencing read if it covers such a repeated sequence. Most NGS techniques generate relatively short reads (nowadays generally between 75 and 300 basepairs) compared to Sanger sequencing (300-1000 basepairs). The chance that a read overlaps at least partially with a unique sequence in the genome is larger if the read is longer and this makes it easier to determine the genomic position of the read. One important improvement of NGS techniques, especially for SV detection, is **paired-end sequencing**. In paired-end sequencing, both ends of a DNA molecule are sequenced, effectively doubling the read length (but also the sequencing costs) (Korbel et al., 2007). Combining two reads (which are called mates) per DNA molecules also improves mapping to the reference genome, because the coordinates of both mates can be determined if just one of them overlaps a unique region. Paired-end sequencing enables three main strategies to detect ("call") SVs in WGS data: read

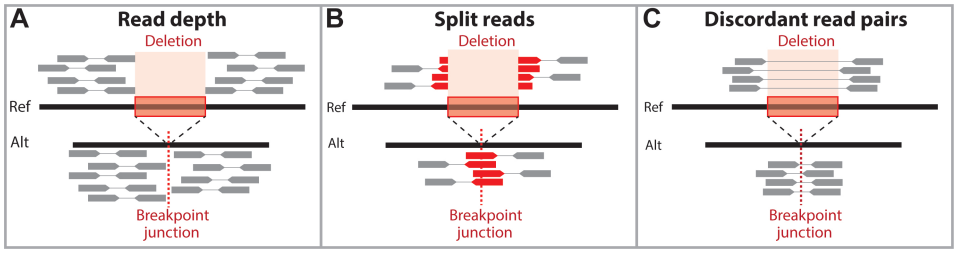


Figure 1.6 | Approaches to detect structural variants in next generation sequencing data. ***(A)*** *Losses or gains of DNA fragments leads to a respectively reduction or increase of the number of reads mapping to the fragment, leading to local changes in the sequencing coverage. Read-depth methods can measure these coverage variations and predict the copy number states in a wide variety of genome sequencing datasets (even WES and single cell sequencing datasets). In addition, sequencing depth of single nucleotide variants overlapping with the SVs can be used to determine the copy number (for example, if the copy number state of a locus is 3 instead of 2 due to a duplication, the ratio between reference and alternative SNVs in the duplication should be 2 to 1 instead of 1 to 1). These approaches are limited to detection of deletions and multiplications.* ***(B)*** *Reads that overlap an SV breakpoint junction will appear to be split if they are mapped to the reference genome. Split-read methods greatly benefit from longer reads (the longer the read, the more chance it overlaps a breakpoint junction and still can be mapped to the reference genome). They can be applied to both single-end and paired-end sequencing data.* ***(C)*** *Paired-end sequencing generates a pair of reads (mates) from a single DNA molecule. The expected distance between these mates can be predicted by generating DNA molecules from a specific insert size during library preparation. For example, if the insert size is 600bp and the read length is 2x150bp, the distance between the two mates should be roughly 300bp. However, if the reads overlap an SV, the distance between the two mates mapped to the reference genome deviates from this 300bp. The mates even map to different chromosomes if there is a translocation. Many different algorithms that make use of one or more of these approaches are available and currently a mix of these algorithms has to be used to detect all types of structural variation.*

depth, discordant-read pairs and split-reads (Figure 1.6) (Alkan et al., 2011). Many different software tools based on these approaches have been developed, but there is no golden standard SV detection method yet (like GATK is for SNV detection). Currently, it is still necessary to apply multiple SV callers to detect all classes of SVs. Most callers generate many false positive calls and further filtering is necessary. Validation of potential pathogenic SV candidates remains necessary and can be performed in the lab using PCR amplification and Sanger sequencing of the breakpoint junctions. Standard short read (2x 150bp) WGS can capture most clinically relevant SVs (Collins et al., 2019), but many SVs (mostly repetitive element insertions of 50-2000 basepairs and inversions over 50Kb) remain difficult to detect (Chaisson et al., 2019; Nelson et al., 2019). It has been shown that combinations of new sequencing methods can detect over 20,000 SVs in a human genome of which over half are not detected by standard WGS (Chaisson et al., 2019; Nelson et al., 2019). These new technologies include long-read sequencing developed by PacBio and Oxford Nanopore, which can generate reads of over 10,000 basepairs (Deamer et al., 2016; Eid et al., 2009). Such long reads can be especially beneficial for SV detection and these long-read technologies may therefore become the golden standard to detect SVs. The technologies are rapidly

evolving, but they still remain more expensive than Illumina short-read sequencing and they still suffer from relatively high error rates. This makes them less suited for the detection of SNVs, which are the cause of most genetic disorders. Short-read WGS can detect most genetic variants and already outperforms all other commonly used genetic tests. Nevertheless, there is still room for improvement, especially in detection and filtering of SVs. Software for analysis of WGS data is still rapidly improving and, with the declining costs of sequencing, WGS is likely to become the standard clinical genetic test in the near future.

1.6 The molecular consequences of structural variants

1.6.1 Direct effects of structural variants on genes

Another challenge, in addition to the identification of disease-causing genetic variants, is the interpretation of these variants. The pathogenicity of variants is mostly determined based on the recurrence of the variant or mutated gene in multiple individuals with similar phenotypes and the absence of the variant in unaffected individuals. The precise molecular mechanisms leading to the phenotype are often not known. SVs can cover large genomic regions, thereby affecting many genes, making it more difficult to solely rely on recurrence. There are several direct and indirect ways SVs can influence the RNA expression of genes (Weischenfeldt et al., 2013). Deletions and duplications can lead to respectively losses and gains of gene copies, which can lead to decreases or increases of RNA expression levels of these genes (Figure 1.7A,B). SVs can also affect portions of genes, which can lead to gene truncations. Depending on the precise location of the breakpoint in the gene, such a truncation can lead to shorter RNA or even non-functional RNA that is degraded. There is a chance that fusions between genes are formed if multiple genes are truncated (Figure 1.7C). Fusions can lead to the formation of entirely new proteins or change the activity of one of the fused fragments, but frequently they are also not functional. Changes in the levels or sequences of RNA can be determined by RNA sequencing (Wang et al., 2009). One other major advantage of next generation sequencing which has not yet been discussed is the flexibility of library preparation. This flexibility makes it possible to use sequencing for many purposes such as the study of interactions between DNA and proteins, the 3D structure of DNA, RNA expression. A wide range of NGS-based applications such as RNA-seq, ChIP-seq, DNase-seq, ATAC-seq and chromatin conformation capture methods have been developed (Karczewski and Snyder, 2018; Schmitt et al., 2016; Zentner and Henikoff, 2014). These techniques have been used to gain more insights in the indirect effects of SVs on genes which are more difficult to detect than direct effects.

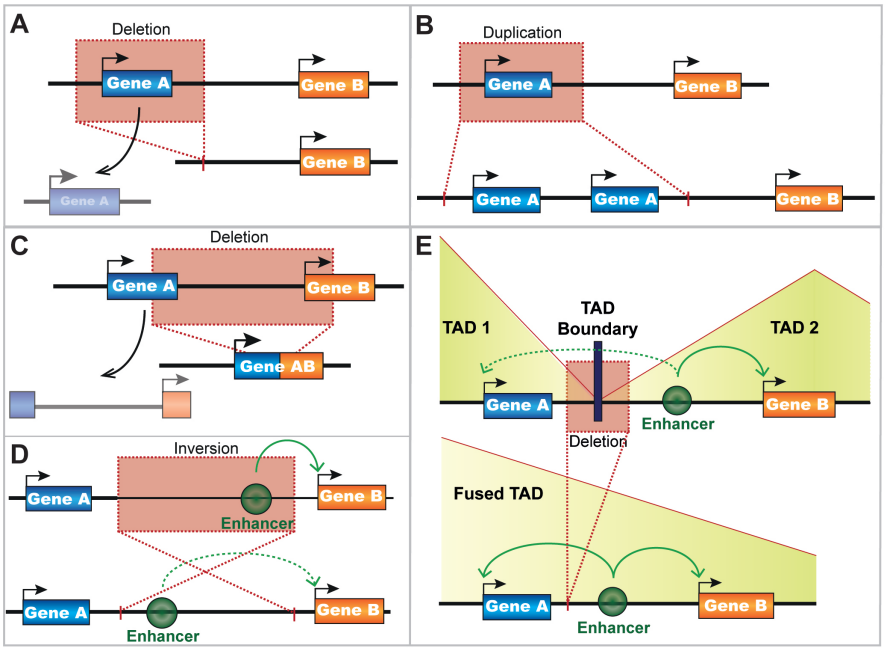


Figure 1.7 | Schematic examples of molecular effects of SVs on genes and regulatory elements. (A) *Deletions (highlighted in red) can cause a loss of one or more gene copies (in this case of gene A), which can lead to a reduction in gene expression.* ***(B)*** *Duplications (highlighted in red) can lead to gains of gene copies (in this case of gene A), which may lead to overexpression of the gene.* ***(C)*** *SVs affecting portions of genes lead to gene truncations. Fusion genes may be formed if multiple genes are truncated and brought together by the SVs.* ***(D)*** *SVs can also have indirect, positional effects on gene expression by affecting regulatory elements. For example, SVs may lead to a relocation of enhancers, thereby causing a loss of interactions between the enhancers and their target genes (for example between the depicted enhancer and gene B) and/or ectopic interactions between the enhancers and different genes (for example between the enhancer and gene A).* ***(E)*** *SVs can also affect chromatin organization for example by disrupting the boundaries of TADs. Disruption of TAD boundaries can lead to ectopic interactions between genes and enhancers that are normally separated from each other by the boundary. For example, the depicted enhancer normally regulates the expression of gene B, which is located in the same TAD as the enhancer. However, deletion of the TAD boundary enables the enhancer to interact with gene A as well, which can lead to expression changes of gene A and /or gene B.*

1.6.2 The non-coding genome

SVs do not only affect gene bodies, but they also affect non-coding parts of the genome. A surprising finding of the human genome project was that less than 2% of the genome contains protein-coding sequences (International Human Genome Sequencing Consortium, 2004). Initially the function of most of the non-coding genome was unknown and sometimes it was even called "junk DNA". Although the precise function of most of the non-coding genome is still not precisely known nowadays, it has been shown that a large part of it shows biochemical activity and is

involved in regulation of gene expression, indicating that much of it has an important function (Kellis et al., 2014). Proteins have to be active at the right times and at the right levels and therefore transcription of DNA to RNA is very tightly regulated. DNA is not freely floating in the nucleus, but it is wrapped around histones forming nucleosomes and it is bound by many other proteins (Lai and Pugh, 2017). DNA and proteins bound to it together form chromatin. Many chromatin regions, called heterochromatin, are tightly packed, preventing other proteins to access the DNA (Allshire and Madhani, 2018). In contrast, euchromatin is more open and other proteins are able to bind to the DNA. This open chromatin for example allows recruitment of transcription factors and subsequent transcription of genes to RNA. The accessibility of chromatin is regulated by chemically modifying the histones or the DNA itself (methylation) (Klemm et al., 2019). The histone and DNA modifications together also form a code which is referred to as the epigenome (Allis and Jenuwein, 2016). In contrast to the stable sequence of the genome, the epigenome is very dynamic (Soshnev et al., 2016). Many different NGS-based technologies have been developed to study the epigenome and the accessibility of chromatin (Zentner and Henikoff, 2014). These technologies give insight in the status of chromatin and therefore in the activity of the genome in the studied cells.

1.6.3 Regulation of gene expression by enhancers

The human genome contains small sequences of ~100 to 1500 nucleotides called **enhancers** that are important in the regulation of the epigenome and transcription (Heinz et al., 2015; Long et al., 2016; Shlyueva et al., 2014). There are hundreds of thousands enhancer regions in the human genome and the expression of a gene is frequently regulated by multiple enhancers (Dunham et al., 2012; Roadmap Epigenomics Consortium et al., 2015; Thurman et al., 2012). Enhancers contain short sequences ("motifs") that can be recognized and bound by proteins called **transcription factors** (Lambert et al., 2018; Spitz and Furlong, 2012). These bound transcription factors can recruit other proteins that can remodel the surrounding chromatin. Activated enhancers can activate transcription of genes, but enhancers can be located hundreds of kilobases away from their target gene. Chromatin has to be folded and physical loops have to be formed to allow enhancers to regulate expression of distantly located target genes. Loops form the **3D structure of the genome** and thousands of such loops are present in a nucleus. The 3D structure of the genome is essential in the regulation of gene expression and it is organized at different layers operating at different genomic scales with different dynamics and cell-type specificities (Bonev and Cavalli, 2016; Rowley and Corces, 2018). Knowledge about the 3D organization of the genome has rapidly increased in recent years due to the development of various chromatin conformation capture (3C) based technologies such as 4C-seq and Hi-C (Dekker et al., 2013; Schmitt et al., 2016).

Enhancers usually do not randomly interact with genes, but it is thought that these interactions are mostly confined within so called insulated neighbourhoods or **topologically associated domains** (TADs) ranging from 200Kb to 1Mb in size (Dixon et al., 2012). TADs are separated from each other by insulating boundaries which promote formation of interactions within domains and inhibit interactions between sequences in neighbouring domains (Dixon et al., 2012). These boundaries are frequently characterized by CTCF binding sites or by highly transcribed genes (Ali et al., 2016; Rao et al., 2014). It has been suggested that chromatin can be pushed or pulled through rings of cohesin (Fudenberg et al., 2016). The extrusion of chromatin through cohesin is blocked when convergent CTCF sites are encountered, leading to the formation of a chromatin loop anchored by CTCF. Thousands of loops of different sizes are formed in this way to order the genome in functional units that are thought to be regulated as modules. In addition to being subdivided into TADs, the genome is also organized into megabase-scaled compartments (Lieberman-Aiden et al., 2009). TADs containing mostly active genes can cluster together to form A compartments and inactive regions form B-compartments. The field of chromatin conformation has been one of the most rapidly developing research areas in molecular biology in recent years, but still much remains to be explored. Novel principles of chromatin folding, new roles for involved proteins (such as WAPL and YY1 (Busslinger et al., 2017; Haarhuis et al., 2017; Weintraub et al., 2018)) and new effects of genetic variants on the 3D genome are being discovered at a fast pace.

1.6.4 Effects of structural variants on regulatory elements can cause disease

Much of the non-coding DNA is involved in gene regulation and therefore SVs that affect non-coding DNA may disturb this regulation (Krijger and de Laat, 2016; Spielmann et al., 2018). In general, loss of a single enhancer only has a mild effect on gene expression and usually has no severe phenotypic consequences (Gasperini et al., 2019; Osterwalder et al., 2018). However, alteration of multiple enhancers targeting the same disease-associated gene by a structural variant can lead to congenital phenotypes. Such indirect effects, in which SVs affect gene expression not by altering the genes themselves, but by disturbing their regulatory contexts, are called position(al) effects (Figure 1.7D,E). The existence and potential influence of position effects on disease has been known for decades, but only in recent years tools have become available to study them at a large scale (Krijger and de Laat, 2016; Spielmann et al., 2018). Deletion of enhancers can lead to reduced expression of a target gene, which has for example been shown for the *SOX9* locus (Benko et al., 2009). Translocations can move enhancers to a different location in the genome, preventing them from regulating their regular target gene (Figure 1.7D). In contrast, duplications of enhancers may lead to overexpression of genes (Dathe et al., 2009). SVs can also have more complex effects

on gene regulation by disrupting TAD boundaries, which can lead to losses and gains of gene-enhancer interactions (Figure 1.7E). Such an impact of SVs on TAD organization was first shown for several SVs overlapping the *WNT6/IHH/EPHA4/PAX3* locus (Lupiáñez et al., 2015). Disruptions of the boundaries of the TAD containing the *EPHA4* gene cause ectopic interactions of surrounding genes with enhancers normally regulating *EPHA4* expression in limb development. These ectopic gene-enhancer interactions lead to increased RNA expression levels of the *Pax3*, *Wnt6* and/or *Ihh* genes (depending on the SV and the affected boundary) in developing limbs of mice, ultimately leading to limb phenotypes such as brachydactyly or polydactyly (Lupiáñez et al., 2015). Not much later, pathogenic position effects of SVs on TADs have been extensively described for other genomic regions such as the *SOX9*, *IHH* and *Pitx* loci (Spielmann et al., 2018).

It is challenging to study the precise effects of SVs on genes, because SVs can affect large genomic regions containing multiple genes and regulatory elements. In addition, the disease-relevant effects of SVs are frequently specific for certain cell types. For example, SVs may affect genes or enhancers that are mainly active during embryonic development of the limbs and therefore the precise molecular consequences such as changes in RNA expression can only be measured in the specific tissue during this developmental phase. Therefore, it is important to study the consequences of SVs in disease-relevant models, such as animal models. However, some neurological phenotypes are difficult to measure in animals. This is one of the reasons why mostly position effects of SVs leading to easily observable phenotypes, such as limb phenotypes, have been studied in detail so far (Spielmann et al., 2018). Additionally, although most genes are conserved between mammals, there is much variation between species in the regulatory landscapes of many genes regulating embryonic development. It is still not clear how frequently effects of SVs on regulatory elements cause neurodevelopmental disorders such as intellectual disability or autism spectrum disorders. Rough estimates suggest that disruptions of TAD boundaries by SVs, leading to "rewiring" of promoter-enhancer interactions within and between the affected TADs, cause the disorders of ~7.3% of patients with balanced SVs (Redin et al., 2017) and of ~11.8% of patients with large rare deletions (Ibn-Salem et al., 2014). Because of the difficulty to determine the effects of SVs, for most potential pathogenic SVs it is not known how they precisely cause the phenotype. Recent advancements in genome editing and culturing of patient-derived cells will help to study the molecular consequences of structural variants in disease-relevant cell types.

1.7 The causes and consequences of *de novo* structural variation

Despite the spectacular developments in the tools to detect genetic variants, the genetic causes of neurodevelopmental disorders remain unknown in half of the patients. In addition, the molecular consequences of genetic variants are frequently unknown. In this thesis we applied **multi-omics** approaches to study the causes of structural variants in embryos and the consequences of *de novo* structural variants in patients with neurodevelopmental disorders.

De novo structural variants can arise if DNA double stranded breaks are not properly repaired in parental gametes or in the early embryo. In **chapter 2** we studied the consequences of sperm DNA damage on the genomic integrity of early embryos using single cell whole genome sequencing.

In chapter 3 and chapter 4 we studied the molecular consequences of *de novo* SVs using multiple sequencing approaches. In **chapter 3** we used whole genome sequencing to detect *de novo* SVs in the genomes of 39 individuals with neurodevelopmental disorders who received an inconclusive diagnosis after regular genetic testing. To improve the molecular diagnosis of these patients, we developed a computational method to predict direct and indirect effects of the SVs on genes at or adjacent to the SVs.

In **chapter 4** we differentiated induced pluripotent stem cells derived from a patient with very complex genomic rearrangements into neural progenitor cells to obtain disease-relevant cells for studying the consequences of SVs. We used RNA-seq, 4C-seq and Hi-C to determine the cell type-specific effects of the *de novo* SVs on adjacent genes.

Discovery and validation of new variants involved in genetic disorders remains an important challenge in clinical genetics. In **chapter 5** we describe our findings of biallelic single nucleotide variants in the *POLR3GL* gene, which had not been associated with disease before, in three individuals with syndromic forms of endosteal hyperostosis.

Finally, in **chapter 6** we reflect on the work presented in this thesis. We will discuss the implications and limitations of using multi-omics approaches to study SVs. Although our approaches improved our understanding of the causes and consequences of *de novo* SVs, still many challenges lie ahead.

Chapter 2

Sperm DNA damage causes genomic instability in early embryonic development

Sjors Middelkamp, Helena van Tol, Diana C.J. Spierings, Sander Boymans, Victor Guryev, Bernard Roelen, Peter M. Lansdorp, Edwin Cuppen**, Ewart Kuijk

*** Corresponding author*

Adapted from:
https://www.biorxiv.org/content/10.1101/681296v1

Abstract

The majority of human embryos are lost early in development due to genetic mosaicism that affects about 70% of the human embryos. In spite of the impact on human health and fecundity, the factors that underlie this embryonic genome instability are largely unknown. Here we examined the consequences of sperm DNA damage on the embryonic genome by single cell genome sequencing of individual blastomeres from two- and eight-cell embryos produced with sperm damaged by radiation. Sperm DNA damage was found to induce a broad spectrum of genomic aberrations through fragmentation of chromosomes in two-cell stage embryos and the induction of segregation errors and heterogoneic cell divisions. Embryos that manage to escape developmental arrest may compromise health by causing mosaic aneuploidies, mixoploidy, uniparental disomies, *de novo* structural variation and possibly other rare genomic disorders of early embryonic origin.

2.1 Introduction

In early embryonic development, there is reduced activity of cell cycle checkpoints and apoptotic pathways until the zygotic genome becomes activated (Braude et al., 1988; Fatehi, 2006; Kiessling et al., 2009; Mantikou et al., 2012; Palmer and Kaldis, 2016; Toyoshima, 2009). As a consequence, mitotic errors, which include chromosome missegregations and spindle abnormalities, are tolerated in the first divisions leading to aneuploidies and subchromosomal aberrations involving one or multiple chromosomes. The result of this genomic instability is mosaicism, i.e. the phenomenon that cleavage stage embryos are composed of multiple genetic lineages. Mosaicism affects approximately three quarters of the human day 3 cleavage stage embryos and contributes to the low success rate of *in vitro* fertilization (IVF) through high miscarriage rates and failed implantations (Chavez et al., 2012; van Echten-Arends et al., 2011; Macklon et al., 2002; Mantikou et al., 2012; Munné et al., 2017; Spinella et al., 2018; Taylor et al., 2014; Vanneste et al., 2009). In addition to early pregnancy loss, embryonic mosaicism can also lead to molar pregnancies and parthenogenetic, androgenetic chimaeric, and mixoploid lineages in live-born humans (Kaiser-Rogers, 2005; Makrydimas et al., 2002; Robinson et al., 2007; Strain et al., 1995; Weaver et al., 2000). Thus, genetically distinct lineages can participate in development and contribute to disease. In spite of the immediate relevance for human health and fertility, the causes for the high mitotic error rates in human preimplantation embryos are largely unknown (van Echten-Arends et al., 2011; Vázquez-Diez and Fitzharris, 2018). Mosaicism is prevalent in human spontaneous abortions of natural pregnancies (Lebedev et al., 2004; Vorsanova et al., 2005), indicating that the causes for the high mitotic error rate in embryos are unrelated to the IVF procedures such as the ovarian stimulation regime, fluctuations in oxygen tension or temperature, and composition of the culture medium (Baart et al., 2007; Bean, 2002; Verpoest et al., 2008). Although advanced maternal age increases the risk for meiotic errors leading to whole embryo aneuploidies, mitotic errors and embryo mosaicism are not correlated with female age (Antonarakis et al., 1993; McCoy et al., 2015; Munné et al., 2017). Important mechanistic insight has come from a genome-wide association study that identified a polymorphism in the *polo-like kinase 4* (*PLK4*) gene that is associated with mitotic errors in development. *PLK4* is involved in centriole duplication and the minor allele is associated with tripolar chromosome segregations (McCoy et al., 2015, 2018). However, it is unlikely that *PLK4* polymorphisms alone can explain the high prevalence of mosaicism in human embryos. Thus far, the role of the sperm cell in embryonic mosaicism has largely been ignored (Colaco and Sakkas, 2018), possibly because paternal effects on the embryonic genome are mostly presumed to be restricted to the zygote stage. A plethora of factors can cause sperm DNA damage, including protamine imbalances, abortive apoptosis, advanced male age, oxidative stress, storage temperatures, and infections (González-Marín et al., 2012). However, sperm

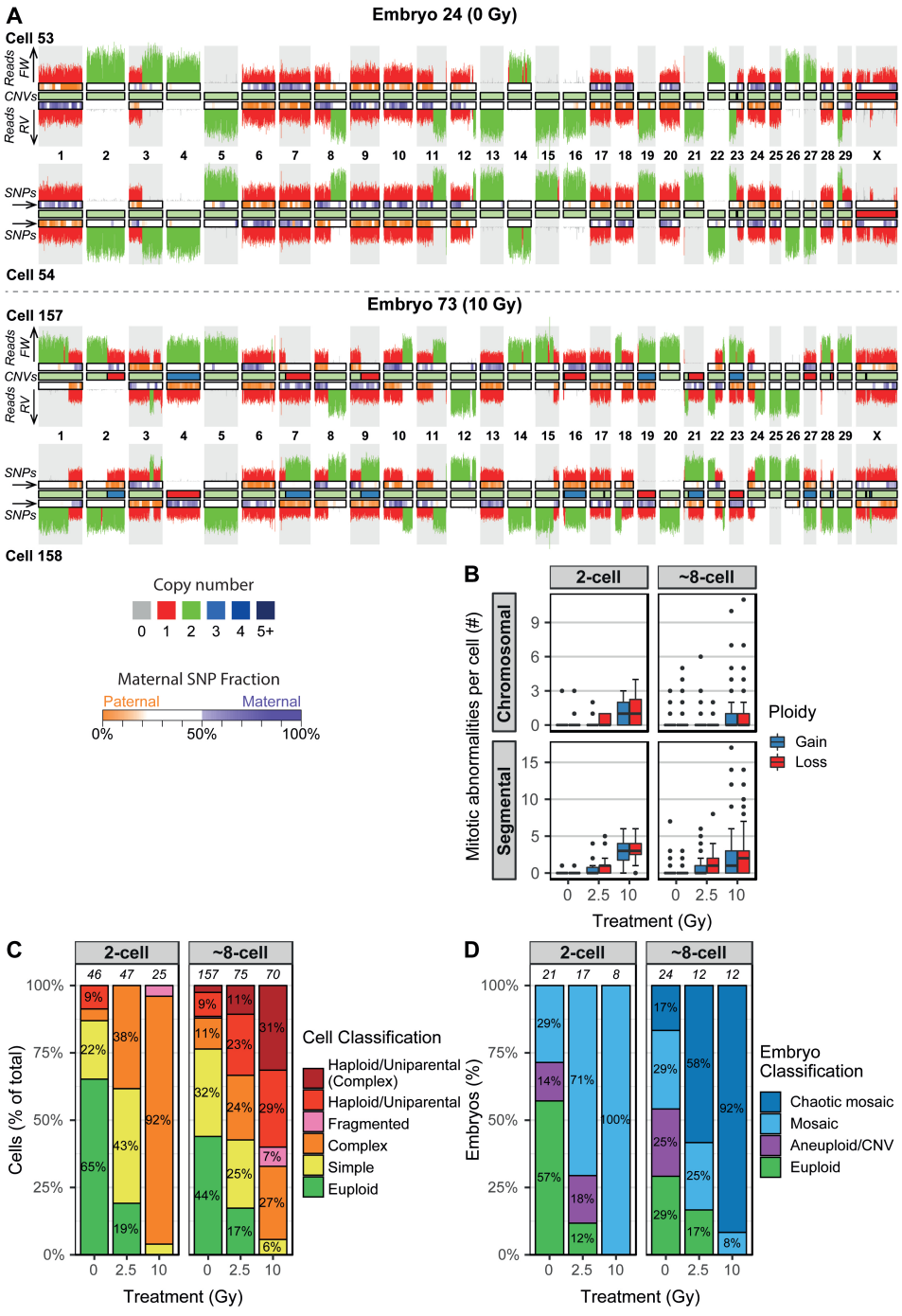


Figure 2.1 | Sperm DNA damage causes mitotic errors and mosaicism in embryos. (A) *Representative examples showing chromosome ideograms with strand-seq copy number profiles. Top: two-cell diploid control embryo; bottom: two-cell mosaic embryo produced with 10Gy-treated sperm.* ***(B)*** *A radiation dosage dependent increase in the number of whole chromosome and* >>>

DNA damage does not necessarily influence seminal parameters, sperm morphology and motility, or impair fertilization of the oocyte (Fatehi, 2006).

Bovine IVF and embryo culture is increasingly recognized as a valuable model system to study genomic instability in mammalian embryos (Destouni et al., 2016; Tšuiko et al., 2017). In this study we took advantage of this system to investigate the consequences of sperm DNA damage on embryonic genome integrity. Single cell whole-genome sequencing of individual blastomeres of two- and eight-cell stage bovine embryos revealed that sperm DNA damage results in reciprocal gains and losses of chromosomes and chromosomal segments in individual blastomeres at the two-cell stage. In addition to these immediate consequences, sperm DNA damage causes genomic instability leading to chaotic mosaicism with a broad variety of genomic aberrations in later-stage embryos.

2.2 Results

To examine the consequences of sperm DNA damage on the developmental competence of embryos, bovine IVF was performed with sperm subjected to γ-radiation. Increasing doses of γ-radiation reduced blastocyst formation rates (Figure S2.1A), but did not have a large effect on cleavage rates (data not shown). In agreement with a previous study (Fatehi, 2006), the main effects of radiation on developmental potential occurred at around the eight-cell stage, which coincides with the activation of the zygotic genome (Graf et al., 2014). Development up to the eight-cell stage thus appears to be a deterministic process regulated by maternally deposited factors that support the first cleavage divisions irrespective of the degree of DNA damage to the sperm cell. The absence of strong selective forces until the eight-cell stage of development allows the formation of genomic aberrations that are non-viable at later stages and therefore these early embryonic stages provide a window of opportunity to study genomic instability in a naive manner.

The consequences of sperm DNA damage on the stability of the embryonic genome were studied by sequencing all individual blastomeres of embryos produced with sperm subjected to a low (2.5 Gy) or high (10 Gy) dose of radiation. First, we employed strand-seq, a single-cell genome sequencing technique in which only DNA strands are sequenced that were used as templates during DNA replication prior

<<< segmental gains and losses per cell in two- and eight-cell stage embryos. ***(C)*** *Classification of all the cells for the different treatment groups. The proportion of cells with multiple genomic abnormalities increases with sperm radiation dose. Cells with three or more chromosomal or segmental abnormalities are classified as complex. Fragmented cells only contain a few chromosomal fragments. Numbers above the bars indicate the number of analyzed cells per group.* ***(D)*** *Classification of all the embryos for the different treatment groups. The majority of embryos derived from fertilization with damaged sperm are mosaic, containing more than a single genotype. Embryos containing cells with a mix of more than three different genotypes are considered chaotic mosaic. The number of analyzed embryos per group is indicated above the bars.*

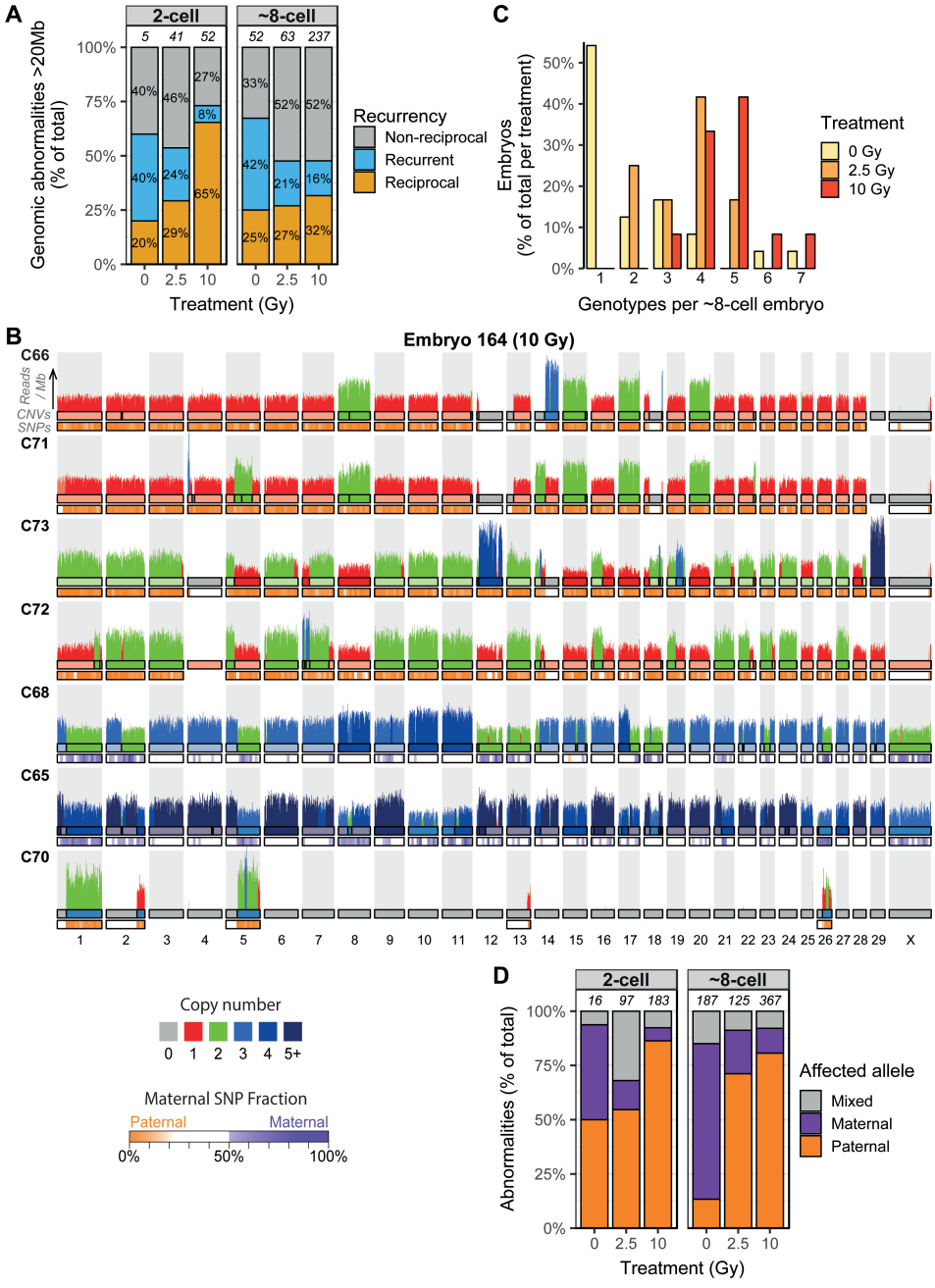


Figure 2.2 | Chaotic mosaicism in eight-cell stage embryos produced with damaged sperm. (A) *Many genomic abnormalities (>20 Mb in size) are mirrored between cells, showing reciprocal gains and losses between at least two cells. The relatively high number of non-reciprocal abnormalities, i.e. restricted to one cell per embryo, is largely due to variation in CNV calls between cells and in some cases due to missing or excluded cells. Numbers above the bars indicate the number of analyzed cells per group.* ***(B)*** *Copy number profiles of seven sequenced cells from embryo E164 showing complex genomic abnormalities. Cells C66, C71, C72 and C73 are uniparental with only* **>>>**

to cell division, to study the genomes of two-cell stage embryos (~17 hours post-fertilization (hpf)). Newly synthesized DNA strands, which have bromodeoxyuridine (BrdU) incorporated during replication, are selectively degraded, which leads to typical Watson-Watson, Watson-Crick, or Crick-Crick strand inheritance patterns in daughter cells after cell division (Falconer and Lansdorp, 2013; Falconer et al., 2012; Porubský et al., 2016). Strand-seq enables detection of copy number changes based on read-depth in single cells (Bakker et al., 2016). The sequenced libraries of 46 individual blastomeres derived from 25 two-cell stage control embryos and 72 individual blastomeres of 47 two-cell stage embryos produced with damaged sperm that passed quality control were analyzed (Figure S2.1B,C). Sister cells displayed the typical complementary strand inheritance patterns (Figure 2.1A).

Copy number analysis revealed few copy number aberrations in the two-cell embryos produced with untreated sperm (Figure 2.1B). Consistent with previous reports (Daughtry et al., 2019; Destouni et al., 2016), ~10% of embryos produced with untreated sperm contained one or more copy number change due to a meiotic error (Figure S2.2) and ~29% of control embryos showed defects due to mitotic errors (Figure 2.1C,D). Strikingly, the majority of embryos (~68%) produced with damaged sperm showed multiple whole chromosome and segmental gains and losses with the number of aberrations increasing in a dose-dependent manner (Figure 2.1). Most of the detected abnormalities were reciprocal between sister cells, where chromosomes or chromosomal segments that were gained in one cell were lost in its sister cell resulting in an average disomic copy number state in the embryo, a phenomenon we refer to as mirrored mosaicism (Figure 2.1A, Figure 2.2A). In agreement with the dose-dependent effects on the developmental potential, increasing radiation dosage resulted in more genomic aberrations.

To examine the genomic consequences at later stages we performed single-cell whole genome sequencing (van den Bos et al., 2019) of individual blastomeres at the ~eight-cell stage of development (~48 hpf) (Figure 2.2B). In total, 302 individual blastomeres of 48 embryos, of which 24 were derived from fertilization with damaged sperm, were successfully sequenced (Figure S2.1B,C). Embryos derived from irradiated sperm contained fewer euploid cells and more cells with complex rearrangements, i.e. affecting at least three chromosomes (Figure 2.1C). Copy number alterations were frequently observed and many were either shared or, as observed in the two-cell stage

<<< paternal chromosomes, whereas C65 and C68 are biparental. C70 is a fragmented cell containing chromosomal fragments that are complementary to copy number losses in C65 and C68. ***(C)*** *Embryos produced with damaged sperm frequently show more than three different genetic lineages around the eight-cell stage of development, indicative of genomic instability through mitotic errors.* ***(D)*** *The majority of copy number changes (>10 Mb) in embryos derived from fertilization with damaged sperm are located on alleles inherited from the father. The copy number changes on the maternal alleles are largely caused by meiotic errors. Numbers above the bars indicate the number of analyzed cells per group.*

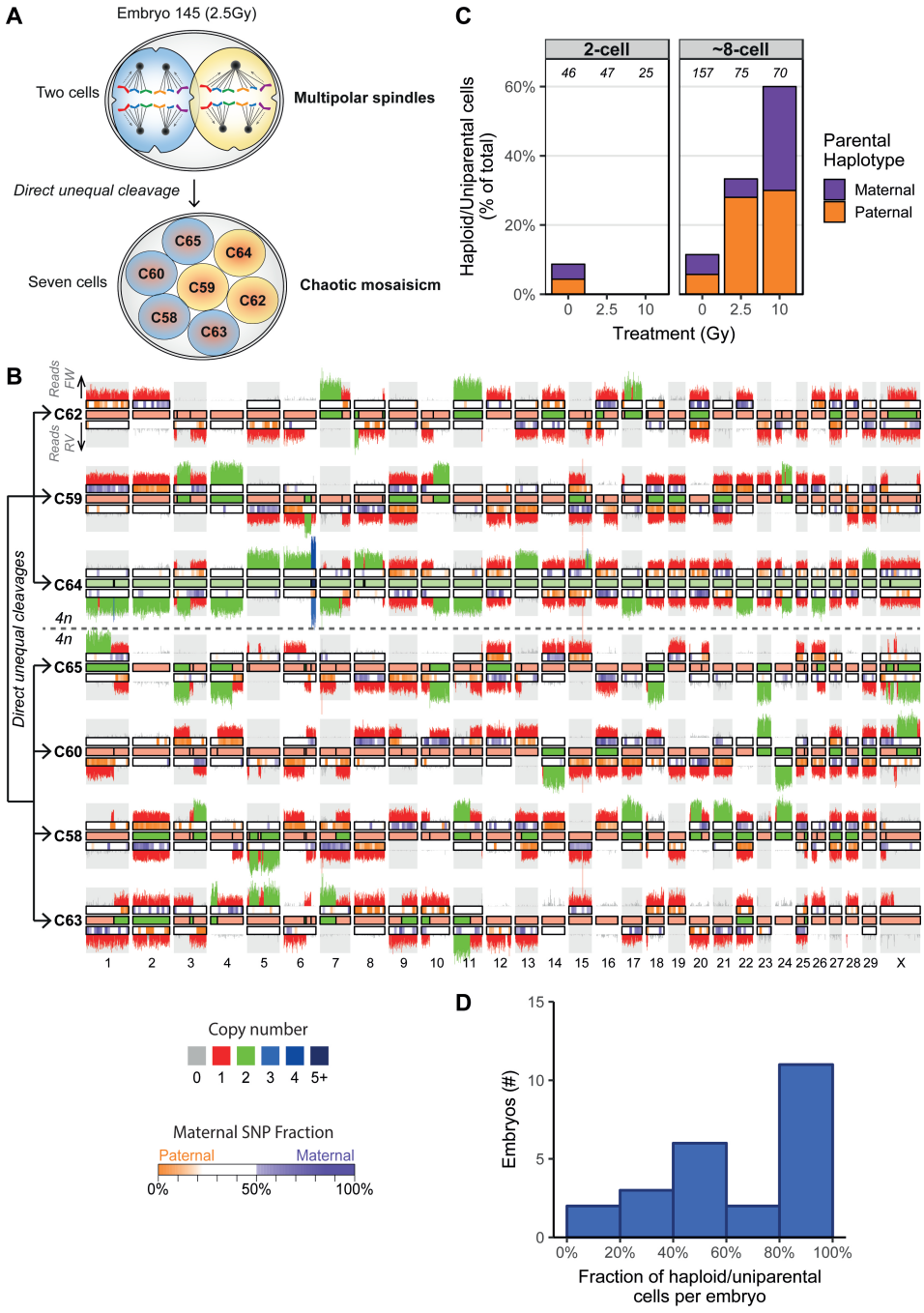


Figure 2.3 | Multipolar cell divisions lead to chaotic mosaicism at the eight-cell stage. (A) *Schematic reconstruction of the direct unequal cleavage divisions of two-cell stage blastomeres into respectively three and four daughter cells in embryo 145 (produced with 2.5 Gy-treated sperm). Sister cells are depicted with the same color.* ***(B)*** *Strand-seq karyogram of seven-cell embryo 145 showing chaotic mosaicism after direct unequal cleavage divisions at the two-cell stage. The strand inheritance patterns detected by strand-seq enable the identification of sister cells and the* >>>

embryos, mirrored between blastomeres from the same embryo (Figure 2.1B, Figure 2.2A). For each condition, the average number of chromosomal aberrations per cell was similar between the two-cell stage and the eight-cell stage (Figure 2.1B), which indicates no further fragmentation of chromosomes from the two-cell stage onwards. However, most eight-cell embryos derived from damaged sperm contain cells representing more than three different genotypes, indicating progressive genomic instability and segregation effects after the two-cell stage (Figure 2.2C).

All IVF experiments were performed with cryopreserved sperm that was derived from the same bull. Bulk whole genome sequencing of the sperm DNA enabled haplotyping of the embryonic single cell sequencing data (see materials and methods) and revealed that copy number alterations were strongly biased towards the paternally-derived chromosomes in both two and eight-cell embryos (Figure 2.2D, Figure S2.3). These results indicate that post-meiotic sperm DNA damage results in fragmentation of the paternal genome followed by distribution of the DNA fragments over both daughter cells during the first embryonic cell division.

Chaotic mosaicism, where blastomeres have seemingly random chromosome complements (Delhanty et al., 1997; McCoy et al., 2015, 2018), was common in embryos produced with irradiated sperm (Figure 2.1D). Some cells even contained very few and mostly fragmented chromosomes, while other blastomeres from the same embryo contained the complementary fragments (Figure 2.2B, Figure S2.3). This phenomenon, described as cellular fragmentation (Daughtry et al., 2019) occurred in 33% of the embryos produced with 10Gy-treated sperm (Figure S2.3B). To further investigate the processes that contribute to chaotic mosaicism we performed strand-seq on individual blastomeres of twelve ~8-cell stage embryos produced with damaged sperm. Strand inheritance patterns enabled a reconstruction of the formation of chaotically mosaic embryos. From the strand-inheritance patterns of two embryos we could deduce that seven cells were formed by direct unequal cleavage of both blastomeres of a two-cell stage embryo that cleaved directly into three and four cells respectively (Figure 2.3A,B, Figure S2.5A) (Zhan et al., 2016). These observations indicate that sperm DNA damage can cause aberrant cleavage divisions at the two-cell stage embryo resulting in chaotic mosaicism at later stages. In the strand-seq libraries we also observed sister cells that

<<< deduction of the preceding division and the distribution of the chromosomal fragments. This analysis reveals that both blastomeres at the 2-cell stage performed a multipolar division; a tripolar division generated the sister cells C62, C59 and C64 and a tetrapolar division generated the sister cells C65, C60, C58 and C63. This led to the random distribution of the tetraploid set of chromosomes over the sister cells. As a consequence, the DNA fragments distributed over the sister cells sum up to a 4n copy number state. ***(C)*** *Embryos generated with damaged sperm contain more haploid and uniparental cells having a genomic content from either the father or the mother, indicating that sperm DNA damage causes heterogoneic cell divisions. Numbers above the bars indicate the number of analyzed cells per group.* ***(D)*** *In many embryos only half of the cells are haploid/uniparental, suggesting that in some cases haploid/uniparental cells may arise after the two-cell stage.*

inherited complementary acentric fragments suggesting that these fragments have been translocated to centromere containing chromosomes to enable their segregation upon replication (Figure 2.3B, Figure S2.5A).

Strikingly, a large number of cells from eight-cell stage embryos produced with damaged sperm lacked X-chromosomes or contained nullisomies, indicating that these cells are (near) haploid or uniparental (Figure 2.2B, Figure S2.3C). To accurately quantify the number of haploid and uniparental cells, we screened for cells that lack heterozygous SNPs. Only a few cells are haploid and/or uniparental in two-cell embryos and in control eight-cell embryos (Figure 2.1C, Figure 2.3C). In contrast, the proportion of haploid/uniparental cells in eight-cell stage embryos produced with damaged sperm increased with radiation dose, amounting to two-thirds of the cells in the embryos produced with 10Gy-treated sperm (Figure 2.1C, Figure 2.3C).

A recent study described complete segregation of maternal and paternal genomes through so-called heterogoneic cell divisions, which were hypothesized to be the result of direct unequal cleavage of the zygote (Destouni et al., 2016). To examine if this process can indeed lead to heterogoneic cell divisions, we sequenced all blastomeres from nine embryos containing three cells that were presumed to have been formed after a direct unequal division of the zygote. In three out of nine three-cell embryos (two control embryos, one from irradiated sperm), all sequenced blastomeres were haploid (Figure S2.5B). These observations indicate that unequal cleavages of zygotes can indeed lead to heterogoneic cell divisions yielding uniparental lineages, but the number seems insufficient to explain the high incidence of haploid and uniparental cells observed at the eight-cell stage in embryos produced with damaged sperm. Because parental genomes still occupy distinct territories at the two-cell stage (Iqbal et al., 2011; Makrydimas, 2002; Reichmann et al., 2018), uniparental and haploid cells may also be formed by unequal cleavages at this stage of development. The observation that frequently half of the cells were haploid/uniparental (Figure 2.3D), indeed suggests these were formed by heterogoneic cell divisions of two-cell stage blastomeres.

2.3 Discussion

Our study demonstrates that sperm DNA damage leads to the fragmentation and random distribution of paternal chromosome segments over both sister cells of two-cell stage embryos. In addition, embryos that have been derived from fertilization with damaged sperm are prone to direct unequal cleavage divisions at the zygote stage or two-cell stage leading to the formation of haploid and uniparental cells. This leads to chaotic mosaicism at the eight-cell stage with blastomeres displaying a variety of genomic abnormalities ranging from aneuploidies, segmental changes, abnormal ploidy states, to cells containing minimal chromosomal content restricted to a few chromosomal fragments, thereby covering the broad spectrum of chromosomal

aberrations that have been previously described in human, primate, and bovine embryos (Daughtry et al., 2019; Destouni et al., 2016; Vanneste et al., 2009). Since sperm DNA damage has adverse effects on fertility (Evenson et al., 1980; Zhao et al., 2014), human IVF embryos may also be biased towards being produced with damaged sperm. Notably, in rhesus macaque embryos, chaotic aneuploidy was correlated with one particular sperm donor (Daughtry et al., 2019), which may indicate a role for sperm DNA damage in these rearrangements.

Gamma radiation results in about 13-37 double strand breaks per diploid human cell per Gy (Costes et al., 2010), causing approximately 33-93 double strand breaks cell when treated with 2.5 Gy. A similar number of breaks will be induced in bovine diploid cells, because the bovine genome size is similar to the human. For haploid sperm cells, the number of induced breaks will be less. Several techniques are available to measure the degree of sperm DNA damage in a population of cells such as the comet assay, the sperm chromatin dispersion assay, terminal deoxyuridine nick end labeling (TUNEL) assay, and sperm chromatin structure assay (Zini and Sigman, 2009). However, the sensitivity of the majority of these assays is limited. Even the comet assay, which is the most sensitive method to detect sperm DNA damage, has an estimated lower bound of 100 double strand breaks per cell (Collins et al., 2008). Our results from the 2.5 Gy embryos demonstrate that even upon the induction of sperm DNA damage close to or below this detection limit leads to embryonic genome instability. Additionally, it should be noted that it is inherently impossible to know the degree of DNA damage of the individual sperm cell that was used to fertilize an oocyte. Consequently, fertilizations with damaged sperm may be an underestimated phenomenon contributing to the widespread genomic instability in human embryos (Vanneste et al., 2009).

Direct unequal cleavage divisions are the result of multipolar spindles. Strikingly, fertilized oocytes that carry the minor allele of *PLK4* are also vulnerable to multipolar spindle formation (McCoy et al., 2015, 2018). Thus, spindle aberrations appear to be an important source of genomic instability in embryos. A recent study demonstrated that two spindles are formed in mouse zygotes, one for each parental genome, and the dual spindles are aligned prior to the first cleavage division (Reichmann et al., 2018). Sperm DNA damage may interfere with this process thereby inducing heterogoneic cell divisions of the zygote (Destouni et al., 2016). Our results suggest that sperm DNA damage also induces multipolar spindles in blastomeres of two-cell stage embryos. As a possible mechanism, sperm DNA damage may induce chromosome misalignments, which have been hypothesized to disturb the integrity of the spindle poles (Maiato and Logarinho, 2014).

Mature sperm cells lack mechanisms of DNA repair and depend on maternal factors for repair that are only available after fertilization. By absence of homologous templates, zygotic repair of paternal double-strand breaks depends on non-

homologous mechanisms that are considered error-prone, generating structural variation when originally distal fragments are joined. Pathogenic *de novo* structural variation that causes severe intellectual disability and other congenital anomalies is mostly of paternal origin (Brandler et al., 2018; Hehir-Kwa et al., 2011; Kloosterman et al., 2015), as is the case for the copy number alterations that were observed in the current study. Previous observations on metaphase spreads of mouse zygotes produced with damaged sperm indicate that sperm DNA damage can indeed lead to structural changes of the genome, although the fate of SVs beyond the zygote stage is unknown (Gawecka et al., 2013). While our single-cell sequencing data is not suited for the reliable detection of balanced structural variation, we did observe sister cells that inherited acentric fragments in the strand-seq libraries suggesting that these fragments have been translocated to other chromosomes.

Here, we have shown that sperm DNA damage induces fragmentation of chromosomes and segregation errors. A consequence of these two processes is chaotic mosaicism of embryos. In support with our findings, chaotic mosaicism is also common in human embryos produced with sperm from men with non-obstructive azoospermia, a condition that is also associated with high levels of sperm DNA damage (Macklon et al., 2009). Complex abnormal mosaic embryos have reduced implantation and clinical pregnancy rates and reduced chances to develop to term (Spinella et al., 2018). Chaotic mosaicism thus appears to be responsible for the well-established correlation between sperm DNA damage and reduced fertility (Evenson et al., 1980; Zhao et al., 2014). The chromosomal aberrations that are induced by damaged sperm include *de novo* structural variation, uniparental disomies, mosaic aneuploidies, and mixoploidy. When embryos escape developmental arrest these aberrations may contribute to congenital disease (Conlin et al., 2010; Kajii and Ohama, 1977; Kloosterman and Cuppen, 2013; Kurtas et al., 2019; van de Laar et al., 2002; Liu et al., 2017; Pellestor, 2014).

2.4 Materials and methods

2.4.1 Bovine IVF and blastomere collection

Fertilization and embryo culture were performed, according to previously described procedures (Aardema et al., 2017). In short, bovine cumulus oocyte complexes were aspirated from 2-8mm antral follicles of ovaries that were obtained from the slaughterhouse. Subsequently, germinal vesicle stage oocytes with an intact multilayered cumulus were selected and matured in M199 supplemented with 26.2 mM $NaHCO_3$, 0.02 IU/ml FSH (Sioux Biochemical Inc., Sioux Center IA, USA), 0.02IU/ml LH (Sioux Biochemical Inc.), 7.7 µg/ml cysteamine, 10 ng/ml epidermal growth factor, and 1% (v/v) penicillin-streptomycin (Gibco BRL) at 39°C in a humidified atmosphere of 5% CO_2 in air. In vitro fertilization was performed at 23 h after maturation with 0.5 ×

10^6 sperm cells per ml sperm. To obtain sperm with damaged DNA, sperm straws were subjected to ionizing radiation from a Gammacell 1000 (Atomic Energy of Canada Limited, Mississauga, Southern Ontario, Canada) prior to IVF. Ionizing radiation allows induction of DNA damage on non-cycling sperm cells while maintaining accurate control over the dosage. Untreated sperm from the same bull was used for the control group. All experiments were performed with sperm from the same donor bull to control for the potential natural variation in DNA damage between individuals. At 18–22 h after sperm addition, the cumulus cells and adhering sperm cells were removed and the denuded zygotes were further cultured in synthetic oviductal fluid (SOF) in a humidified incubator at 39°C with 5% CO_2 and 7% O_2. To obtain blastocysts, cleaved embryos were transferred to fresh SOF at day 5 and cultured until day 8. For strand-seq experiments at the 2-cell stage, bromodeoxyuridine (BrdU) was added to the embryo culture medium from the start of the embryo culture. Blastomeres were collected from 17h after fertilization (hpf) onwards. For strand-seq experiments at the 8-cell stage, 4-cell stage embryos (at 29-33 hpf) were transferred to medium containing BrdU and cultured until the 8-cell stage (at 48 hpf) when individual blastomeres were collected. For single-cell whole genome sequencing of 8-cell stage embryos, embryos were cultured in medium without BrdU and blastomeres were collected from 48hpf onwards. To collect individual blastomeres, embryos were placed in a droplet of pronase. After the zona pellucida was dissolved, the embryos were transferred to a droplet of Trypsin EDTA to dissociate the blastomeres. Blastomeres were transferred to single wells from a 96-well plate containing 5µl cryoprotectant consisting of 50% PBS0, 42.5% ProFreeze (Lonza), and 7.5% DMSO. Full plates were stored at -80°C until further processing.

2.4.2 Single-cell genome sequencing and primary data processing

Strand-seq and single-cell whole genome sequencing libraries were generated as previously described by respectively (Falconer et al., 2012) and (van den Bos et al., 2019). Libraries were pooled (192 libraries per rapid run flow cell lane) and sequenced on the Illumina HiSeq 2500 sequencing platform. Raw sequencing reads were mapped to the Bos taurus UMD3.1 (bt8) reference genome using Bowtie2 (Langmead and Salzberg, 2012) and BamUtil was used to filter duplicated reads. The median read count was 353,350 reads (with a mapping quality of more than 10) per cell after primary data processing.

2.4.3 Single cell copy number variant calling and filtering

The BAM files for all single cell libraries were merged to generate a composite BAM file using Samtools merge. Bedtools intersect was used to calculate the coverage per 100kb genomic bins. A blacklist for CNV calling (included with the scripts) was generated by selecting the 3% bins with the highest and 3% of the bins with the lowest read counts on the autosomes and the bins with the top 5% and bottom 3% read

counts on the X chromosome. The R-package AneuFinder (v1.8.0) was used to count the reads (with a minimal mapping quality of 10) in fixed-width bins of 1Mb and to call copy number variants using the "edivisive" method (Bakker et al., 2016). The genomic sequence provided by the R-package BSgenome.Btaurus.UCSC.bosTau8 (v1.4.2) was used for GC-correction applied by Aneufinder. CNV calls with a limited change in read count compared to the median read count per bin per cell were excluded (decrease of <25% for presumed losses and increase of <25% for gains). Subsequently, CNV calls for each cell were merged based on a variable overlap threshold dependent on CNV size into one CNV call set per embryo (e.g. larger CNV require a higher percentage of overlap to merge than smaller CNVs). CNV calls occurring in more than 15% of the high-quality control libraries (with more than 200,000 reads), which likely correspond to common population variants or reference genome artifacts, were removed from the call sets. CNVs were considered to be reciprocal if the ratio of gains versus losses within the embryo is more than 0.1. Sequenced libraries with more than 100,000 reads, 10 or less filtered chromosomal or segmental abnormalities, less than 80 segments detected by Aneufinder and, if applicable, alternating Watson/Crick strand inheritance patterns (whose mother cell incorporated BrdU during replication and underwent mitosis) were used as high-quality libraries for further analyses.

2.4.4 Bulk whole genome sequencing of bovine sperm DNA

Sperm DNA was extracted with the guanidine thiocyanate method (Griffin, 2013). A Covaris sonicator was used to shear the isolated DNA to fragments of 400-500 basepairs. Libraries for whole genome sequencing were prepared using the TruSeq DNA Nano Library Prep Kit (Illumina) according to the manufacturer's protocol. Paired-end 2x150 basepair read whole genome sequencing was performed on a Illumina Hiseq X sequencer to a mean genome coverage depth of 34x. Reads were aligned to the Bos taurus UMD3.1 reference genome using BWA-0.7.5a with settings BWA-MEM -t 12 -c 100 -M -R (Li and Durbin, 2009). Reads were realigned with GATK IndelRealigner (McKenna et al., 2010) and duplicate reads were flagged with Sambamba markdup (Tarasov et al., 2015).

2.4.5 SNP genotyping of sperm and blastomere DNA

All non-reference single nucleotide variants (SNVs) were called from the composite BAM file (containing all the reads from the sequenced single-cell libraries) using bcftools mpileup and bcftools call (Li, 2011). All heterozygous SNVs with more than 2 reference and 2 alternative allele counts and with a maximum coverage depth of 50 were selected to generate a list of 2,713,984 embryonic single nucleotide polymorphisms (SNPs). Subsequently, the paternal sperm WGS data was genotyped for the embryonic SNP positions using bcftools. To enable classification of SNPs in single embryonic cells as maternal (non-paternal), only the SNP positions that are homozygous in the father

(with a coverage depth between 10 and 75 in the sperm WGS data) were selected. All the single cells were genotyped for these 986,063 homozygous sperm SNP positions using bcftools. A SNP was classified as maternally-inherited if the genotype is different from the homozygous genotype in the father.

2.4.6 Determination of the ploidy status of single blastomeres

Haploid and uniparental cells were identified based on several parameters. First all cells were genotyped for the 2,713,984 variable embryonic SNP positions in the composite BAM file (see above). Cells were considered to be uniparental if less than 15% or more than 50% of the called SNPs in the cell were different from the homozygous SNPs in the father (Figure S2.3A). Additionally, haploid cells were detected by a loss of heterozygous SNP positions. Haploid/uniparental cells with more than 3000 covered SNPs were required to have less than one heterozygous REF/ALT SNP (excluding SNPs overlapping copy number gains) per 1000 called SNPs. Strand-seq libraries of haploid cells were recognized by the absence of bins with reads on both the Watson and Crick strands (haploid cells should only contain reads on one strand per bin after Strand-seq). Cells classified as haploid/uniparental were considered to be haploid (with a copy number state of one) if the majority (>80%) of called copy number losses are nullisomies.

2.5.7 Classifications of individual blastomeres and embryos

Cells were classified based on their ploidy status and the number of segmental and whole chromosome copy number changes. Cells containing three or more segmental or whole chromosome abnormalities were classified as complex. Cells with more than 10,000 reads and more than 25% of their reads on a single chromosome were considered to be fragmented. To determine the presence of different genotypes within each embryo, copy number changes (>20Mb) were compared between cells. Cells sharing more than 75% of their CNVs are considered to be of the same genotype. Embryos containing more than one or more than three different genotypes are classified as respectively mosaic and chaotic mosaic.

2.5.8 Statistical Analysis

Univariate ANOVAs was used to determine significant differences between experimental groups and the control group and the p-value was adjusted for multiple testing.

2.6 Supplemental materials

2.6.1 Acknowledgments

We thank Wigard Kloosterman for helpful discussions and Roel Janssen for bioinformatics support. We would also like to thank the Hartwig Medical Foundation for whole genome sequencing.

2.6.2 Funding

This work was supported by the funding provided by the Netherlands Science Foundation (NWO) Vici grant (865.12.004) to Edwin Cuppen and provided by De Snoo-van 't Hoogerhuijs Stichting to Ewart Kuijk.

2.6.3 Author contributions

HvT and EK performed wet-lab experiments. DS, VG, and PL performed single cell sequencing. SM, HvT, DS, and EK performed data analysis. SM, HvT, DS, BR, PL, EC, and EK were involved in the conceptual design of the study. PL, EC, and EK acquired financial support for the study. SM, EC, and EK wrote the manuscript with input from all authors.

2.6.4 Competing interests

The authors declare no competing interests.

2.6.5 Data and materials availability

All sequencing data have been deposited in the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under accession number PRJEB32696. Custom code used in this study is available on GitHub (https://github.com/UMCUGenetics/Bovine_Embryo/). Supplementary Table S2.1 is available on bioRxiv (https://www.biorxiv.org/content/10.1101/681296v1.supplementary-material).

2.6.6 Supplemental figures

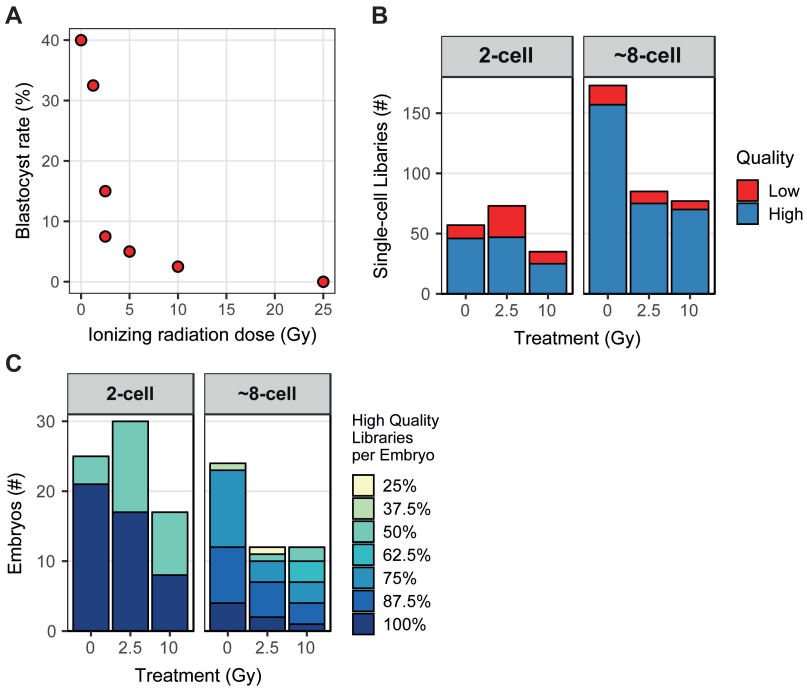
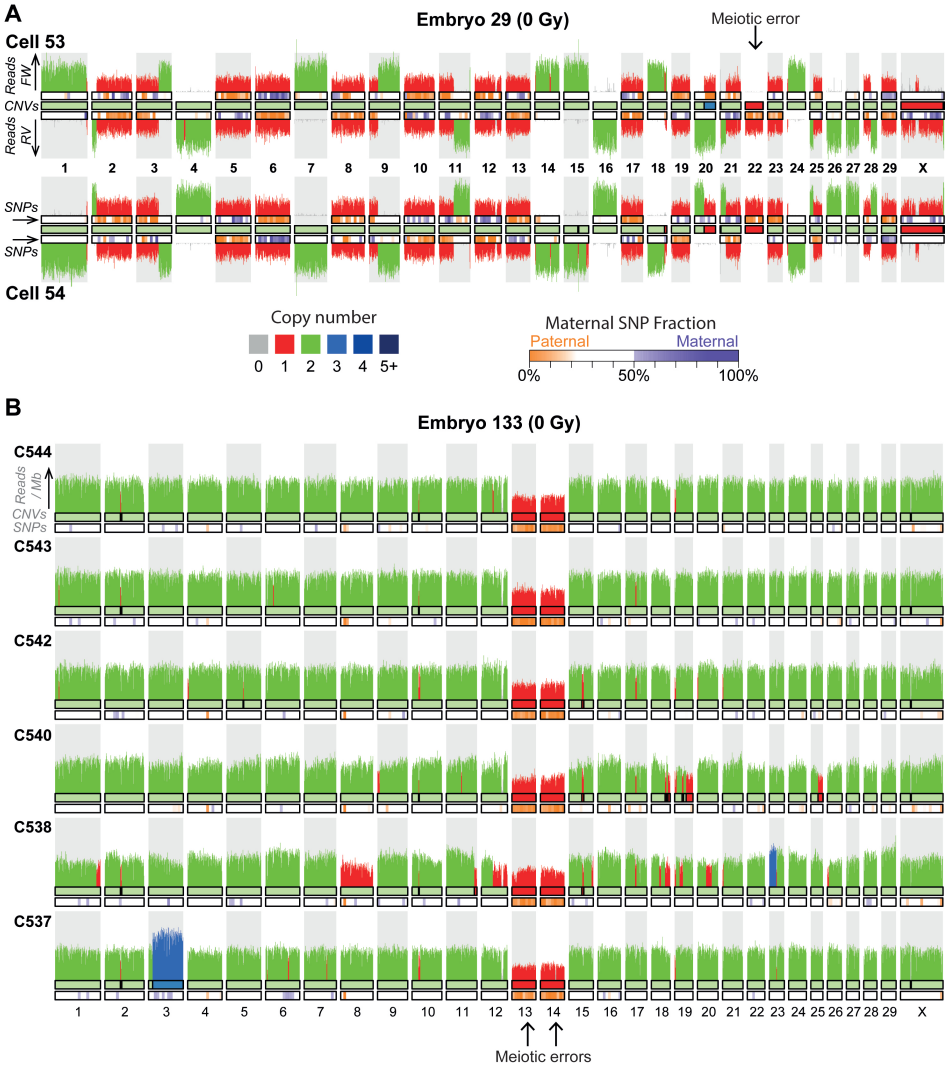


Figure S2.1 | Characteristics of blastocysts and single cell libraries of analyzed bovine embryos. (A) *Percentage of blastocysts that develop from fertilization with sperm treated with different doses of gamma-radiation. Forty embryos per treatment group were produced with sperm radiated with 0, 1.25, 2.5, 5, 10 and 25 Gy. The number of blastocysts was counted at day 8 after fertilization. The results were obtained from two independent fertilization experiments.* ***(B)*** *Number of successfully sequenced single-cell libraries per developmental stage. High quality libraries have more than 100,000 non-duplicate reads with a mapping quality of more than 10 and 10 or less filtered chromosomal or segmental abnormalities. Strand-seq libraries additionally required the typical strand inheritance patterns. Sequencing results for low quality libraries with more than 10,000 reads are also included in the karyograms, because they can be informative for identifying sister cells, but they are excluded for further quantitative analyses.* ***(C)*** *Percentage of sequenced high quality single cell libraries per embryo.*


A
Embryo 29 (0 Gy)
Meiotic error
Cell 53
Reads FW
CNVs
Reads RV
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 X
SNPs
SNPs
Cell 54
Copy number
0 1 2 3 4 5+
Maternal SNP Fraction
Paternal
Maternal
0%
50%
100%
B
Embryo 133 (0 Gy)
C544
Reads / Mb
CNVs
SNPs
C543
C542
C540
C538
C537
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 X
Meiotic errors
C
Embryos with meiotic errors (% of total)
20%
15%
10%
5%
0%
2-cell
~8-cell
Embryo stage
Type of error
Segmental gain+loss
Segmental loss
Chromosomal gain+loss
Chromosomal loss
D
ND
2
Paternal
6
Allele affected by meiotic error
15
Maternal

>>>

<<<

Figure S2.2 | Incidence and parental origin of meiotic errors in early embryos. (A) *Example of a two-cell embryo analyzed with strand-seq containing a loss of chromosome 22 due to a meiotic error in the maternal germline. The remaining chromosome 22 is enriched for paternal SNPs. The embryo also contains a reciprocal mitotic copy number change on chromosome 20.* ***(B)*** *Karyogram of six sequenced cells from one embryo showing meiotic losses of chromosomes 13 and 14. Only the paternally-inherited copies of chromosome 13 and 14 are present, indicating that the meiotic errors occurred on the maternal alleles.* ***(C)*** *Quantification of the number of cells containing different classes of meiotic abnormalities (>10Mb) per embryonic stage.* ***(D)*** *Number of meiotic errors (>10Mb) on maternally and paternally-inherited chromosomes. The bias towards the maternal alleles is consistent with previous findings showing an enrichment for meiotic errors in the maternal germline (Nagaoka et al., 2012).*

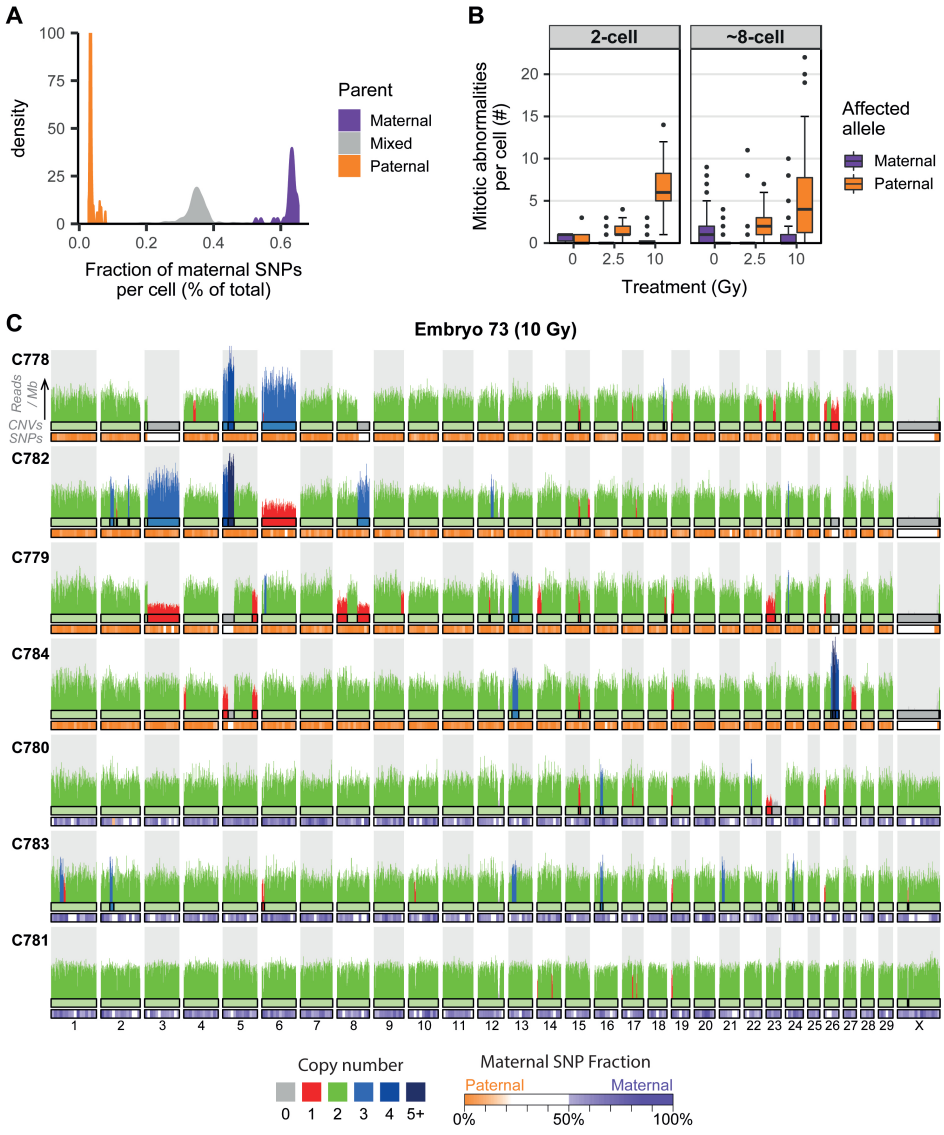


Figure S2.3 | Bulk sperm DNA sequencing allows parental haplotyping of embryonic cells. (A) *Proportion of SNPs in the single blastomeres corresponding to and diverging from homozygous SNPs in the genome of the father. Bulk whole genome sequencing enabled the detection of homozygous SNP positions in the genome of the bull whose sperm was used for all IVF experiments. SNPs in the blastomeres that are different from the SNPs in the father are considered to be maternally-inherited SNPs. SNPs overlapping between the father and the blastomeres can be paternally or maternally inherited (if the mother has the same SNP). The paternal peak is capped at y=100 to improve visibility of the maternal and mixed peaks.* ***(B)*** *Sperm DNA damage leads to genomic abnormalities (>10Mb) on the paternally-inherited chromosomes.* ***(C)*** *Example karyogram of a seven-cell embryo produced with damaged sperm containing a segregation of uniparental maternal and paternal cells, suggesting a heterogoneic cell division of the zygote. Copy number changes are mostly present in the cells containing a paternal genome (C778, C782, C779 and C784) and cells with maternally-inherited chromosomes are relatively unaffected (although there is a loss of chr23 in cell C780).*

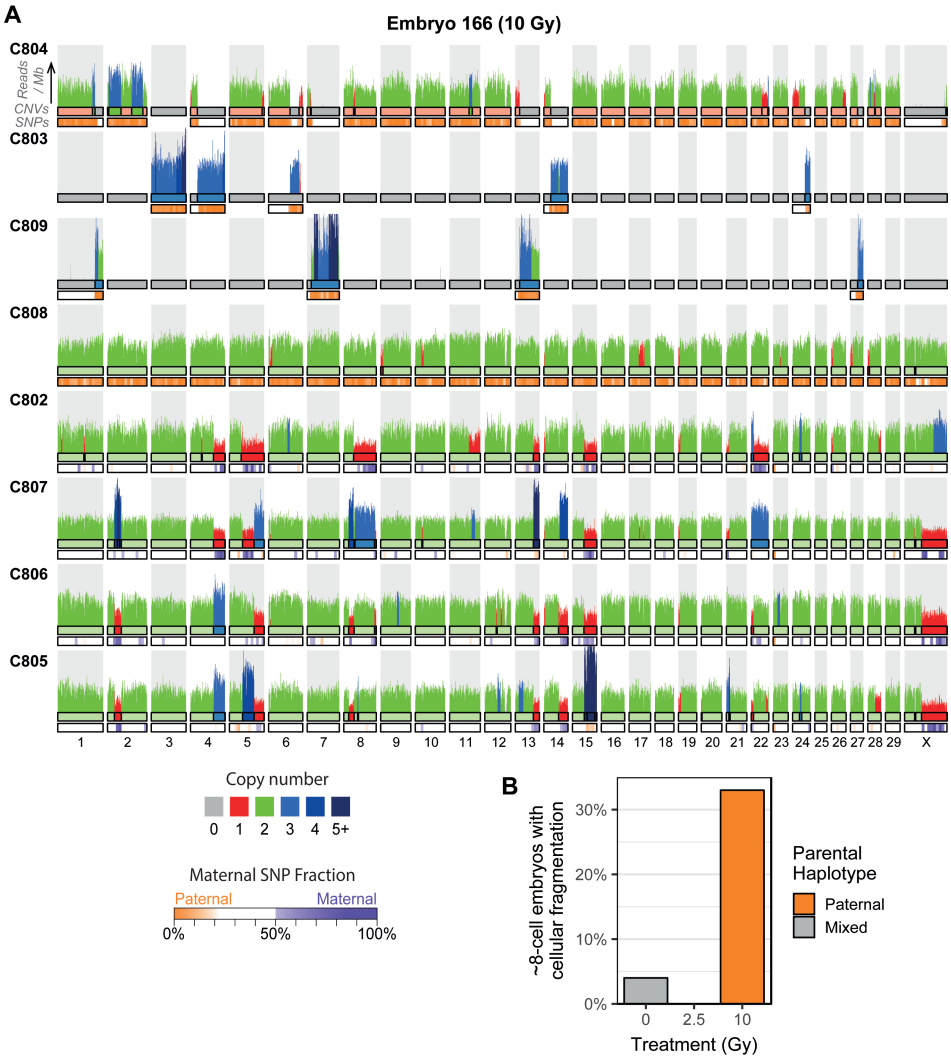


Figure S2.4 | Cellular fragmentation is common in embryos derived from damaged sperm. ***(A)*** *Example of an eight-cell stage embryo (E166) produced with damaged sperm (10Gy) showing cellular fragmentation. The mother cell of C804 is fragmented into three cells of which two cells (C803 and C809) only contain a few (respectively 5 and 4) chromosomal fragments. The top four cells only contain paternally inherited chromosomes. Cells C802, C807, C806 and C805 are diploid, but show many segmental gains and losses due to mitotic errors.* ***(B)*** *Quantification of the eight-cell stage embryos containing fragmented cells. Around a third of the eight-cell stage embryos produced with 10Gy-treated sperm contain fragmented cells with only paternally inherited chromosomes.*

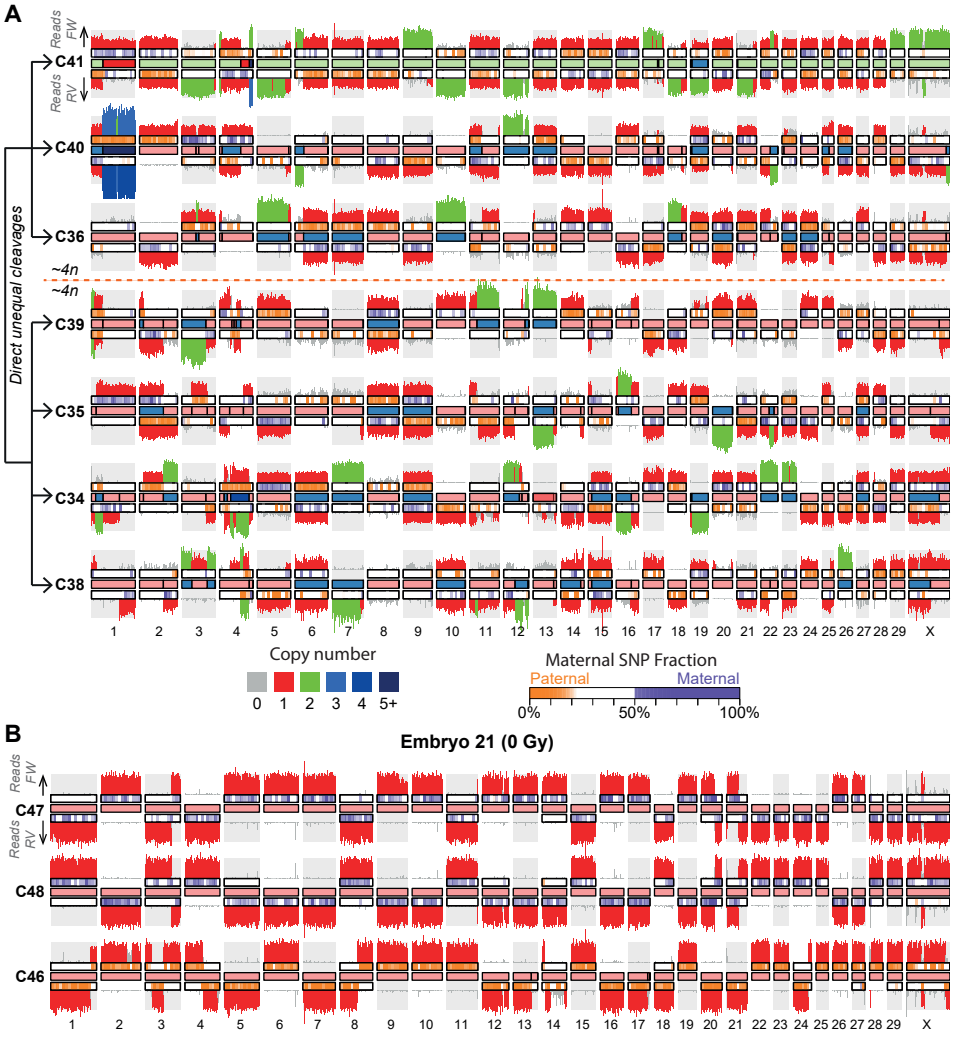
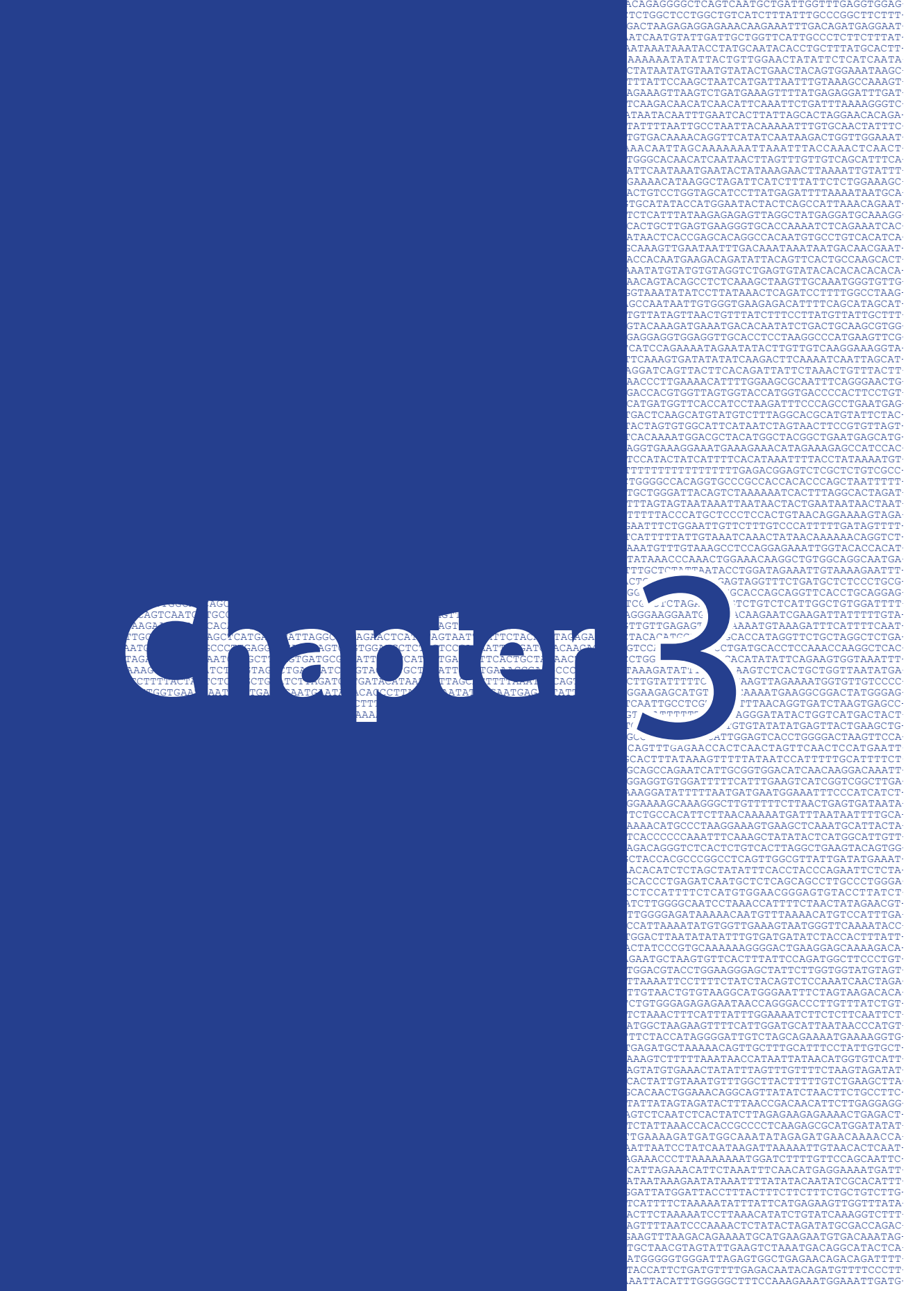


Figure S2.5 | Direct unequal cell divisions at zygote and two-cell stage cause complex genomic rearrangements. (A) *Karyogram of a seven-cell embryo analyzed by strand-seq showing the results of direct unequal cell divisions at the two-cell stage. The strand inheritance patterns indicate that cells C41, C40 and C36 are sister cells as are cells C39, C35, C34 and C38. The DNA fragments distributed over the sister cells sum up to a 4n copy number state.* ***(B)*** *Example of three haploid cells originating from a heterogoneic cell division at the zygote stage, which lead to segregation of the paternal and maternal genomes. Sister cells C47 and C48 contain a haploid maternal genome. The sister cell of C46 may have been lost during collection of the single blastomeres.*


Chapter 3

Prioritization of genes driving congenital phenotypes of patients with *de novo* structural variants

Sjors Middelkamp*, Judith M. Vlaar*, Jacques Giltay, Jerome Korzelius, Nicolle Besselink, Sander Boymans, Roel Janssen, Lisanne de la Fonteijne, Ellen van Binsbergen, Markus J. van Roosmalen, Ron Hochstenbach, Daniela Giachino, Michael E. Talkowski, Wigard Kloosterman, Edwin Cuppen**

** Equal contribution*

*** Corresponding author*

Adapted from:

https://www.biorxiv.org/content/10.1101/707430v1

Abstract

Background: Structural variants (SVs) can affect many genes and regulatory elements and the molecular mechanisms and the involved genes driving the congenital phenotypes of patients with *de novo* structural variants are frequently unknown.

Results: We applied a combination of systematic experimental and bioinformatic methods to improve the molecular diagnosis of individuals with *de novo* SVs who had an inconclusive diagnosis after regular genetic testing. First, we performed whole genome sequencing (WGS) on 39 patients with *de novo* SVs and detected additional disease-relevant complexities of the SVs missed by microarray testing in 15% of these cases. Next, we developed a computational tool to predict effects on genes directly affected by SVs and on genes indirectly affected due to changes in chromatin organization and impact on regulatory mechanisms. Combining these functional predictions with extensive phenotype information, identified candidate driver genes in 16 of the 39 (41%) included individuals. Subsequently, we applied this computational method to a collection of 382 patients with *de novo SVs* and identified candidate driver genes in 210 cases (54%), leading to 32 potential new diagnoses. Potential pathogenic positional effects were predicted in 25% of the cases with balanced SVs and in 8% of the cases with CNVs. Interestingly, in eight of the cases evidence was found for involvement of multiple affected candidate drivers contributing to different parts of the complex phenotypes.

Conclusions: These results show that identification of driver genes based on integration of WGS data with phenotype association and chromatin organization datasets can improve the molecular diagnosis of individuals with *de novo* SVs.

Keywords: Structural variation, Copy number variants, Neurodevelopmental disorders, Driver genes, Whole genome sequencing, Transcriptome sequencing, Topologically associated domains, Positional effects

3.1 Background

De novo germline structural variations (SVs) including deletions, duplications, inversions, insertions and translocations are important causes of (neuro-)developmental disorders such as intellectual disability and autism. Clinical genetic centres routinely use microarrays or karyotyping to detect SVs at kilo- to megabase resolution (Wright et al., 2018). The pathogenicity of an SV is generally determined by finding overlap with SVs in other patients with similar phenotypes (Hehir-Kwa et al., 2013; Nowakowska, 2017). SVs can affect large genomic regions which can contain many genes and non-coding regulatory elements (Weischenfeldt et al., 2013). This makes it challenging to determine which and how specific affected gene(s) and regulatory elements contributed to the phenotype of a patient. Therefore, the causative genes driving the phenotype are frequently unknown for patients with *de novo* SVs which can hamper conclusive genetic diagnosis.

SVs can have a direct effect on the expression and functioning of genes by altering their copy number or by truncating their coding sequences (Weischenfeldt et al., 2013). In addition, SVs can also indirectly influence the expression of adjacent genes by disrupting the interactions between genes and their regulatory elements (Krijger and de Laat, 2016). New developments in chromatin conformation capture (3C) based technologies such as Hi-C have provided the means to study these indirect effects (Dekker et al., 2013). Most of the genomic interactions (loops) between genes and enhancers occur within megabase-sized topologically associated domains (TADs). These domains are separated from each other by boundary elements characterized by CTCF-binding, which limit interactions between genes and enhancers that are not located within the same TAD (Bonev and Cavalli, 2016; Rowley and Corces, 2018). For several loci, such as the *EPHA4* (Lupiáñez et al., 2015), *SOX9* (Franke et al., 2016), *IHH* (Will et al., 2017), *Pitx* (Kragesteen et al., 2018) loci, it has been demonstrated that disruption of TAD boundaries by SVs can cause rewiring of genomic interactions between genes and enhancers, which can lead to altered gene expression during embryonic development and ultimately in disease phenotypes (Spielmann et al., 2018). Although the organization of TADs appears to be relatively stable across cell types, sub-TAD genomic interactions between genes and regulatory elements have been shown to be more dynamic and cell type-specific (Dixon et al., 2015). Disruptions of genomic interactions are therefore optimally studied in disease-relevant cell types, which may be obtained from mouse models or from patient-derived induced pluripotent stem cells. However, it is not feasible to study each individual locus or patient with such elaborate approaches. Therefore, it is not yet precisely known how frequently positional effects contribute to the phenotypes of patients with developmental disorders.

It has been shown that the use of computational methods based on combining phenotypic information from the Human Phenotype Ontology (HPO) database

("phenomatching") with previously published chromatin interactions datasets can help to improve the molecular diagnoses of patients with *de novo* SVs (Ibn-Salem et al., 2014; Yauy et al., 2018; Zepeda-Mendoza et al., 2017). These approaches have largely been based on data derived from a small set of cell types and techniques. Here, we further expand these approaches by integrating detailed phenotype information with genome-wide chromatin conformation datasets of many different cell types. By combining this method with whole genome and transcriptome sequencing we could improve the molecular diagnosis of 16 out of 39 individuals with *de novo* SVs who had an inconclusive diagnosis after regular genetic testing. By applying the computational method on larger cohorts of patients we estimated the significance of positional effects for both balanced and unbalanced SVs.

3.2 Results

3.2.1 WGS reveals hidden complexity of *de novo* SVs

We aimed to improve the genetic diagnosis of 39 individuals with multiple congenital abnormalities and/or intellectual disability (MCA/ID) who had an inconclusive diagnosis after regular genetic testing or who have complex genomic rearrangements. The phenotypes of the individuals were systematically described by Human Phenotype Ontology (HPO) terms (Köhler et al., 2009, 2017, 2019). The included individuals displayed a wide range of phenotypic features and most individuals (82%) presented neurological abnormalities including intellectual disability (Figure 3.1A, Table S3.1). All individuals carried *de novo* SVs which were previously detected by ArrayCGH, SNP arrays, karyotyping or long-insert mate-pair sequencing (Figure S3.1A). First, we performed whole genome sequencing (WGS) to screen for potential pathogenic genetic variants that were not detected by the previously performed genetic tests. No known pathogenic single nucleotide variants (SNVs) were detected in the individuals analyzed by patient-parents trio-based WGS (individuals P1 to P20), except for one pathogenic SNV that is associated with a part (haemophilia) of the phenotype of individual P1. A total of 46 unbalanced and 219 balanced *de novo* SVs were identified in the genomes of the individuals (Figure3.1B, Figure S3.1B, Table S3.2). The detected SVs range from simple SVs to very complex genomic rearrangements that range from 4 to 40 breakpoint junctions per individual. Importantly, WGS confirmed all previously detected *de novo* SVs and revealed additional complexity of the SVs in 7 (39%) of the 18 cases who were not studied by WGS-based techniques before (Figure 3.1C, Figure 3.2, Table S3.2). The previously identified duplications in 4 of 8 individuals appeared to be more complex in the WGS data, suggesting that the complexity of especially duplications is frequently underestimated by microarray analysis. In these cases, the duplications are not arranged in a tandem orientation, but instead they are inserted in another genomic region, which can have far-reaching consequences for the molecular

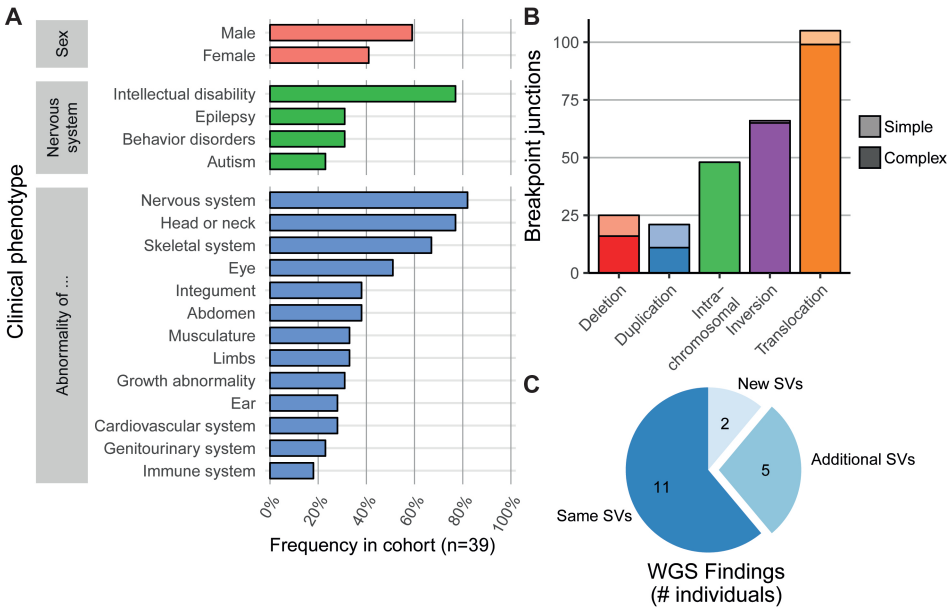


Figure 3.1 | Characterization of de novo SVs in a cohort of individuals with neurodevelopmental disorders. ***(A)*** *Frequencies of clinical phenotypic categories described for the 39 included individuals based on categories defined by HPO. Nervous system abnormalities are divided into four subcategories.* ***(B)*** *Number of de novo break junctions per SV type identified by WGS of 39 included patients. Most detected de novo SVs are part of complex genomic rearrangements involving more than three breakpoint junctions.* ***(C)*** *Number of cases in which WGS analysis identified new, additional or similar SVs compared to microarray-based copy number profiling.*

diagnosis for these individuals (Figure 3.2) (Brand et al., 2015; Gilissen et al., 2014; Nazaryan-Petersen et al., 2018). For example, in one case (P11) a previously detected 170 kb duplication from chromosome 9 was actually inserted upstream 82 kb of the *SOX3* gene on chromosome X (Figure 3.2, Figure S3.2). This inserted fragment contains a super-enhancer region that is active in craniofacial development (Wilderman et al., 2018) (Figure S3.2). The insertion of the super-enhancer may have disturbed the regulation of *SOX3* expression during palate development, which may have possibly caused the orofacial clefting in this individual (Brewer et al., 2016; Bunyan et al., 2014; DeStefano et al., 2013; Haines et al., 2015; Zhu et al., 2011). The detection of these additional complexities in the genomes of seven of the individuals exemplify the added value that WGS analyses can have for cases that remain unresolved after standard array diagnostics (Gilissen et al., 2014).

3.2.2 Phenodriver approach links directly affected genes to phenotypes

Subsequently, we determined if the phenotypes of the patients could be explained by direct effects of the *de novo* SVs, most of which were previously classified as VUS, on genes. In total, 332 genes are directly affected (deleted, duplicated or truncated) by the SVs in the cohort (Figure S3.1D). The Phenomatch tool was used to match the HPO

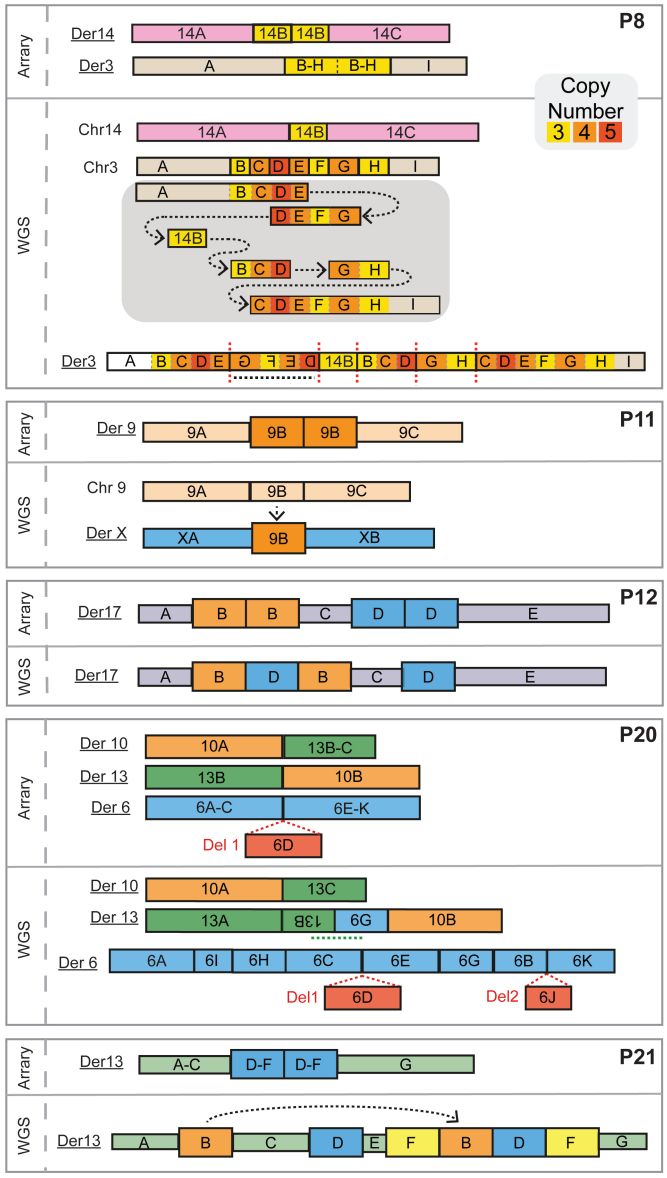


Figure 3.2 | Schematic representation of additional genomic rearrangements that were observed by WGS in five individuals. *For each patient, the top panel shows the de novo SVs identified by arrays or karyotyping and bottom panel shows the structures of the SVs detected by WGS. The WGS data of individual P8 revealed complex chromoanasynthesis rearrangements involving multiple duplications and an insertion of a fragment from chr14 into chr3. Individual P11 has an insertion of a fragment of chr9 into chrX that was detected as a duplication by array-based analysis (Figure S3.2). The detected duplications in individuals P12 and P21 show an interspersed orientation instead of a tandem orientation. The translocation in patient P20 appeared to be more complex than previously anticipated based on karyotyping results, showing 11 breakpoint junctions on three chromosomes.*

terms associated with these genes with the HPO terms used to describe the phenotypes of the individuals (Ibn-Salem et al., 2014; Zepeda-Mendoza et al., 2017). Genes were considered as candidate driver genes based on the height of their Phenomatch score, the number of phenomatches between the HPO terms of the gene and the patient, recessive or dominant mode of inheritance, Loss of Function constraint score (pLI) (Lek et al., 2016), Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2013) and the presence in OMIM and/or DD2GP (Firth et al., 2009) databases (Table 3.1). Directly affected genes strongly or moderately associated with the phenotype are classified as respectively tier 1 (T1) and tier 2 (T2) candidate driver genes (Figure 3.3A, Table 3.1). Genes with limited evidence for contribution to the phenotype are reported as tier 3 (T3) genes. In the cohort of 39 patients, this approach prioritized 2 and 13 of the 332 directly affected genes as T1 and T2 candidate drivers, respectively (Figure 3.3B). In three cases, the identified directly affected T1/T2 candidate drivers are associated with most (>75%) of the HPO terms of the individuals and are therefore predicted to explain most of the phenotypes (Table S3.4). In six other cases directly affected T1/T2 candidate drivers were identified that are only associated with parts (>20% of the patient's HPO terms) of the phenotypes (Table S3.4).

Subsequently, we performed RNA sequencing on primary blood cells or lymphoblastoid cell lines derived from the individuals to determine the impact of the *de novo* SVs on the RNA expression of the candidate driver genes. RNA sequencing confirmed that most expressed genes directly affected by the *de novo* deletions show a reduced RNA expression (97 of 107 genes with a median reduction of 0.46-fold compared to non-affected individuals) (Figure 3.3D). Although duplicated genes show a median 1.44-fold increase in expression, only 14 of 43 (~30%) of them are significantly overexpressed compared to expression levels in non-affected individuals. In total, 87 genes are truncated by SVs and four of these are classified as T1/T2 candidate drivers. The genomic rearrangements lead to 12 possible fusions of truncated genes and RNA-seq showed an increased expression for two gene fragments due to formation of a fusion gene (Figure S3.3, Table S3.3). However, none of genes involved in the formation of fusion genes were associated with the phenotype of the patients. Three deleted and two duplicated T1/T2 candidate drivers were expressed and all of them were differentially expressed compared to controls. The RNA sequencing data suggests that most genes affected by *de novo* deletions show a reduced RNA expression and limited dosage compensation, but increased gene dosage by *de novo* duplications does not always lead to increased RNA expression, at least in blood cells of patients.

3.2.3 Prediction of positional effects of *de novo* SVs on neighbouring genes

In 28 of the included cases (72%) our prioritization method did not predict T1/T2 candidate driver genes that are directly affected by the *de novo* SVs. Therefore, we

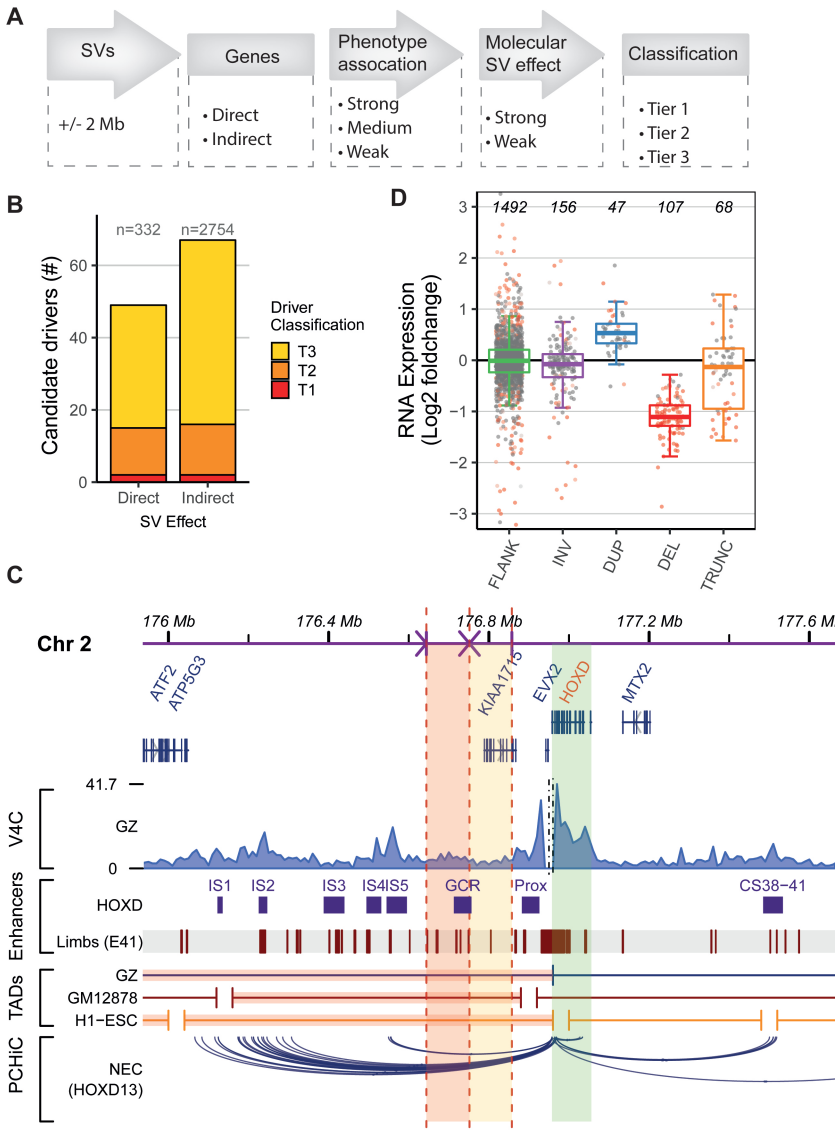
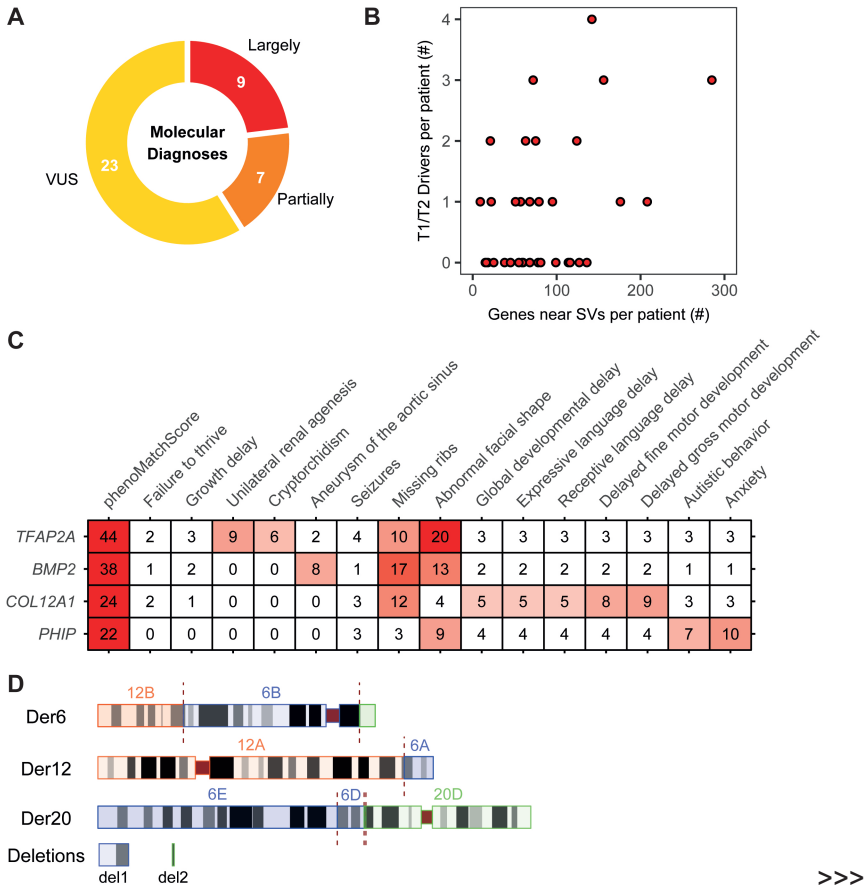


Figure 3.3 | Prediction of candidate driver genes directly and indirectly affected by the SVs. ***(A)*** *Schematic overview of the computational workflow developed to detect candidate driver genes. Classification of genes at (direct) or surrounding (indirect) the de novo SVs is based on association of the gene with the phenotype and the predicted direct or indirect effect on the gene (Table 3.1). The predicted effects of SVs on adjacent genes are based on integration of chromatin organization datasets of multiple cell types (Figure S3.4).* ***(B)*** *Total number of identified tier 1, 2 and 3 candidate driver genes predicted to be directly or indirectly affected by an SV.* ***(C)*** *Genome browser overview showing the predicted disruption of regulatory landscape of the HOXD locus in individual P22. A 107 kb fragment (red shading) upstream of the HOXD locus (green shading) is translocated to a different chromosome and a 106 kb fragment (yellow shading) is inverted. The SVs affect the TAD centromeric of the HOXD locus which is involved in the regulation of gene expression in developing digits. The translocated and inverted fragments contain multiple embryonic limb enhancers, including the global control region (GCR). Disruptions of these developmental enhancers likely contributed to* >>>

investigated positional effects on the genes surrounding the *de novo* SVs to explain the phenotypes of the cases that were not fully explained by directly affected candidate driver genes. We extended the candidate driver gene prioritization analysis by including all the protein-coding genes located within 2 Mb of the breakpoint junctions. Of the 2,754 genes adjacent to the SVs, 117 are moderately to strongly associated with the specific phenotypes of the individuals. To determine if the regulation of these genes is affected, we first evaluated RNA expression levels. Three-quarters (81/117) of the genes linked to the phenotypes were expressed, but only 8 and 1 of those showed respectively reduced or increased expression (Figure 3.3D). However, RNA expression in blood may not always be a relevant proxy for most neurodevelopmental phenotypes (Cai et al., 2010; Tylee et al., 2013). Therefore, we developed an extensive *in silico* strategy to predict potential disruption of the regulatory landscape of the genes surrounding the SVs (Figure S3.4). Because the interactions between genes and their regulatory elements are highly cell-type specific a large collection of tissue-specific Hi-C, TAD, promoter capture Hi-C (PCHiC), DNase hypersensitivity site, RNA and ChIP-seq datasets was included (Table S3.3). Several embryonic and neural cell type (such as fetal brain and neural progenitor cells) datasets are included that may be especially relevant to study the neurodevelopmental phenotypes in our cohort.

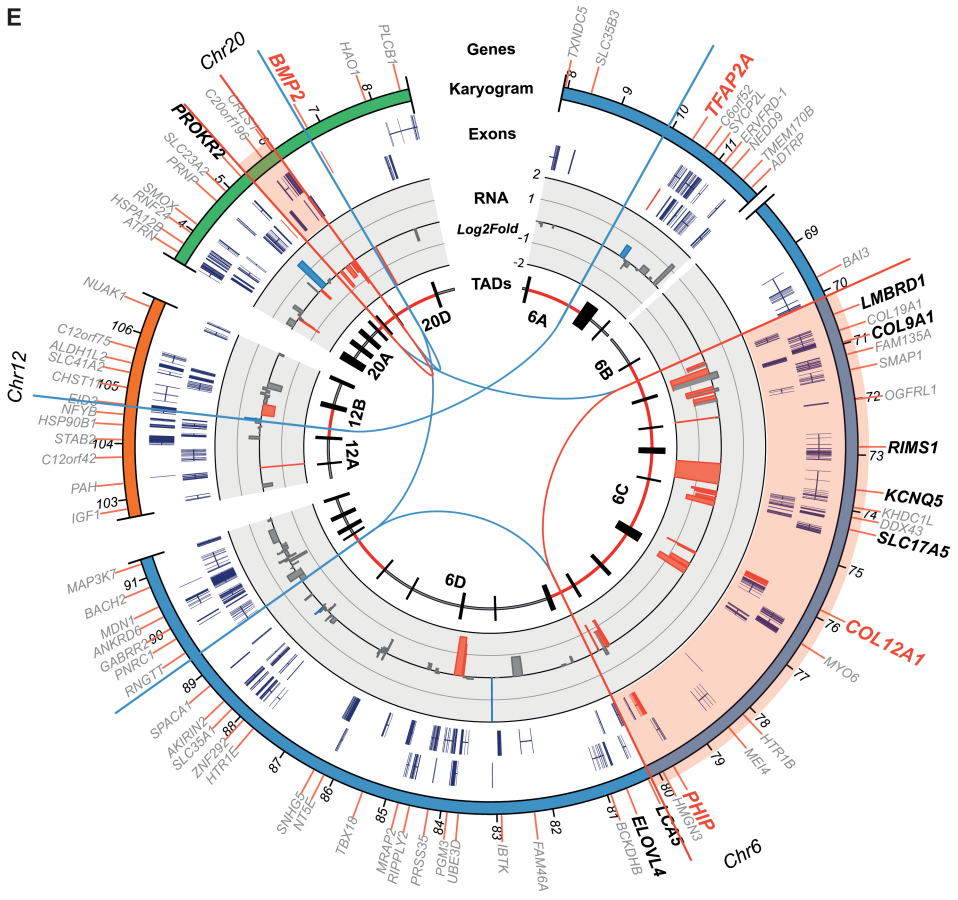
First, we determined which TADs of 20 different cell types overlap with the *de novo* SVs and which genes are located within these disrupted TADs (Schmitt et al., 2016; Wang et al., 2018; Won et al., 2016) (Figure S3.4B). One third of the genes surrounding the SVs (884/2754) are located within TADs that are disrupted in more than half of the assessed cell types. Subsequently, we determined if the disrupted portions of the TADs contain regulatory elements that may be relevant for the genes located in these TADs. For each gene, we selected the three cell types in which the gene is highest expressed based on RNA-seq data from the Encode/Roadmap projects (Roadmap Epigenomics Consortium et al., 2015), because not all included cell types and their cell-type specific regulatory elements may be relevant for each gene (Figure S3.4C). For each gene, the number of active enhancers (determined by chromHMM analysis of Encode/Roadmap ChIP-seq data (Roadmap Epigenomics Consortium et al., 2015)) in the TADs up- and downstream of the breakpoint junction was counted (Figure S3.4D). Because the coordinates of TAD boundaries can be dependent on the calling method and the resolution of the Hi-C (Dali and Blanchette, 2017; Yardımcı et al., 2019; Zufferey et al., 2018) and because a significant portion of genomic interactions crosses TAD

<<< the limb phenotype of the patient. ***(D)*** *RNA expression levels of genes at or adjacent to de novo SVs. Log2 fold RNA expression changes compared to controls (see methods) determined by RNA sequencing for expressed genes (RPKM >0.5) that are located within 2 Mb of SV breakpoint junctions (FLANK) or that are inverted (INV), duplicated (DUP), deleted (DEL) or truncated (TRUNC). Differentially expressed genes ($p < 0.05$, calculated by DESeq2) are displayed in red.*



boundaries, we also performed virtual 4C (v4C) for each gene by selecting the rows of the normalized Hi-C matrices containing the transcription start site coordinates of the genes as viewpoints. The v4C profiles were overlapped with the breakpoint junction to determine the portion of interrupted Hi-C interactions of the gene (Figure S3.4E). In addition, promoter capture Hi-C data of 22 tissue types (Cairns et al., 2016; Freire-Pritchett et al., 2017; Javierre et al., 2016; Rubin et al., 2017) and DNAse-hypersensitivity site (DHS) connections (Thurman et al., 2012) were overlapped with the SV breakpoints to predict disruption of long range interactions over the breakpoint junctions (Figure S3.4F). Integrated scores for TAD disruption, v4C disruption, potential enhancer loss, disruption of PCHiC interactions and DHS connections were used to calculate a positional effect support score for each gene (Figure S3.4). Finally, indirectly affected genes were classified as tier 1, 2 or 3 candidate drivers based on a combination of their association with the phenotype and their support score (Figure 3.3A, Table 3.1).

Based on this data integration, 16 of the 117 genes that are associated with the phenotypes and are located within 2 Mb of the SVs are predicted to be T1/T2



<<<

Figure 3.4 | SVs can affect multiple candidate drivers which jointly contribute to a phenotype. ***(A)*** *Number of patients whose phenotype can be partially or largely explained by the predicted T1/T2 candidate drivers. These molecular diagnoses are based on the fraction of HPO terms assigned to the patients that have a phenomatchScore of more than five with at least one T1/T2 driver gene.* ***(B)*** *Scatterplot showing the number of predicted T1/T2 candidate drivers compared to the total number of genes at or adjacent (< 2Mb) to the de novo SVs per patient.* ***(C)*** *Heatmap showing the association of the four predicted T1/T2 candidate drivers with the phenotypic features (described by HPO terms) of individual P25. The numbers correspond to score determined by Phenomatch. The four genes are associated with different parts of the complex phenotype of the patient.* ***(D)*** *Ideogram of the derivative (der) chromosomes 6, 12 and 20 in individual P25. WGS detected complex rearrangements with six breakpoint junctions and two deletions on chr6 and chr20 of respectively ~10 Mb and ~0.6 Mb.* ***(E)*** *Circos plot showing the genomic regions and candidate drivers affected by the complex rearrangements in individual P25. Gene symbols of T1/T2 and T3 candidate drivers are shown in respectively red and black. The break junctions are visualized by the lines in the inner region of the plot (red lines and highlights indicate the deletions). The middle ring shows the log2 fold change RNA expression changes in lymphoblastoid cells derived from the patient compared to controls measured by RNA sequencing. Genes differentially expressed (p<0.05) are indicated by red (log2 fold change < -0.5) and blue (log2 fold change > 0.5) bars. The inner ring shows the organization of the TADs and their boundaries (indicated by vertical black lines) in germinal zone (GZ) brain cells (Wang et al., 2018). TADs overlapping with the de novo SVs are highlighted in red.*

candidate driver genes (Figure 3.3B). The validity of the approach was supported by the detection of pathogenic positional effects that have been identified in previous studies. For example, the regulatory landscape of *SOX9* was predicted to be disturbed by a translocation 721 Kb upstream of the gene in individual P5, whose phenotype is mainly characterized by acampomelic campomelic dysplasia with Pierre-Robin Syndrome (PRS) including a cleft palate (Figure S3.6). SVs in this region have been predicted to disrupt interactions of *SOX9* with several of its enhancers further upstream, leading to phenotypes similar to the phenotype of individual P5 (Amarillo et al., 2013; Benko et al., 2009). In individual P39, who has been previously included in other studies, our method predicted a disruption of *FOXG1* expression regulation by a translocation (Figure S3.4), further supporting the hypothesis that deregulation of *FOXG1* caused the phenotype of this individual (Mehrjouy et al., 2017; Redin et al., 2017).

Another example of a predicted positional effect is the disruption of the regulatory landscape of the *HOXD* locus in individual P22. This individual has complex genomic rearrangements consisting of 40 breakpoint junctions on four different chromosomes likely caused by chromothripsis (Cretu Stancu et al., 2017). One of the inversions and one of the translocations are located in the TAD upstream (centromeric) of the *HOXD* gene cluster (Figure 3.3C). This TAD contains multiple enhancers that regulate the precise expression patterns of the *HOXD* genes during the development of the digits (Andrey et al., 2013; Fabre et al., 2017; Rodríguez-Carballo et al., 2017). Deletions of the gene cluster itself, but also deletions upstream of the cluster are associated with hand malformations (Mitter et al., 2010; Montavon et al., 2012; Svensson et al., 2007). The translocation in individual P22 disrupts one of the main enhancer regions (the global control region (GCR)), which may have led to altered regulation of the expression of *HOXD* genes, ultimately causing brachydactyly and clinodactyly in this patient.

Our approach predicted positional effects on T1/T2 candidate driver genes in 10 included cases (26%). Most of these predicted positional effects were caused by breakpoint junctions of balanced SVs, suggesting that these effects may be especially important for balanced SVs.

3.2.4 Prediction of driver genes improves molecular diagnosis

By combining the directly and indirectly affected candidate drivers per patient we found possible explanations for the phenotypes of 16/39 (41%) complex and/or previously unsolved cases (Figure 3.4A). Interestingly, in eight cases we found evidence for multiple candidate drivers that are individually only associated with part of the phenotype, but together may largely explain the phenotype (Figure 3.4B). For example, we identified four candidate drivers in individual P25, who has a complex phenotype characterized by developmental delay, autism, seizures, renal agenesis, cryptorchidism

and an abnormal facial shape (Figure 3.4C). This individual has complex genomic rearrangements consisting of six breakpoint junctions and two deletions of ~10 Mb and ~0.6 Mb on three different chromosomes (Figure 3.4D). The 6q13–6q14.1 deletion of ~10 Mb affects 33 genes including the candidate drivers *PHIP* and *COL12A1,* which have been associated with developmental delay, anxiety and facial dysmorphisms in other patients (Engwerda et al., 2018; Webster et al., 2016). In addition to the two deleted drivers, two genes associated with other parts of the phenotype were predicted to be affected by positional effects (Figure 3.4E). One of these is *TFAP2A*, whose TAD (characterized by a large gene desert) and long-range interactions overlap with a translocation breakpoint junction. Rearrangements affecting the genomic interactions between TFAP2A and neural crest enhancers located in the *TFAP2A* TAD have recently been implicated in branchiooculofacial syndrome (Laugsch et al., 2019). The regulation of *BMP2*, a gene linked to agenesis of the ribs and cardiac features, is also predicted to be disturbed by a complex SV upstream of the gene (Kostina et al., 2018; Tan et al., 2017). Altogether, these candidate driver genes may have jointly contributed to the phenotype of this individual (Figure 3.4D). This case illustrates the challenge of identifying the causal genes driving the phenotypes of patients with structural rearrangements and highlights the notion that multiple genes should be considered for understanding the underlying molecular processes and explaining the patient's phenotype.

3.2.5 *In silico* prediction of candidate driver genes in larger patient cohorts

The candidate driver prioritization approach identified many candidate drivers in previously unresolved cases, but these complex cases may not be fully representative for the patient population seen in clinical genetic diagnostics. Therefore, we applied our prediction method to larger sets of patients with *de novo* SVs to further assess the validity and value of the approach. First, we determined the effects of largely balanced structural variants in 228 previously described patients (Figure S3.7A) (Redin et al., 2017). In 101 (44%) of the cases the detected *de novo* SVs were previously classified as pathogenic or likely pathogenic and in all but four of these diagnosed cases one or more candidate driver genes have been proposed (Figure S3.7b). Our approach identified 46 T1 and 92 T2 candidate phenodrivers out of 7406 genes located within 1 Mb of the SVs (Figure S3.7C,D). More than half (85/138) of the identified T1/T2 candidate drivers were not previously described as driver genes. In contrast, 23/114 (22%) previously described pathogenic or likely pathogenic drivers were classified as T3 candidates and 38/114 (33%) were not reported as driver by our approach (Figure 3.5A), mostly because the phenomatch scores are below the threshold (46%) or because the genes are not associated with HPO terms (41%) (Figure S3.7E). T1/T2 candidate drivers are identified in 99/225 (44%) of the individuals with mostly balanced SVs, including 31 individuals with SVs that were previously classified as VUS (Figure 3.5B, Figure S3.8).

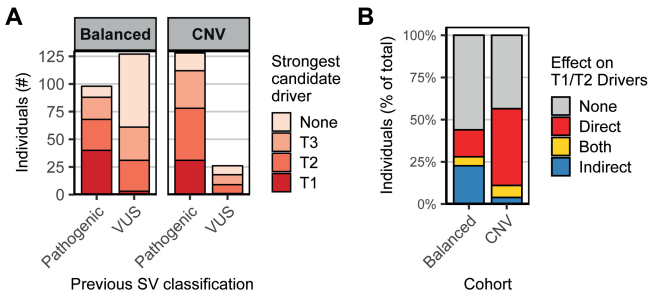


Figure 3.5 | In silico prediction of candidate drivers in larger cohorts of patients with de novo SVs. (A) *Comparison between previous SV classifications with the strongest candidate driver (located at or adjacent (<1 Mb) to these SVs) predicted by our approach. Two different patient cohorts, one containing mostly balanced SVs (Redin et al., 2017) and one containing copy number variants, were screened for candidate drivers. Our method identified T1/T2 candidate drivers for most SVs previously classified as pathogenic or likely pathogenic. Additionally, the method detected T1/T2 candidate drivers for some SVs previously classified as VUS, which may lead to a new molecular diagnosis.* ***(B)*** *Quantification of the predicted effects of the SVs on proposed T1/T2 candidate driver genes per cohort. Individuals with multiple directly and indirectly affected candidate drivers are grouped in the category described as "Both". Indirect positional effects of SVs on genes contributing to phenotypes appear to be more common in patients with balanced SVs compared to patients with copy number variants.*

Positional effect on genes moderately to strongly associated with the phenotypes are predicted in 63 of the cases (28%).

Subsequently, we also assessed the value of our driver prioritization approach for individuals with unbalanced copy number variants. We collected genetic and phenotypic information of 154 individuals with *de novo* copy number variants (<10 Mb) identified by clinical array-based copy number profiling (Figure S3.7A,B). The CNVs in the majority (83%) of these individuals have been previously classified as pathogenic according to clinical genetic diagnostic criteria (Figure S3.7B). These criteria are mostly based on overlap of the SVs with SVs of other individuals with similar phenotypes and the causative driver genes were typically not previously specified. Our method identified T1/T2 candidate driver genes in 87/154 (56%) individuals, including 9/26 individuals with CNVs previously classified as VUS (Figure 3.5A). Interestingly, support for positional effects on candidate drivers was only found in 11% of the cases with CNVs, suggesting that pathogenic positional effects are more common in patients with balanced SVs than in patients with unbalanced SVs (Figure 3.5B). No driver genes were identified for 39% of the previously considered pathogenic CNVs (based on recurrence in other patients). In some cases, potential drivers may remain unidentified because of incompleteness of the HPO database or insufficient description of the patient's phenotypes, but given the WGS results described for our patient cohort, it is also likely that some complexities of the CNVs may have been missed by the array-based detection method. It also suggests that many disease-causing genes or mechanisms are still not known or that some SVs are incorrectly classified as pathogenic.

3.3 Discussion

About half of the patients with neurodevelopmental disorders do not receive a diagnosis after regular genetic testing based on whole exome sequencing and microarray-based copy number profiling (Wright et al., 2018). Furthermore, the molecular mechanisms underlying the phenotype often remain unknown, even when a genetic variant is diagnosed as (potentially) pathogenic in a patient, as this is often only based on recurrence in patients with similar phenotype. Here, we applied an integrative method based on WGS, computational phenomatching and prediction of positional effects to improve the diagnosis and molecular understanding of disease aetiology of individuals with *de novo* SVs.

Our WGS approach identified additional complexities of the *de novo* SVs previously missed by arrays in 7 of 18 cases, supporting previous findings that WGS can have an added value in identifying additional SVs that are not routinely detected by arrays (Gilissen et al., 2014). The WGS results suggest that especially duplications can be more complex than can be determined with arrays, which is in line with previous studies (Brand et al., 2015; Newman et al., 2015). WGS can therefore be a valuable follow-up method to improve the diagnosis particularly of patients with duplications classified as VUS. Knowing the exact genomic location and orientation of SVs is important for the identification of possible positional effects.

To systematically dissect and understand the impact of *de novo* SVs, we developed a computational tool based on integration of HiC, RNA-seq and ChIP-seq datasets to predict positional effects of SVs on the regulation of gene expression. We combined these predictions with phenotype association information to identify candidate driver genes. In three of the cases we identified candidate drivers that are directly affected by the SVs. Positional effects of SVs have been shown to cause congenital disorders, but their significance is still unclear and they are not yet routinely screened for in genetic diagnostics (Spielmann et al., 2018). Our method predicted positional effects on genes associated with the phenotype in 25% and 8% of all studied cases with balanced and unbalanced *de novo* SVs, respectively. Previous studies estimated that disruptions of TAD organization may be the underlying cause of the phenotypes of ~7.3% patients with balanced rearrangements (Redin et al., 2017) and of ~11.8% of patients with large rare deletions (Ibn-Salem et al., 2014). Our method identified a higher contribution of positional effects in patients with balanced rearrangements mainly because our method included more extensive chromatin conformation datasets and also screened for effects that may explain smaller portions of the phenotypes. Our method, although it incorporates many published chromatin conformation datasets on untransformed human cells, focuses mainly on disruptions of interactions, which is a simplification of the complex nature of positional effects. It gives an insight in the potential effects that may lead to the phenotypes and prioritizes

candidates that can be followed up experimentally, ideally in a developmental context. SVs can affect many genes and multiple of these genes may together contribute to the phenotype. Indeed, in eight cases we found evidence for multiple candidate drivers affected by one or more *de novo* SVs. In many of the studied cases our method did not detect candidate drivers. This may be due to incomplete knowledge about disease-causing genes and/or due to missing disease associations in the used databases. Additionally, *de novo* SVs are also frequently identified in individuals without severe developmental disorders (Brandler et al., 2018; Collins et al., 2019; Kloosterman et al., 2015) and some of the detected SVs in the patients may be benign. The datasets underlying our computational workflow can be easily updated, enabling routine reanalysis of previously identified SVs. Moreover, our approach can be extended to study the consequences of SVs in different disease contexts such as cancer, in which SVs also play a major causal role.

3.4 Conclusions

Interpretation of SVs is important for clinical diagnosis of patients with developmental disorders, but it remains a challenge because SVs can have many different effects on multiple genes. We developed an approach to gain a detailed overview of the genes and regulatory elements affected by *de novo* SVs in patients with congenital disease. We show that WGS can be useful as a second-tier test to detect variants that are not detected by exome- and array-based approaches.

3.5 Methods

3.5.1 Patient selection and phenotyping

A total of 39 individuals with *de novo* germline SVs and an inconclusive diagnosis were included in this study. Individuals P1 to P21 and their biological parents were included at the University Medical Center Utrecht (the Netherlands) under study ID NL55260.041.15 15-736/M. Individual P22, previously described by Redin *et al.* as UTR22 (Redin et al., 2017), and her parents were included at the San Luigi University Hospital (Italy). For individuals P23 to P39, lymphoblastoid cell lines (LCL cell lines) were previously derived as part of the Developmental Genome Anatomy Project (DGAP) of the Brigham and Women's Hospital and Massachusetts General Hospital, Boston, Massachusetts, USA (Redin et al., 2017). Written informed consent was obtained for all included individuals and parents and the studies were approved by the respective institutional review boards.

3.5.2 DNA and RNA extraction

Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood samples

of individuals P1 to P22 and their biological parents using a Ficoll-Paque Plus gradient (GE Healthcare Life Sciences) in SepMate tubes (STEMCELL Technologies) according to the manufacturer's protocols. LCL cell lines derived from individuals P23 to P39 were expanded in RPMI 1640 medium supplemented with GlutaMAX (ThermoFisher Scientific), 10% fetal bovine serum, 1% penicillin, 1% streptomycin at 37°C. LCL cultures of each individual were split in three flasks and cultured separately for at least one week to obtain technical replicate samples for RNA isolation. Genomic DNA was isolated from the PBMCs or LCL cell lines using the QIASymphony DNA kit (Qiagen). Total RNA was isolated using the QIAsymphony RNA Kit (Qiagen) and RNA quality (RIN > 8) was determined using the Agilent RNA 6000 Nano Kit.

3.5.3 Whole-genome sequencing

Purified DNA was sheared to fragments of 400-500 bp using a Covaris sonicator. WGS libraries were prepared using the TruSeq DNA Nano Library Prep Kit (Illumina). WGS libraries were sequenced on an Illumina Hiseq X instrument generating 2x150 bp paired-end reads to a mean coverage depth of at least 30x. The WGS data was processed using an in-house Illumina analysis pipeline (https://github.com/UMCUGenetics/IAP). Briefly, reads were mapped to the CRCh37/hg19 human reference genome using BWA-0.7.5a using "BWA-MEM -t 12 -c 100 -M -R" (Li and Durbin, 2009). GATK IndelRealigner (McKenna et al., 2010) was used to realign the reads. Duplicated reads were removed using Sambamba markdup (Tarasov et al., 2015).

3.5.4 Structural variant calling and filtering

Raw SV candidates were called with Manta v0.29.5 using standard settings (Chen et al., 2015) and Delly v0.7.2 (Rausch et al., 2012) using the following settings: "-q 1 -s 9 -m 13 -u 5". Only Manta calls overlapping with breakpoint junctions called by Delly (+/- 100 basepairs) were selected. Rare SVs were selected by filtering against SV calls of 1000 Genomes (Sudmant et al., 2015) and against an inhouse database containing raw Manta SV calls of ~100 samples (https://github.com/UMCUGenetics/vcf-explorer). *De novo* SVs were identified in individuals P1 to P22 by filtering the SVs of the children against the Manta calls (+/- 100 basepairs) of the father and the mother. Filtered SV calls were manually inspected in the Integrative Genome Viewer (IGV). *De novo* breakpoint junctions of individuals P1 to P21 were validated by PCR using AmpliTaq gold (Thermo Scientific) under standard cycling conditions and by Sanger sequencing. Primers were designed using Primer3 software (Table S3.3). Breakpoint junction coordinates for individuals P22 to P39 were previously validated by PCR (Cretu Stancu et al., 2017; Redin et al., 2017).

3.5.5 Single nucleotide variant filtering

Single nucleotide variants and indels were called using GATK HaplotypeCaller. For

individuals P1 to P21 (whose parents were also sequenced), reads overlapping exons were selected and the Bench NGS Lab platform (Agilent-Cartagenia) was used to detect possible pathogenic *de novo* or recessive variants in the exome. *De novo* variants were only analyzed if they affect the protein structure of genes that are intolerant to missense and loss-of-function variants. Only putative protein changing homozygous and compound heterozygous variants with an allele frequency of <0.5% in ExAC (Lek et al., 2016) were reported.

3.5.6 RNA-sequencing and analysis

RNA-seq libraries were prepared using TruSeq Stranded Total RNA Library Prep Kit (Illumina) according to the manufacturer's protocol. RNA-seq libraries were pooled and sequenced on a NextSeq500 (Illumina) in 2x75bp paired-end mode. Processing of RNA sequencing data was performed using a custom in-house pipeline (https://github.com/UMCUGenetics/RNASeq). Briefly, reads were aligned to the CRCh37/hg19 human reference genome using STAR 2.4.2a (Dobin et al., 2013). The number of reads mapping to genes and exons were counted using HTSeq-count 0.6.1 (Anders et al., 2015). Data obtained from the PBMCs (Individuals P1 to P22) and the LCL cell lines (Individuals P23 to P39) were processed as separate datasets. The R-package DESeq2 was used to normalize raw read counts and to perform differential gene expression analysis for both datasets separately (Love et al., 2014). Genes with more than 0.5 reads per kilobase per million mapped reads (RPKM) were considered to be expressed.

3.5.7 Gene annotation

Gene information (including genomic positions, Ensembl IDs, HGNC symbols and Refseq IDs) was obtained from Ensembl (GRCh37) using the R-package biomaRt (v2.38) (Durinck et al., 2009). Genes containing a RefSeq mRNA ID and a HGNC symbol were considered as protein-coding genes. Genomic coordinates for the longest transcript were used if genes contained multiple RefSeq mRNA IDs. The list of 19,300 protein-coding genes was further annotated with 1) pLI, 2) RVIS, 3) haploinsufficiency (HI) scores, 4) OMIM identifiers and 5) DD2GP information for each gene (see Table S3.3 for data sources). A phenotypic score based on these five categories was determined for each gene. Modes of inheritance for each gene were retrieved from the HPO and DD2GP databases.

3.5.8 Computational prediction of effects of SVs on genes

For each patient, the genes located at or adjacent (< 2Mb) were selected. The HPO terms associated with these genes were matched to each individual HPO term assigned to the patient and to the combination of these HPO terms. For each gene, the number of phenomatchScores higher than 5 ("phenomatches") with individual patient HPO terms was calculated. The strength of the association (none, weak, medium or strong)

of each selected gene with the phenotype of the patient was determined based on the total phenomatchScore, the number of phenomatches, the mode of inheritance and the phenotypic score (Figure S3.4A, Table 3.1). Subsequently, potential direct and indirect effects of the SVs (none, weak or strong) on the genes were predicted (Figure S3.4A, Table 3.1). The predictions of indirect, positional effects are based on overlaps of TADs, enhancers, DHS connections, PCHiC interactions and V4C profiles with the SVs (Figure S3.4B-F, Table 3.1). These chromatin organization and epigenetic datasets of many different cell types were obtained from previous publications (see Table S3.3 for data sources). Genes with at least a weak phenotype association and a weak SV effect are considered as T3 candidate genes. Genes were classified as T1 candidate drivers if they have a strong association with the phenotype and are strongly affected by the SV. Genes classified as T2 candidate driver can have a weak/medium phenotype association combined with a strong SV effect or they can have a medium/strong phenotype association with a weak SV effect.

1. Phenotype association				
		Weak	Medium	Strong
Phenotypic score	pLI > 0.9 RVIS < 10 HI < 10 DD2GP OMIM	> 0	> 0	> 2
Mode of inheritance			AD/XD/XR+XY	AD/XD/XR+XY
phenomatchScore		> 0	> 4	> 10
Phenomatches with individual HPO terms	Score > 1 Score > 5	> 0	>10%	> 25%

2. Effect of SV on gene					
		Weak		Strong	
Gene location		Adjacent	DUP	Adjacent	DEL/TRUNC
Support score	TAD disrupted V4C disrupted PCHiC disrupted DHS disrupted RNA expression	> 1	NA	> 3	NA

3. Driver classification				
Classification	T3	T2		T1
Phenotype association	Weak	Strong	Medium	Strong
Effect of SV on gene	+ Weak	+ Weak	+ Strong	+ Strong

Table 3.1 | Cutoffs used to classify affected genes as T1, T2 or T3 candidate driver genes. *Driver classification is based on the phenotype association and the predicted effect of the SV on the gene (Fig. S3.4).*

3.5.9 SV and phenotype information large patient cohorts

Breakpoint junction information and HPO terms for 228 individuals (excluding the individuals already included in this study for WGS and RNA-seq analysis) with mostly balanced SVs were obtained from Redin et al (Redin et al., 2017). Phenotype and genomic information for 154 patients with *de novo* copy number variants ascertained by clinical genomic arrays were obtained from an inhouse patient database from the University Medical Center Utrecht (the Netherlands).

3.6 Supplements

3.6.1 List of abbreviations

HPO: Human Phenotype Ontology; kb: kilobase; RPKM: reads per kilobase per million mapped reads; SNV: Single nucleotide variant; SV: Structural variant; TAD: topologically associated domain; VUS: Variant of unknown significance; WGS: Whole-genome sequencing

3.6.2 Availability of data and material

Whole genome sequencing and RNA sequencing datasets generated during the study have been deposited in the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega) under accession number EGAS00001003489. All custom code has been made available on https://github.com/UMCUGenetics/Complex_SVs/. Additional file 2 containing the Supplemental Tables S3.1-S3.7 is available on bioRxiv (https://www.biorxiv.org/content/10.1101/707430v1.supplementary-material).

3.6.3 Competing interests

The authors declare that they have no competing interests.

3.6.4 Funding

This work is supported by the funding provided by the Netherlands Science Foundation (NWO) Vici grant (865.12.004) to Edwin Cuppen.

3.6.5 Authors' contributions

SM and JV performed experiments and bioinformatic analyses. JG and RH ascertained and enrolled individuals P1 to P21 and provided phenotypic information. JK and SM cultured LCL cell lines. NB and LdlF performed DNA and RNA isolations and lab support. SB, RJ, MJvR and WK provided computational support. EvB collected genomic and phenotypic information of individuals U1-U111. DG provided material for individual P22. MET provided LCL cell lines and information of individuals P22-P39. SM, JV, JG, JK, WK and EC designed the study. SM, JV and EC wrote the manuscript. All authors read

and approved the final manuscript.

3.6.6 Acknowledgements

We wish to thank all individuals who participated in this research. We thank Giulia Pregno and Giorgia Mandrile for the clinical and biological study of individual P22. We thank Utrecht Sequencing Facility (USEQ) for providing RNA-sequencing service and data. USEQ is subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University. We would also like to thank the Hartwig Medical Foundation for providing whole genome sequencing services.

3.6.7 Supplemental figures

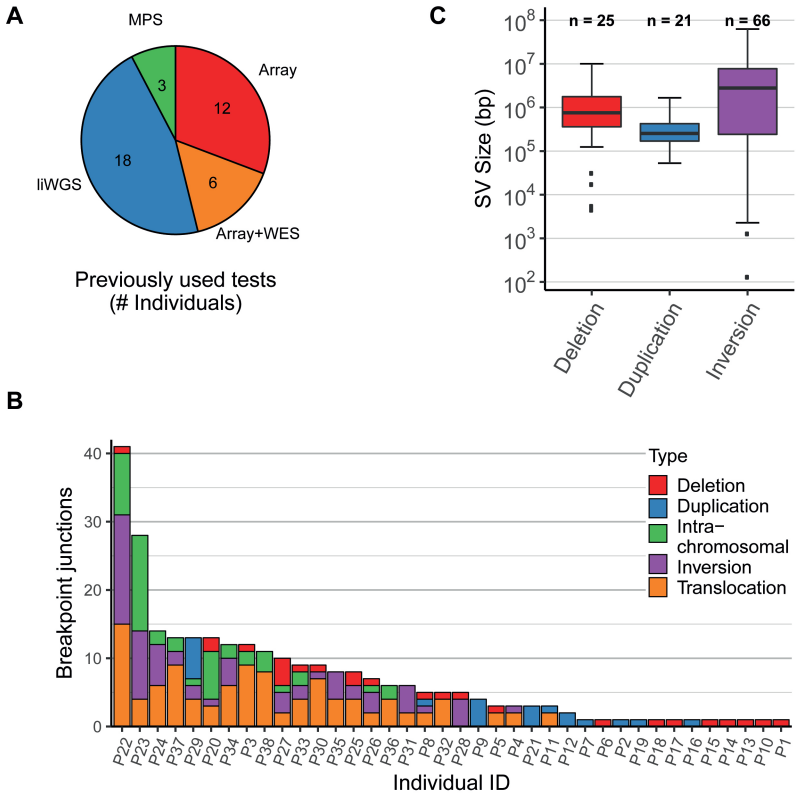


Figure S3.1 | Detected de novo germline SVs in 39 included patients. (A) *Genetic tests previously used in a clinical setting to identify the de novo SVs in the included individuals. Microarrays (ArrayCGH or SNP arrays) were used to detect the deletions and duplications in 18 of the included individuals. MPS: Mate-pair sequencing, WES: Whole Exome Sequencing, liWGS: long-insert Whole Genome Sequencing.* ***(B)*** *Number of identified de novo SV breakpoint junctions per individual.* ***(C)*** *Size distribution in base pairs (bp) of the identified de novo deletions (median size 757,378 bp), duplications (median size 253,729 bp) and inversions (median size 2,295,988 bp).*

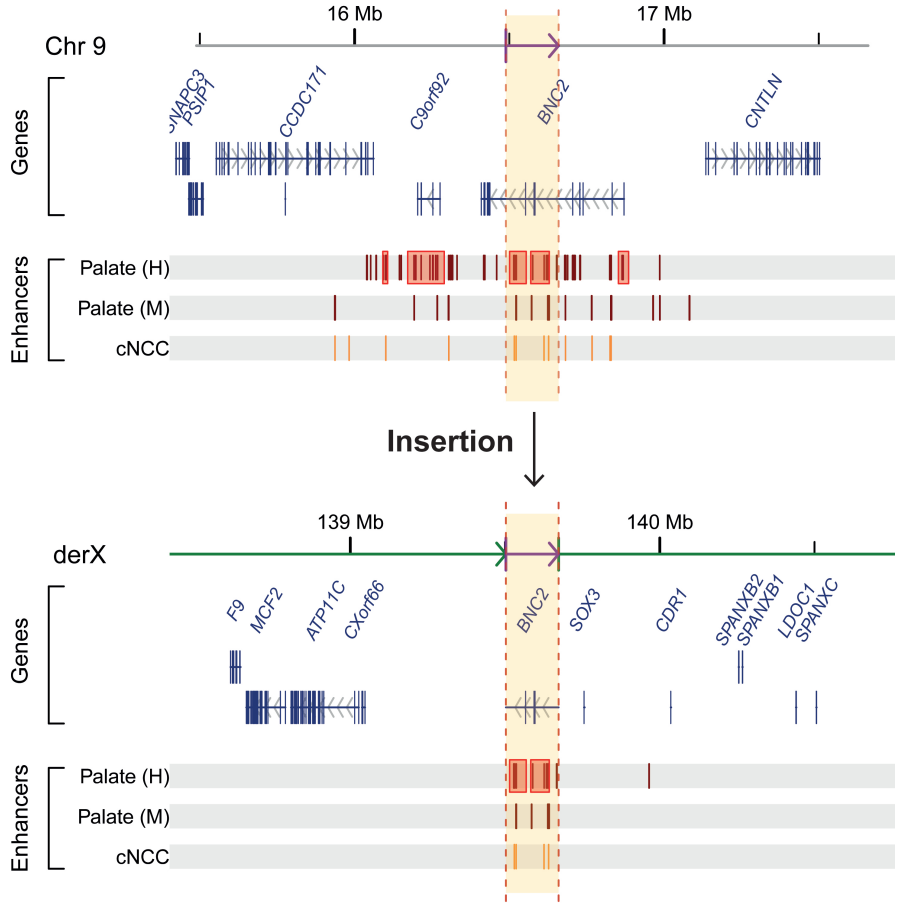


Figure S3.2 | Insertion of a super-enhancer region upstream of SOX3 detected by WGS in individual P11. *A 170kb duplication in the BNC2 gene body at chr9 was reported by array-based analysis (top panel), but WGS detected that this duplication is actually inserted in chrX (bottom panel). The fragment (highlighted in yellow) is inserted 82 kb upstream of the SOX3 gene. This locus at chrX contains a palindromic sequence that is susceptible for formation of genomic rearrangement. Multiple patients with varying phenotypes and different insertions at this locus have been described (Brewer et al., 2016; Bunyan et al., 2014; DeStefano et al., 2013; Haines et al., 2015; Zhu et al., 2011). The inserted fragment from chr9 contains multiple enhancers, including two previously described super-enhancer clusters (highlighted by red boxes), that are active in human (Palate (H), Carnegie stage 13) and mouse (Palate (M), embryonic day 11.5) craniofacial development and human cultured cranial neural crest cells (cNCC) (Attanasio et al., 2013; Prescott et al., 2015; Wilderman et al., 2018). The inserted enhancers may disturb the normal expression of the SOX3 gene and/or the surrounding genes, which may have led to the cleft palate phenotype in this patient. Genomic coordinates of mouse (mm9) embryonic craniofacial enhancers (determined by p300 ChIP-seq (Attanasio et al., 2013)) were converted to hg19 coordinates using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).*

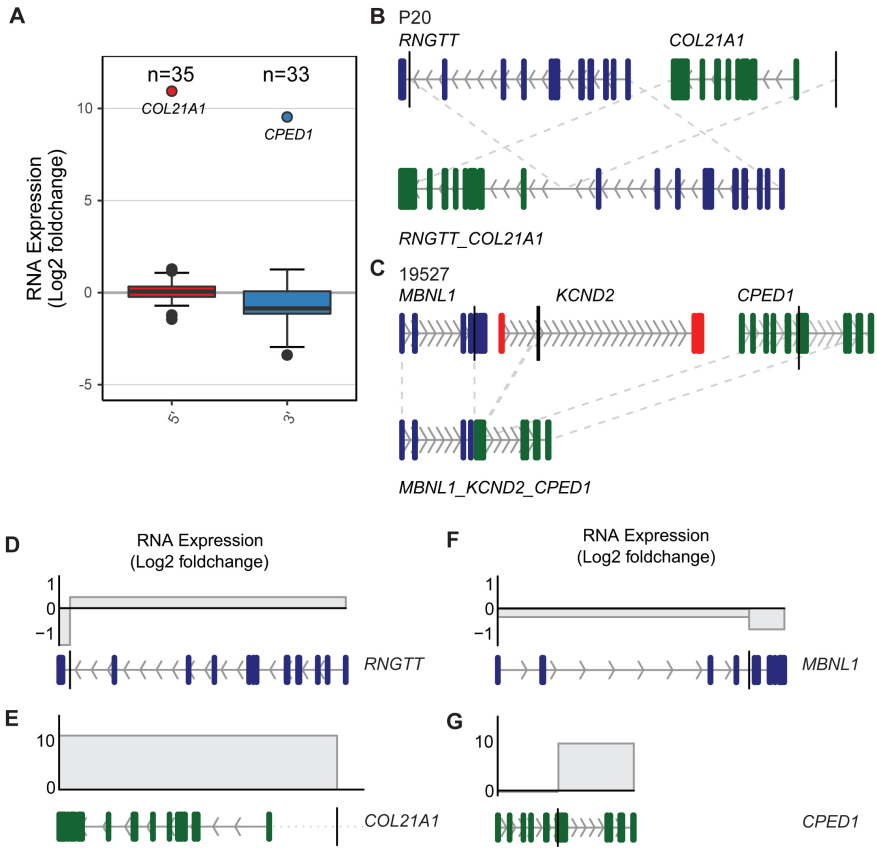
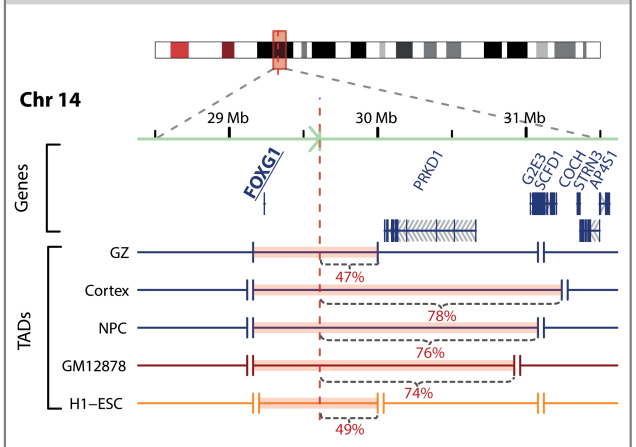


Figure S3.3 | RNA expression of genes truncated by de novo germline SVs. (A) *Log2 fold change expression values (compared to expression of the exons in control individuals) for 5' gene fragments and 3' gene fragments of truncated genes. The 5' fragment of COL21A1 and the 3' fragment of CPED1 show a strong overexpression due to a gene fusion.* ***(B)*** *Schematic representation of the RNGTT_COL21A1 fusion gene caused by genomic rearrangements in individual P20. The breakpoint junctions near the RNGTT (ENST00000369485) and COL21A1 (ENST00000244728) gene bodies are depicted by the vertical black lines.* ***(C)*** *Schematic reconstruction of the MBNL1_KCND2_CPED1 fusion gene in individual P34. Breakpoint junctions in the truncated genes MBNL1 (ENST00000324210), KCND2 (ENST00000331113) and CPED1 (ENST00000310396) are represented by the vertical black lines.* ***(D-G)*** *RNA log2 fold change expression values (compared to the expression in unaffected individuals) for the fragments of the truncated genes RNGTT, COL21A1, MLBNL1 and CPED1.*

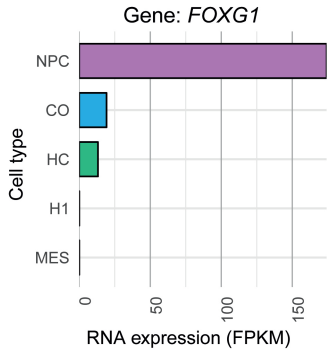
A Determine association between phenotype and genes at/adjacent to SVs

Gene	Seizures	Severe global developmental delay	Absent speech	Inability to walk	Blindness	phenoMatchScore	Phenomatches	pLI	RVIS	HI	DDG2P	OMIM	Inheritance	Phenotypic score
FOXG1	14.2	15.4	18.4	17.6	0.4	23.8	4/5		29.1	3.1	Yes	Yes	AD	3
DPYD	8.3	10	11.9	8	7.7	21.2	5/5	0	59	1.8		Yes	AR	2
AP4S1	8.1	9.3	11.7	16.7	0.9	21.1	4/5	0	61.3	19.1	Prob	Yes	AR	2
PCDH15	2.8	4.5	4.3	2.8	11.8	16.2	1/5	0	54.9	23.2		Yes	AR	1
WDR19	0.9	1.7	1.5	0.9	11.2	12.4	1/5	0	35.7	40.6	Yes	Yes	AR	2
LIAS	5.1	9.3	6	4.7	0.1	10.9	3/5	0.1	29.9	12.4	Prob	Yes	AR	2
SLC25A24	2.4	2.4	2.4	2.4	9	10.8	1/5	0	38.8	51.7	Yes	Yes	AD	2
NTNG1	5.3	5	5	5	0.1	6.6	3/5	0.4	28.7	3.5		Yes	ND	2
PRKD1	1.4	2.8	3.8	1.4	1.1	5	0/5	0	20.5	2.8	Yes	Yes	AD	3
ALG14	1.9	2.5	2.5	2.5	0.1	3.1	0/5	0	54.8	62.5		Yes	AR	1
COCH	0.4	0.4	0.4	0.4	0	0.4	0/5	0	46.9	15.1		Yes	AD	1

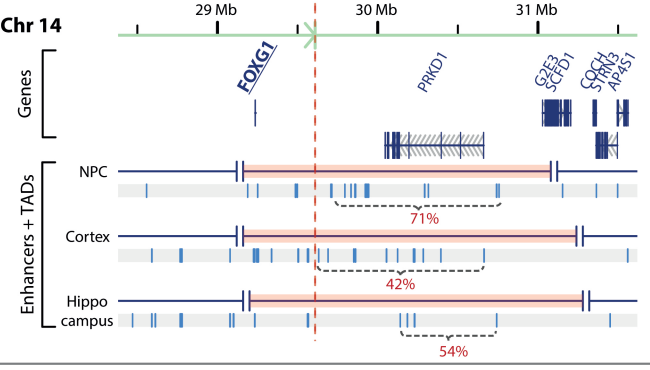
B Determine which TADs of 20 cell types overlap with the SVs and which percenge of the TAD is disrupted



C Select the 3 cell types with the highest RNA expression for each gene



D Determine the number of cell-type specific enhancers in the disrupted TADs up- and downstream of the breakpoint



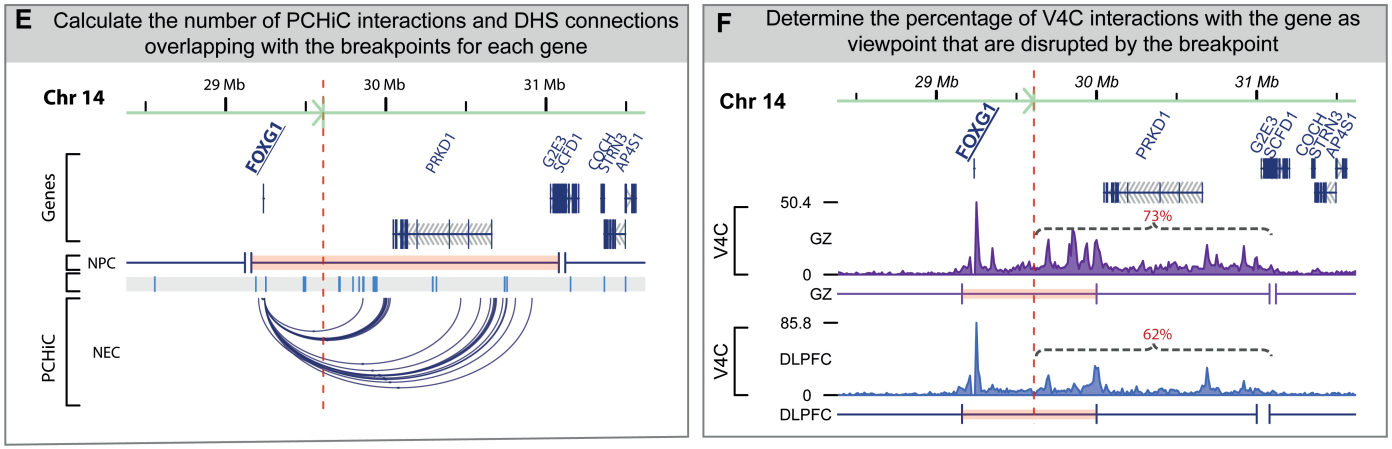
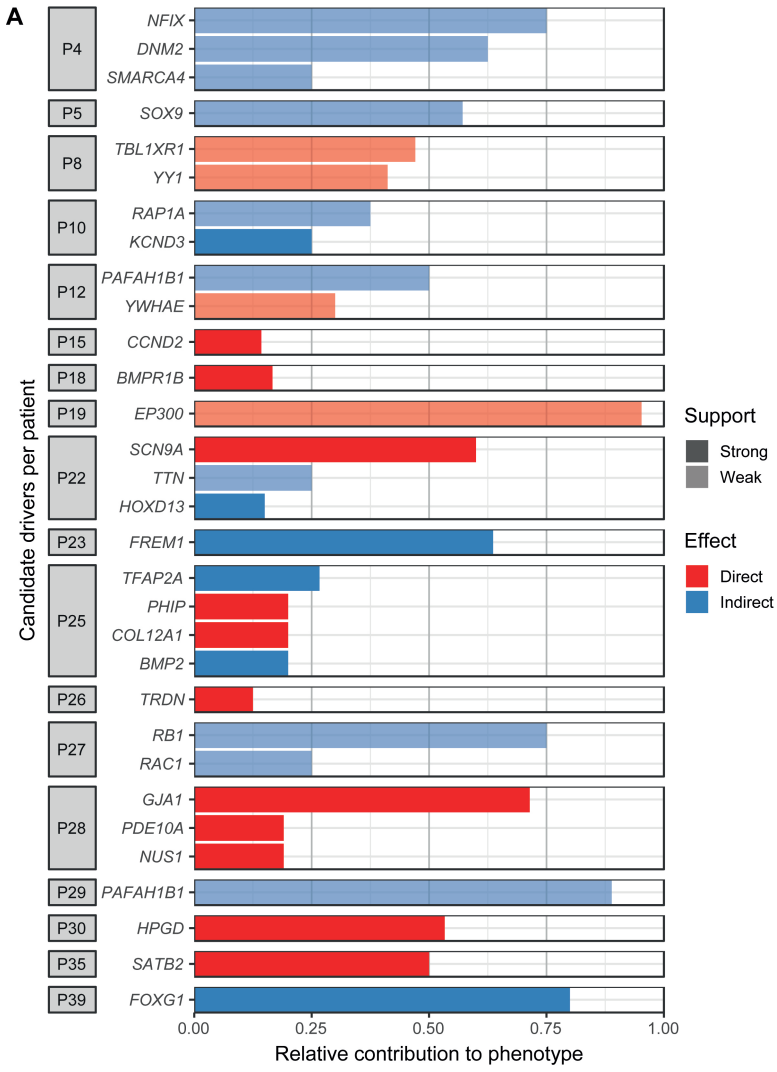
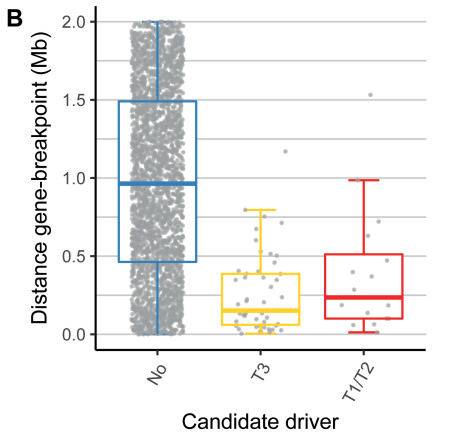
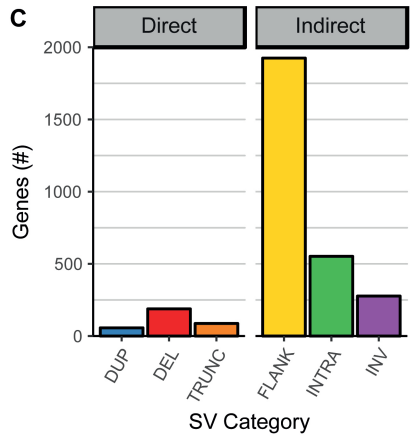


Fig. S4 Schematic overview of computational strategy used to predict positional effects. ***(A)*** *The association of a gene at or adjacent to an SV with the patient-specific phenotype is based on the Phenomatch score, the number of phenomatches, mode of inheritance and the phenotypic score. Each gene has a fixed phenotypic score (ranging from 1 to 5) based on its pLI (>0.9), RVIS (<10) and haploinsufficiency (HI, <10) scores and the presence of the gene in DDG2P and OMIM.* ***(B)*** *The TADs of 20 different cell types are overlapped with the SV breakpoint junctions of the individual. The TADs affected by a breakpoint are split into fragments up- and downstream of the breakpoint and the size of each fragment (relative to the size of the intact TAD) is calculated. Subsequently the genes located on each fragment are determined and each gene receives a score based on the relative size of the fragment. For example, 47% of the TAD in germinal zone (GZ) cells containing FOXG1 is considered disrupted.* ***(C)*** *For each gene the 3 cell types with the highest RNA expression (FPKM: Fragments Per Kilobase Million) based on the Encode/Roadmap ChIP-seq data are selected.* ***(D)*** *For each gene, enhancers from the three selected cell types are overlapped with the disrupted TAD fragments. The number of enhancers in the disrupted part of the TAD is compared to the number of enhancers in the TAD fragment containing the gene. This ratio is considered at the percentage of enhancers moved away from the gene (for example, the location of 71% of neural progenitor cell enhancers in the FOXG1TAD is changed).* ***(E)*** *For each gene, PCHiC interactions of 22 cell types and promoter-DHS connections are overlapped with the breakpoint junctions. The number of interactions overlapping with the junctions is divided by the total number of interactions of the gene. For example, for FOXG1all 107 PCHiC interactions (in multiple cell types) overlap with the breakpoint junction. f Virtual 4C profiles were generated for each gene and these were overlapped with the breakpoint junctions to determine the percentage of interactions that are located up- or downstream of the breakpoint junction. For FOXG1, 73% of the V4C interactions in dorsolateral prefrontal cortex (DLPFC) cells are considered to be disrupted.*


A
P4
NFIX
DNM2
SMARCA4
P5
SOX9
P8
TBL1XR1
YY1
P10
RAP1A
KCND3
P12
PAFAH1B1
YWHAE
P15
CCND2
P18
BMPR1B
P19
EP300
P22
SCN9A
TTN
HOXD13
P23
FREM1
P25
TFAP2A
PHIP
COL12A1
BMP2
P26
TRDN
P27
RB1
RAC1
P28
GJA1
PDE10A
NUS1
P29
PAFAH1B1
P30
HPGD
P35
SATB2
P39
FOXG1
Candidate drivers per patient
0.00
0.25
0.50
0.75
1.00
Relative contribution to phenotype
Support
Strong
Weak
Effect
Direct
Indirect


B
2.0
1.5
1.0
0.5
0.0
Distance gene-breakpoint (Mb)
No
T3
T1/T2
Candidate driver


C
Direct
Indirect
2000
1500
1000
500
0
Genes (#)
DUP
DEL
TRUNC
FLANK
INTRA
INV
SV Category

>>>

<<<

Figure S3.5 | Overview of the detected candidate driver genes. (A) *Relative contributions of the candidate drivers to the phenotypes of the individuals. The contributions are based on the number of Phenomatch hits (phenomatchScore > 5) of a gene with each individual HPO term assigned to an individual, e.g. a gene with a contribution of 0.75 is associated with 75% of the HPO terms of an individual. Shading indicates if there is relatively weak or strong evidence for an effect on the candidate driver.* ***(B)*** *Genomic distance (in base pairs) between the indirectly affected candidate driver genes (adjacent to the SVs) and the closest breakpoint junction. Most predicted candidate drivers are located within 1 Mb of a breakpoint junction.* ***(C)*** *Total number of analysed genes per SV category. DUP: Duplication, DEL: Deletion, TRUNC: Truncation, FLANK: Flanking region (+/- 2Mb), Intra: Intrachromosomal rearrangement, INV: Inversion.*

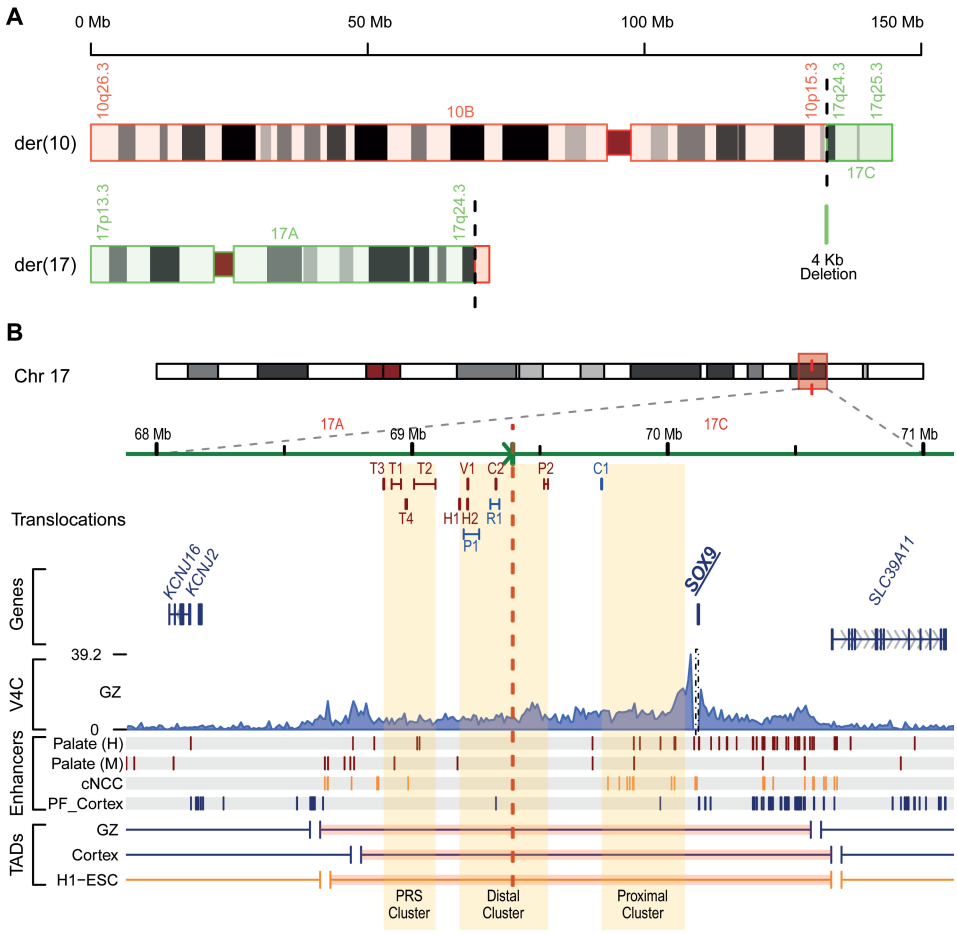


Figure S3.6 | Prediction of positional effects of a translocation on SOX9 in individual P5. (A) *Ideogram of the derivative chromosomes in individual P5. WGS identified a de novo translocation between chromosome 10 and 17 (46,XY,t(10;17)(p15;q24)). The breakpoint on chr17 (chr17:69395684, indicated by the vertical dotted red line) is 721 kb upstream of SOX9. A small 4kb fragment from chr17 is deleted (chr17:69391279-69395683).* ***(B)*** *Genome browser overview showing region surrounding the translocation breakpoint (red dotted line) at chromosome 17 in individual P5. The phenotype of this individual is characterized by acampomelic campomelic dysplasia and Pierre-Robin Syndrome including cleft palate, micrognathia and a long philtrum. SVs including translocations have been detected upstream of SOX9 in individuals with various phenotypes including campomelic dysplasia. The translocations found in patients with phenotypes including cleft palate are shown in red and translocations found in patients with different phenotypes are depicted in blue. These translocations are predicted to separate SOX9 from enhancers active in the developing palate, which may lead to the cleft palate phenotypes. Information about the other patients was obtained from the following publications: T1+T2+T3 (Benko et al., 2009); T4 (Jakobsen et al., 2007); C1+C2 (Leipoldt et al., 2007); P1+P2 (Fonseca et al., 2013); V1 (Velagaleti et al., 2005); R1 (Refai et al., 2010); H1+H2 (Hill-Harfe et al., 2005).*

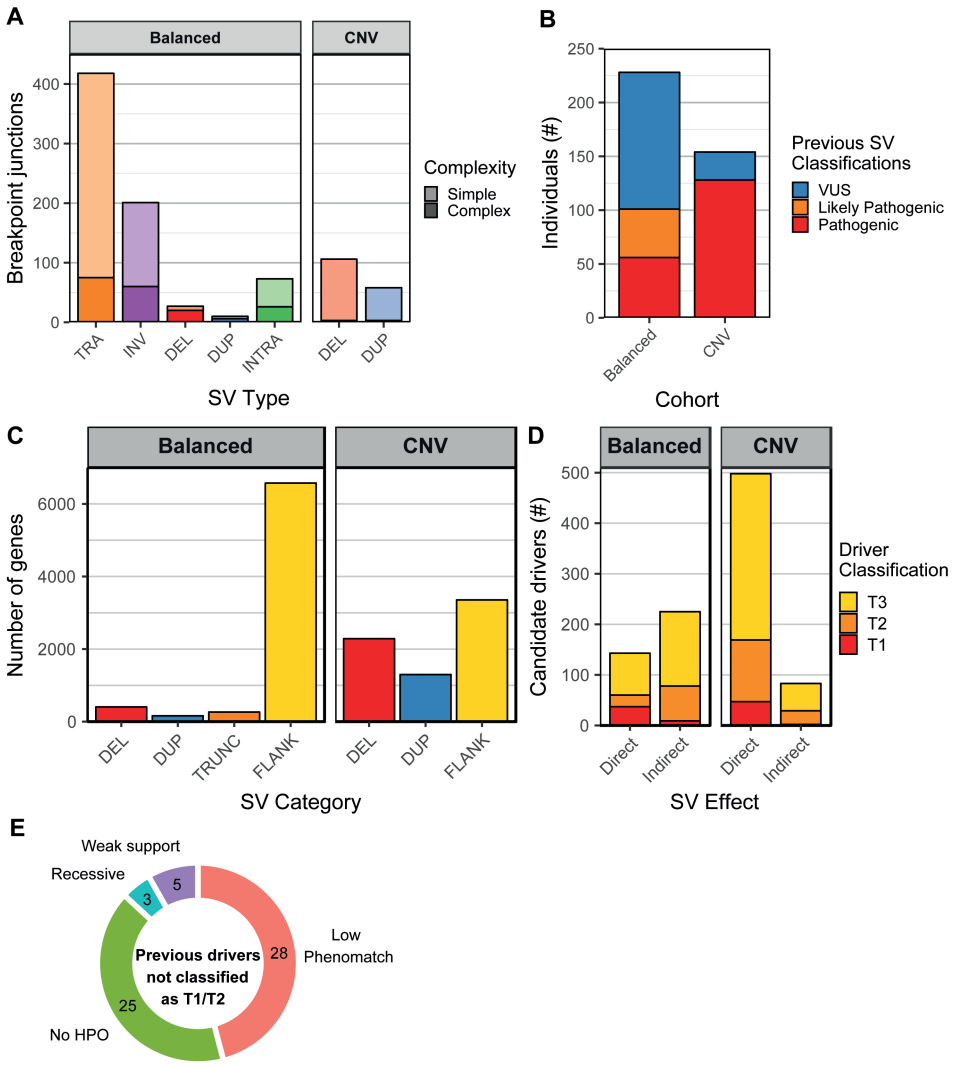


Figure S3.7 | Overview of SVs and candidate drivers in two cohorts of patients with de novo SVs. (A) *Quantification of previously identified de novo SVs in a cohort containing patients with mostly balanced SVs and a cohort containing patients with copy number variants (CNV). De novo translocations (TRA), inversions (INV) and intra-chromosomal rearrangements (INTRA) are most prevalent in the cohort of patients with balanced SVs. Some patients have complex genomic rearrangements (>3 SVs) including some deletions (DEL) or duplications (DUP). The cohort labelled as "CNV" consist of patients with relatively simple deletions and duplications (<10 Mb in size).* ***(B)*** *Number of patients whose de novo SVs were previously classified as pathogenic, likely pathogenic or variant of unknown significance (VUS) per cohort.* ***(C)*** *Total number of analysed genes per SV category in the two cohorts. Dup: Duplicated, Del: Deleted, Trunc: Truncated, Flank: Flanking SVs (<1 Mb).* ***(D)*** *Total number of predicted directly and indirectly affected candidate drivers per cohort.* ***(E)*** *Quantification of the genes that were previously classified as pathogenic or likely pathogenic (by (Redin et al., 2017)), but not identified as T1 or T2 candidate driver by our approach. These classification differences may be caused by a lack of HPO terms associated with the gene, low phenomatch scores below the threshold of our method, insufficient (weak) support for an effect of an SV on the gene detected by our method or a presumed recessive mode of inheritance.*

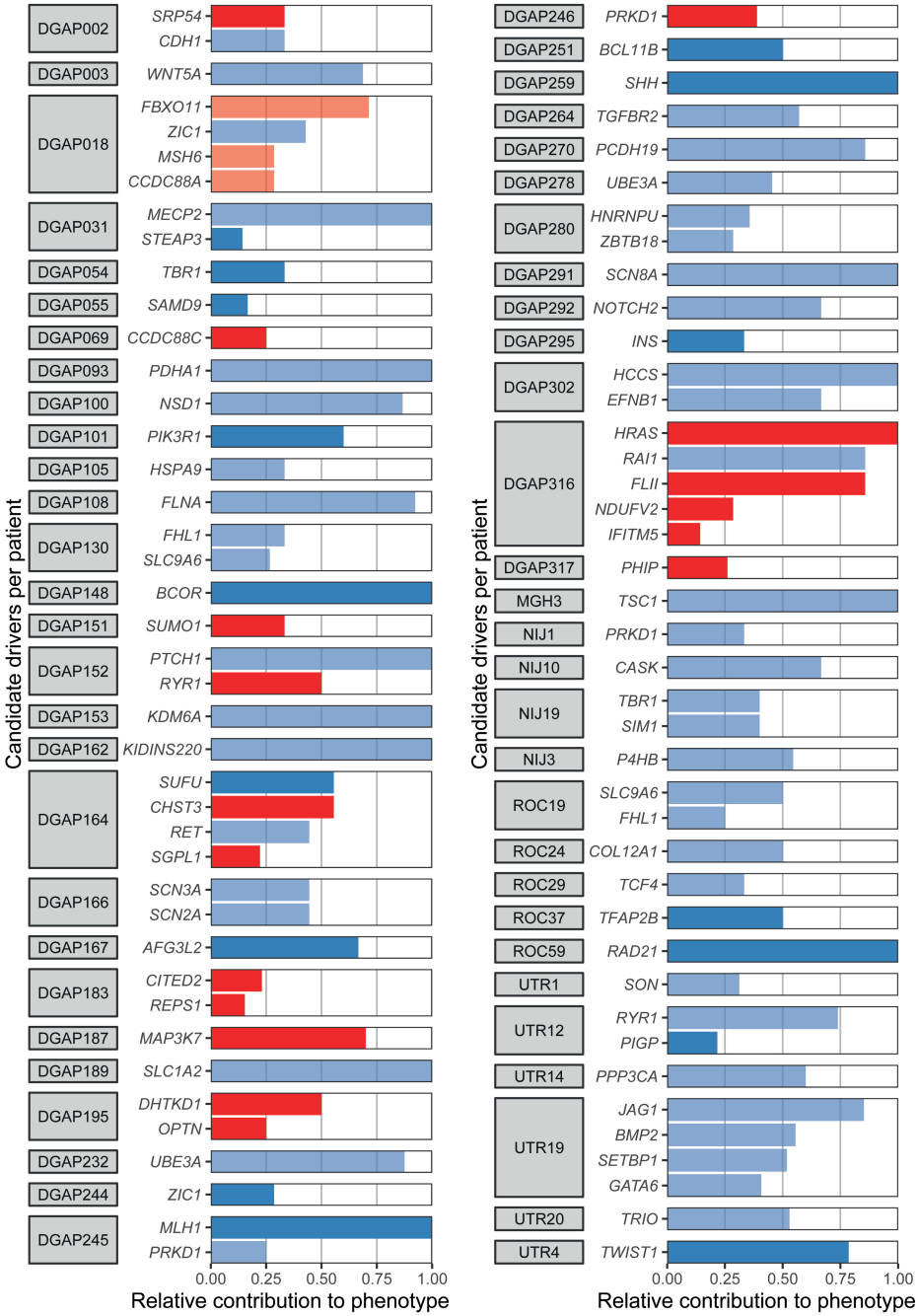
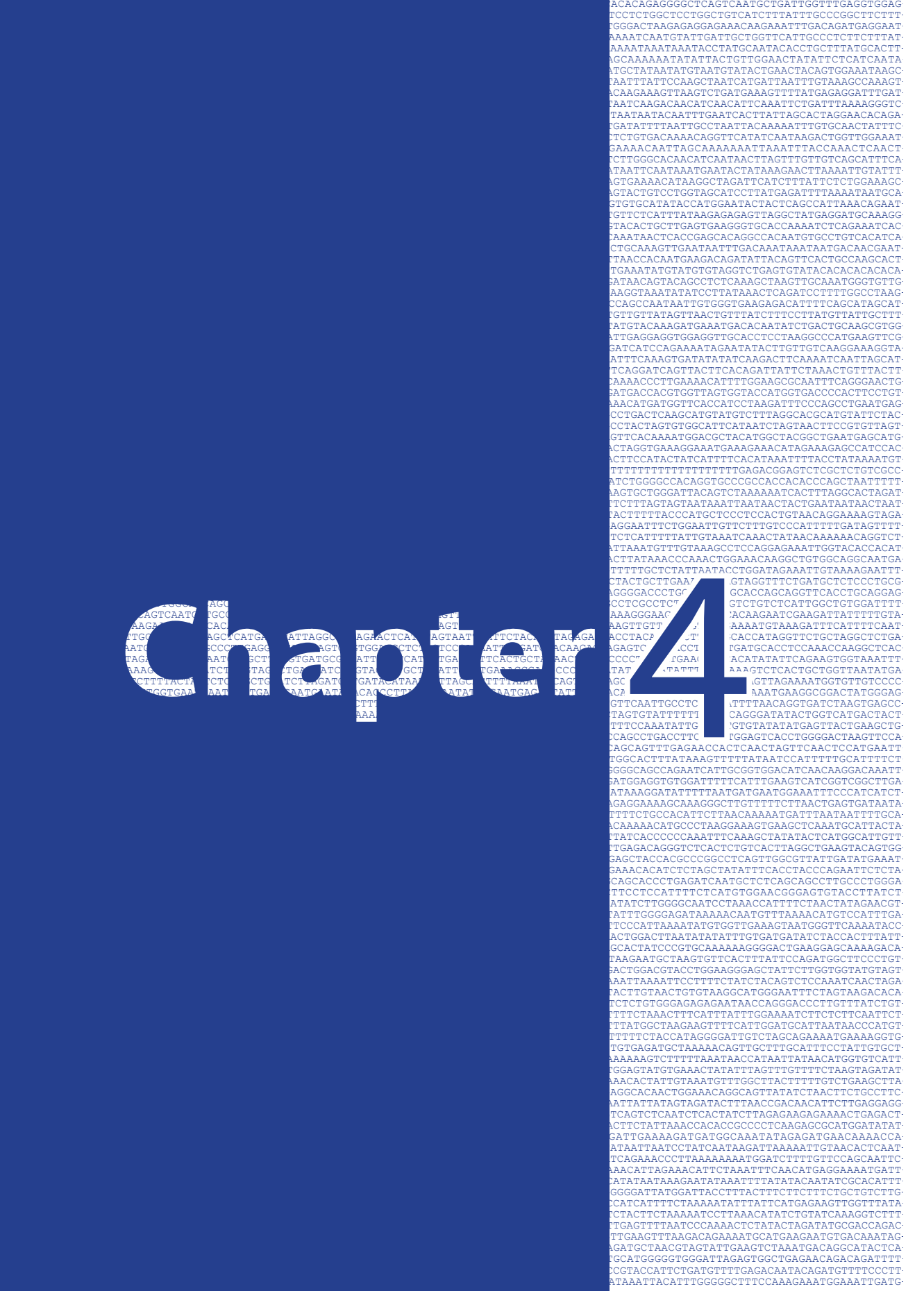


Figure S3.8 | Predicted contributions of candidate drivers to the phenotypes of patients with balanced structural variants of unknown significance. *T1/T2 candidate drivers were detected in 31 patients whose de novo SVs were previously classified as VUS by (Redin et al., 2017). The contributions to the phenotypes are based on the number of Phenomatch hits (phenomatchScore > 5) of the gene with each individual HPO term used to describe the phenotype of a patient. Shading indicates if there is relatively weak or strong evidence for an effect on the candidate driver.*


Chapter 4

Molecular dissection of germline chromothripsis in a developmental context using patient-derived iPS cells

Sjors Middelkamp, Sebastiaan van Heesch, A. Koen Braat, Joep de Ligt, Maarten van Iterson, Marieke Simonis, Markus J. van Roosmalen, Martijn J. E. Kelder, Evelien Kruisselbrink, Ron Hochstenbach, Nienke E. Verbeek, Elly F. Ippel, Youri Adolfs, R. Jeroen Pasterkamp, Wigard P. Kloosterman, Ewart W. Kuijk**, Edwin Cuppen**

*** Corresponding authors*

Adapted from:
Genome Medicine 9, 9 (2017)

Abstract

Background: Germline chromothripsis causes complex genomic rearrangements that are likely to affect multiple genes and their regulatory contexts. The contribution of individual rearrangements and affected genes to the phenotypes of patients with complex germline genomic rearrangements is generally unknown.

Methods: To dissect the impact of germline chromothripsis in a relevant developmental context, we performed trio-based RNA expression analysis on blood cells, induced pluripotent stem cells (iPSCs), and iPSC-derived neuronal cells from a patient with *de novo* germline chromothripsis and both healthy parents. In addition, Hi-C and 4C-seq experiments were performed to determine the effects of the genomic rearrangements on transcription regulation of genes in the proximity of the breakpoint junctions.

Results: Sixty-seven genes are located within 1 Mb of the complex chromothripsis rearrangements involving 17 breakpoints on four chromosomes. We find that three of these genes (*FOXP1*, *DPYD*, and *TWIST1*) are both associated with developmental disorders and differentially expressed in the patient. Interestingly, the effect on *TWIST1* expression was exclusively detectable in the patient's iPSC-derived neuronal cells, stressing the need for studying developmental disorders in the biologically relevant context. Chromosome conformation capture analyses show that *TWIST1* lost genomic interactions with several enhancers due to the chromothripsis event, which likely led to deregulation of *TWIST1* expression and contributed to the patient's craniosynostosis phenotype.

Conclusions: We demonstrate that a combination of patient-derived iPSC differentiation and trio-based molecular profiling is a powerful approach to improve the interpretation of pathogenic complex genomic rearrangements. Here we have applied this approach to identify misexpression of *TWIST1*, *FOXP1*, and *DPYD* as key contributors to the complex congenital phenotype resulting from germline chromothripsis rearrangements.

Keywords: Chromothripsis, Complex genomic rearrangements, Congenital disorders, Induced pluripotent stem cells, Neuronal differentiation, RNA-sequencing, Chromosome conformation capture, *TWIST1*, Craniosynostosis, Personal genomics

4.1 Background

Disruption of the genomic architecture by structural rearrangements such as translocations, deletions, duplications, and inversions is an important cause of congenital disease (Stankiewicz and Lupski, 2010). It has been estimated that approximately 15% of patients with multiple congenital abnormalities and/or mental retardation (MCA/MR) have a clinically relevant structural genomic rearrangement (Cooper et al., 2011; Hochstenbach et al., 2011; Kaminsky et al., 2011; Miller et al., 2010). Some of these patients have very complex combinations of structural variants resulting from chromothripsis, the local shattering and reassembly of one or a few chromosomes in a single event (Chiang et al., 2012; Kloosterman et al., 2012; Stephens et al., 2011). Chromothripsis can occur in both somatic cells, where it can contribute to cancer, and germline cells, where it can lead to congenital disorders (Kloosterman and Cuppen, 2013; Kloosterman et al., 2011; Stephens et al., 2011). Congenital chromothripsis cases with up to 57 breakpoints involving one to five chromosomes have been described (Kloosterman et al., 2012; Redin et al., 2017). Determining the molecular and phenotypic consequences of genomic rearrangements is a major challenge, especially for patients with complex rearrangements that involve large genomic regions of several megabases on multiple chromosomes containing many genes and regulatory elements (Kloosterman and Hochstenbach, 2014; Weischenfeldt et al., 2013). Structural rearrangements may lead to altered gene expression, gene fusions, disruption of regulatory elements such as enhancers and boundaries of topologically associated domains (TADs), and/or unmasking of recessive mutations in the unaffected allele (Kloosterman and Hochstenbach, 2014; Lupiáñez et al., 2016; Poot and Haaf, 2015; Spielmann and Mundlos, 2013; Weischenfeldt et al., 2013). Due to the large number of potentially affected genes in patients with complex rearrangements, the molecular mechanisms that have contributed to their congenital phenotypes are often unknown. Transcriptome analysis is a powerful method to determine the functional molecular consequences of structural rearrangements (Blumenthal et al., 2014; van Heesch et al., 2014; Luo et al., 2012; Schlattl et al., 2011). Patients' blood cells are commonly used as the source for RNA-seq analysis because of the relatively easy accessibility of this material. However, genes potentially involved in the disease of a patient may be expressed differently or not at all in blood compared to the disease-relevant tissue (Cai et al., 2010; Tylee et al., 2013). In addition, congenital disorders are typically the result of defects in developmental programs and it is questionable whether deregulation of developmental gene expression patterns persists in adult tissues. One approach that circumvents these concerns is to recapitulate certain developmental processes by generating induced pluripotent stem cells (iPSCs) from patients and differentiate these towards disease-relevant cell types (Avior et al., 2016; Bellin et al., 2012; Grskovic et al., 2011). This strategy has been applied successfully to improve our understanding of the molecular mechanisms underlying several (neuro-)

developmental diseases such as schizophrenia and Rett syndrome (Cundiff and Anderson, 2011; Dolmetsch and Geschwind, 2011).

We previously performed RNA-seq on blood samples of patients with germline chromothripsis and identified several molecular phenotypes caused by the genomic rearrangements (van Heesch et al., 2014). These included a hyper-activated trophoblast-specific miRNA cluster that interferes with embryonic brain development when ectopically expressed (van Heesch et al., 2014). However, in a second patient with MCA/MR the relevance of the identified molecular effects to the phenotype could not be entirely resolved due to the complexity of the rearrangements (van Heesch et al., 2014). In this study we further dissected the molecular consequences of chromothripsis by analyzing RNA expression and genome architecture in disease-relevant cell types derived from iPSCs from this patient and both parents.

4.2 Methods

4.2.1 Derivation and cultivation of iPSCs

Peripheral blood samples were obtained from a family trio consisting of the patient (child) with germline chromothripsis and both parents who served as controls. Peripheral blood mononuclear cells (PBMCs) were isolated by separation on a Ficoll-Paque TM PLUS gradient (GE Healthcare) with a density of 1.077 g/ml. Subsequently, CD34-positive cells were magnetically labeled with CD34-microbeads and purified with a CD34 Microbead kit (Miltenyi). The purified CD34-positive cells were resuspended in PBMC medium consisting of Iscove's modified Dulbecco's medium (ThermoFisher Scientific) with 5% fetal calf serum, 50 ng/ml stem cell factor, 50 ng/ml FLT3-ligand, 50 μM β-mercaptoethanol, 10 μg/ml penicillin, 10 μg/ml streptomycin, and 2 mM L glutamine, and plated in flat bottom 96-well ultra-low attachment plates. After 5 days, cells were passaged and the PBMC medium was further supplemented with 20 ng/ml interleukin (IL)-6 and 20 ng/ml thrombopoietin (TPO). After 7 days, cells were spin-transduced with 1 ml OSKM-dTOMATO lentivirus (Warlich et al., 2011) supplemented with 8 μg/ml polybrene, 50 ng/ml stem cell factor, 50 ng/ml FLT3 ligand, 20 ng/ml IL-6, and 20 ng/ml TPO at 1800 rpm at 32 °C for 100 minutes. Cells were subsequently incubated for 3 h, after which medium was changed to PBMC medium supplemented with IL-6 and TPO. The spin-transductions were repeated at day 9 and day 10 and cultures continued in PBMC medium supplemented with IL-6 and TPO. Subsequently all cells were seeded on irradiated mouse embryonic fibroblasts (Amsbio) and cultured in human embryonic stem cell (hESC) medium consisting of DMEM-F12 supplemented with 20% knock-out serum replacement, 10 μg/ml penicillin, 10 μg/ml streptomycin, 2 mM L-glutamine, 0.1 mM MEM-NEAA, 0.1 mM β-mercapthoethanol, and 10 ng/ ml basic fibroblast growth factor. The hESC medium was refreshed daily. Three clonal iPSC lines were derived from the patient, two lines from the father and one

from the mother. The iPSCs were subsequently adapted to and cultured on Geltrex-coated plastic (ThermoFisher Scientific) in serum- and feeder-free Essential-8 medium (ThermoFisher Scientific) with 1× penicillin-streptomycin (ThermoFisher Scientific). All cell lines were free of mycoplasm contamination.

4.2.2 Differentiation of iPSCs towards the neural lineage

Differentiation of the iPSCs to neural progenitors was performed according to the protocol by Shi et al. (Shi et al., 2012) with several modifications. iPSCs were prepared for neural induction by culturing cells in three wells of a six-well plate to 90% confluency on Vitronectin-coated plates using the Essential-8 medium, after which cells were passaged in a 1:2 ratio to Geltrex-coated six-well plates. Cells were then cultured until 95–100% confluency, upon which the medium was switched to neural induction medium. Neural induction medium was prepared with a 1:1 mixture of DMEM/F-12-Glutamax (Life Technologies) and Neurobasal medium (Life Technologies) with added 1× N-2 supplement (Life Technologies), 1× B-27 supplement (Life Technologies), 5 μg/ml insulin (Sigma), 2 mM L-glutamine (Life Technologies), 1× non-essential amino acids (Life Technologies), 100 μM β-mercaptoethanol (Life Technologies), 1 μM dorsomorphin (Sigma), and 10 μM SB431242 (Tocris Bioscience). Medium was replaced daily. RNA was collected at days 0, 7, and 10 of differentiation. At day 10, cells were passaged to laminin-coated coverslips for later immunofluorescent staining. Medium was then switched to neural maintenance medium (neural induction medium without dorsomorphin and SB431242), in which cells were cultured until formation of neural rosettes on day 15 after neural induction.

4.2.3 Immunofluorescent labeling of cultured cells

For immunofluorescent staining, cells were grown on coverslips, after which they were fixed in 4% paraformaldehyde for 15 minutes at room temperature (RT). Coverslips were then washed briefly in PBST (90% phosphate-buffered saline (PBS), 10% fetal bovine serum (FBS), 0.05% Triton X-100), permeabilized in permeabilization buffer (90% PBS, 10% FBS, 0.5% Triton X-100) for 15 minutes and blocked in PBST at RT for 1 h. Cover slips were incubated with primary antibody solution at RT for 1 hr. Primary antibodies were diluted in PBST to a concentration of 2 μg/ml. The primary antibodies used were mouse anti-NANOG (MABD24, EMD Millipore), Goat anti-OCT3/4 (sc-8628, Santa Cruz), Rabbit anti-SOX2 (AB5603, Chemicon), and Goat anti-PAX6 (PRB-278P-100, Covance Inc.). The coverslips were then washed three times with PBST at RT for 10 minutes. Next, the secondary antibody diluted in PBST to a concentration of 2 μg/ml was added and the samples were incubated in the dark at RT for 1 h. Secondary antibodies used are donkey anti-rabbit 488 (A-21206, Invitrogen), donkey anti-goat 568 (A-11057, Invitrogen), goat anti-mouse 633 (A-21050, Invitrogen) and rabbit anti-goat 488 (A-11055, Invitrogen). The coverslips were again washed three times with PBST at RT for

10 minutes. Finally, the coverslips were mounted using 3 µl Vectashield mounting medium with DAPI (H-1200, Vectorlabs), after which fluorescence was detected by confocal microscopy (Leica TCS SPE). The same acquisition settings were used for all samples throughout each experiment.

4.2.4 RNA extraction and sequencing

Samples for RNA sequencing were collected at days 0, 7, and 10 of neural differentiation of cell lines UMCU14 and UMCU15 from the patient, UMCU30 from the mother, and UMCU23 (with technical replicate) and UMCU32 from the father. RNA extraction was performed with Trizol (Life Technologies) according to the manufacturer's protocol. The isolated RNA was poly(A) selected with the MicroPoly(A) Purist Kit (Life Technologies) and a subsequent CAP-selection was performed with the mRNA ONLY Eukaryotic mRNA isolation kit (Epicentre/ Illumina). Next, the RNA was heat sheared followed by hybridization and ligation to the SOLID adapters according to the SOLID sequencing protocol. The RNA was subsequently reverse transcribed using the SOLID RT primer. After size-selection of the complementary DNA, it was amplified using a SOLID PCR primer and a unique barcoding primer for each library. Samples were sequenced on a SOLID Wildfire. RNA sequencing of patient and parental blood samples was performed previously (van Heesch et al., 2014).

4.2.5 Analysis of RNA sequencing data

Reads were mapped to the human reference genome (GRCh37/hg19) using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). The R package GenomicAlignments v1.6.3 was used to count reads overlapping exons (Lawrence et al., 2013). DESeq v1.22.1 was used to normalize read counts for library size and differential expression was calculated using the DESeq nBinomtest function (Anders and Huber, 2010). Hierarchical clustering based on the expression of the 500 genes with highest variance between all iPSC and neural progenitor cell (NPC) samples was performed using heatmap.2 from the gplots R package v2.17.0 (https://cran.r-project.org/web/packages/gplots/). Expression profiles of day 7 and day 10 NPCs clustered together and were therefore merged for downstream analysis (Figure S4.1). Genes with more than ten normalized counts were considered expressed genes. Molecular effects were defined as gene expression differences of at least twofold between patient and parents. Circos plots for data visualization were generated using Circos software (Krzywinski et al., 2009).

4.2.6 Hi-C data generation and analysis

iPSC-derived NPCs of the patient (lines UMCU14 and UMCU15) and the father

(UMCU23 and UMCU32) were crosslinked with 2% formaldehyde for 10 minutes. The crosslinking reaction was quenched by 0.125 M glycine. Following the crosslinking procedure, samples were centrifuged at 400 g at 4 °C for 8 minutes. Pelleted cells were washed with PBS and centrifuged again at 400 g at 4 °C for 5 minutes. Cells were lysed in 1 mL freshly prepared lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% Triton X-100, and 1× complete EDTA-free Protease Inhibitor Cocktail (Roche)) on ice for 10 minutes. Nuclei were washed twice in cold PBS after completion of the cell lysis.

Isolated and cross-linked NPC nuclei were digested with the DpnII restriction enzyme (New England Biolabs). Subsequently, the proximity ligation of interacting fragments was performed using T4 DNA ligase (Roche) to produce the 3C template, according to a previously described protocol by Simonis et al. (Simonis et al., 2006). After reverse crosslinking and precipitation, 10 μg of the template was sheared in microtubes (AFA Fiber Pre-Slit Snap-Cap 6 × 16 mm, 520045) using the Covaris S2 sonicator (1 cycle of 25 s; duty cycle 5%, intensity 3, 200 cycles per burst, frequency sweeping). Fragments that ranged in size from 500 to 1500 bp were selected using a 2% agarose gel. Size-selected fragments (1.1 μg) were used as the input for the TruSeq DNA Low Sample (LS) protocol (Illumina). Constructed libraries were size-selected using a LabChip XT DNA 750 Assay Kit (Caliper), resulting in libraries between 800 and 950 bp. These Hi-C libraries were sequenced in a paired-end manner on the Illumina HiSeq 2500, resulting in 2 × 100-bp reads. Sequenced read pairs were mapped independently using Burrows-Wheeler Aligner (BWA-0.7.5a; settings were bwa mem -c 100 -M) (Li and Durbin, 2009) to the human reference genome (hg19). Reads were further processed as previously described (Krijger et al., 2015).

4.2.7 4C-seq

4C-seq libraries were generated from crosslinked iPSC-derived NPCs of the patient (lines UMCU14 and UMCU15) and the father (UMCU23 and UMCU32) as previously described (Splinter et al., 2012). DpnII was used as primary restriction enzyme and NlaIII as secondary restriction enzyme. We PCR amplified 1.6 μg of each 4C template for each of the viewpoints using the primers listed in Table S4.1. The amplified 4C libraries were pooled, spiked with 30% Phi X 174 DNA, and sequenced on the Illumina NextSeq500 platform in paired-end mode. Data were processed as previously described (Van De Werken et al., 2012). The 4C-seq reads were normalized based on the total number of captured reads per sample. We analyzed 1.3 to 4.3 million mapped reads per viewpoint.

Locations of TADs in H1-hESC cells were determined by (Dixon et al., 2012) and obtained from http://promoter.bx.psu.edu/hi-c/download.html. Enhancer activity determined by expanded 18-state ChromHMM analysis of H1-derived NPCs

(E007) and primary foreskin fibroblasts (E056) was obtained from the Roadmap Epigenomics Mapping Consortium (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final). The dataset for the primary foreskin fibroblasts (E056) was selected because these cells have the highest *TWIST1* RNA expression of all cell types analyzed by the Roadmap Consortium (data not shown).

4.2.8 Molecular cloning

CNTN3 was amplified from a *CNTN3*-containing plasmid (RG221979 Origene). An In Fusion cloning kit (Clontech) was used to clone the amplicon into an empty plasmid with a pCAG promoter. High expression and proper cellular localization of *CNTN3* were confirmed by transfection of the pCAG *CNTN3* plasmid into HEK293 cells followed by western blotting and immunofluorescence with an antibody that recognizes CNTN3 (AF5539; R&D Systems; data not shown).

4.2.9 *In utero* electroporations of CNTN3 overexpression plasmids

Animal use and care was in accordance with institutional and national guidelines (Dierexperimentencommissie). At E14.5, pregnant C57Bl/6 mice were anesthetized using isoflurane (induction 3–4%, surgery 1.5–2%) and sedated with 0.05 mg/kg buprenorfin hydrochloride in saline. The abdominal cavity was opened and the uterine horns containing the embryos were carefully exposed. The lateral ventricles of the embryos were injected with linearized pCAG-*CNTN3* or control DNA (linearized Nes714tk/lacZ) vectors dissolved in 0.05% Fast Green using glass micro-pipettes (Harvard Apparatus). Nes714tk/lacZ was a gift from Urban Lendahl (Addgene plasmid #47614) (Lothian and Lendahl, 1997). pCAG-GFP was co-injected with the vectors to identify successfully electroporated cells. Developing cortices were targeted by electroporation with an ECM 830 Electro-Square-Porator (Harvard Apparatus) set to five unipolar pulses of 50 ms at 30 V (950-ms interval) using a platinum tweezer electrode holding the head (negative poles) and a third gold-plated Genepaddle electrode (positive pole) on top of the head (Fisher Scientific). Embryos were placed back into the abdomen and abdominal muscles and skin were sutured separately.

4.2.10 Immunofluorescent staining and analysis of brain sections

In utero electroporated embryos were collected at E16.5 and heads were fixed in 4% paraformaldehyde and submerged in 30% sucrose followed by freezing in 2-methylbutane. Sections of 20 μm were cut on a cryostat, mounted on Superfrost Plus slides (Fisher Scientific), air-dried, and stored at −20 °C until used for immunofluorescence. The sections were then blocked with 3% bovine serum albumin in PBS and 0.1% Triton, followed by an overnight incubation in rabbit

anti-GFP (A11122, ThermoFisher Scientific) diluted in blocking solution. After washing with PBS the sections were incubated in goat anti-rabbit 488 diluted in blocking solution. Finally, the sections were counterstained with Hoechst and embedded in Fluorsafe before mounting on the coverslips. Cortices were imaged using conventional confocal microscopy using a Zeiss confocal microscope. Adobe Illustrator was used to place consistent rectangles divided in eight equal square bins on top of the acquired images, so that bin 1 starts at the ventricle border of the tissue and bin 8 ends at the pial surface. The number of GFP-positive cells were counted in each bin and divided by the total amount of cells in the rectangle.

4.3 Results

4.3.1 Complex genomic rearrangements caused by chromothripsis in an MCA/MR patient

Previously we performed RNA-seq on blood samples of an MCA/MR patient with germline chromothripsis and both parents. The phenotype of this patient includes craniosynostosis (premature fusion of one or more cranial sutures), facial dysmorphisms, duplication of the right thumb, pre- and postnatal growth retardation, and intellectual disability. Mate-pair and breakpoint junction sequencing showed that the genome of the patient contains 17 breakpoints on chromosomes 1, 3, 7, and 12 (Figure 4.1A) (Kloosterman et al., 2012). Molecular phenotypes detected in blood could not entirely explain the patient's phenotype. Not all genes in proximity to the breakpoints were expressed in the patient's blood samples, so we hypothesized that essential molecular effects that may have contributed to the patient phenotype were undetectable in the patient blood samples.

To obtain cell types relevant for the disease phenotype we generated three iPSC lines from the germline chromothripsis patient and differentiated two of these to the neural lineage (Figure 4.1B). iPSCs were generated by reprogramming CD34-positive peripheral blood mononuclear cells (PBMCs) by transduction of a multicistronic lentiviral construct containing the canonical reprogramming factors (Takahashi et al., 2007; Warlich et al., 2011). We also successfully generated two control iPSC lines from the father and one line from the mother. Karyotyping confirmed the presence of all four derivative chromosomes in the patient's iPSC lines (Figure S4.2). One of the patient's cell lines contained a duplication of derivative chromosome 1 (Figure S4.2B). The paternal lines contained normal chromosome numbers, but the cell line of the mother has a translocation between chromosome 20 and part of chromosome 1 (Figure S4.2C). Because these karyotype abnormalities are distant from the breakpoints and because three of the five lines had the expected karyotypes, we concluded that

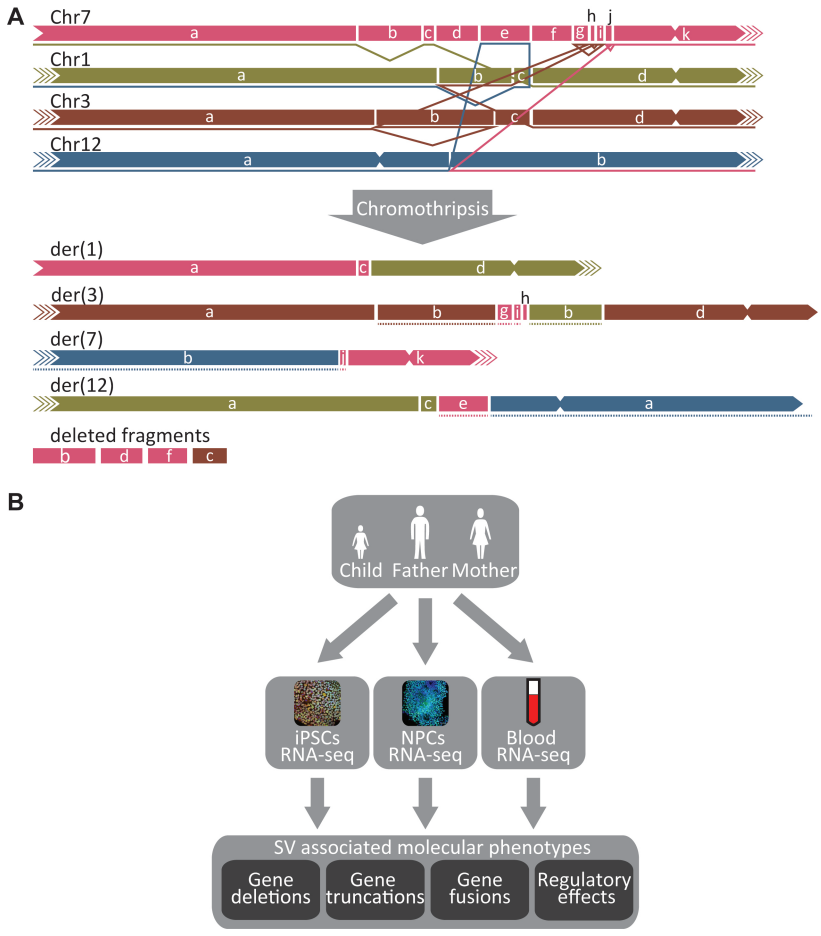


Figure 4.1 | Overview of complex chromosomal rearrangements in the patient with MCA/MR and study design. (A) *The breakpoint locations and genomic rearrangements on the four affected chromosomes in the germline chromothripsis patient determined by mate-pair and breakpoint fusion sequencing. Inversions are depicted with dashed lines beneath the derivative chromosomes. The four deleted fragments are shown below the derivative chromosomes. This illustration is adapted from (van Heesch et al., 2014).* ***(B)*** *Overview of the experimental setup of this study. Molecular effects of the chromosomal rearrangements on deleted, truncated, and fused genes and genes within 1 Mb of the rearrangements were determined by trio-based RNA-seq of iPSCs and iPSC-derived neuronal cells from the patient and both parents. These data were compared with previously generated expression data of blood samples of the patient and parents to identify molecular phenotypes that contribute to the patient's phenotype but are not detectable in blood (van Heesch et al., 2014).*

these lines were suitable to study the effects of the rearrangements within 1 Mb of the breakpoints. All iPSCs expressed the pluripotency-associated factors OCT4, SOX2, and NANOG, as determined by immunofluorescence and western blotting (Figure S4.3A,B). RNA-seq confirmed high expression of pluripotency factors in the iPSCs (Figure S4.3C). Neural progenitor cells (NPCs) derived from the patient's and parents' iPSCs formed neural rosettes and expressed early neural markers such as

PAX6, *OTX1*, *OTX2*, *SOX1*, and *SOX11* (Figure S4.4).

4.3.2 Molecular profiling of iPSC-derived neural progenitors

To identify molecular consequences of the chromothripsis rearrangements we performed RNA-seq on the iPSC lines and the iPSC-derived NPCs of the patient and the parents. We systematically analyzed the expression patterns of deleted genes, genes with disrupted coding sequences, and differentially expressed genes in close proximity to the breakpoints. Sixty-seven protein-coding genes are located across or within 1 Mb from the rearrangements (Figure 4.2; Table S4.2). Sixty (89%) of these are expressed in at least one of the samples. Ten genes are located on three deleted fragments (Figure 4.3; Figure S4.5). Four of these hemizygously deleted genes (*SNX13* (OMIM:606589), *TMEM106B* (OMIM:613413), *AHR* (OMIM:600253) and *ARL4A* (OMIM:604786)) show a consistent reduced expression in all patient's samples compared to the parents' samples (Figure 4.3; Figure S4.5). Although in theory the loss of these genes on the affected paternal alleles may have contributed to the patient's phenotype through haploinsufficiency, none of these genes have previously been associated with any of the patient's symptoms in the literature and were therefore considered unlikely to have played a major role in disturbing the development of the patient (Figure 4.3; Table S4.3).

4.3.3 Expression-dependent molecular effects on broken genes

The coding sequences of six genes have been interrupted by the rearrangements (Figure 4.4). Of these six disrupted genes, only *AGMO* (*TMEM195*) is not expressed in any of the assessed cell types. The 5′ part of *FOXP1* is fused to an inverted region on chromosome 7 containing parts of the *HDAC9* gene. The two disrupted genes are fused in opposite orientation and therefore do not directly form a fusion protein. However, we previously showed that there is read-through transcription from the 5′ part of *FOXP1* to the other strand of chromosome 7, leading to expression of a short fusion protein (van Heesch et al., 2014). The 5′ fused part of *FOXP1* is expressed at higher levels in the cells derived from the patient in comparison with the cells of the parents (Figure 4.4A). In contrast, the 3′ fragment of *FOXP1* shows a reduction in expression of 55% on average in the patient's cells (Figure 4.4A). The 3′ part of *ETV1* is fused to the 5′ part of *DPYD* and this *DPYD-ETV1* fusion gene shows strong expression in blood cells (van Heesch et al., 2014) but not in the iPSCs and iPSC-derived neural progenitors (Figure 4.4B,C). The expression of *DPYD-ETV1* is driven by the activity of the *DPYD* promoter, which is strong in blood but low in iPSCs and neural progenitors. The unaffected maternal *ETV1* allele is only expressed in the iPSCs and iPSC-derived neural progenitors, but at the RNA level expression of this allele cannot completely compensate for the loss of the paternal allele in these cell types

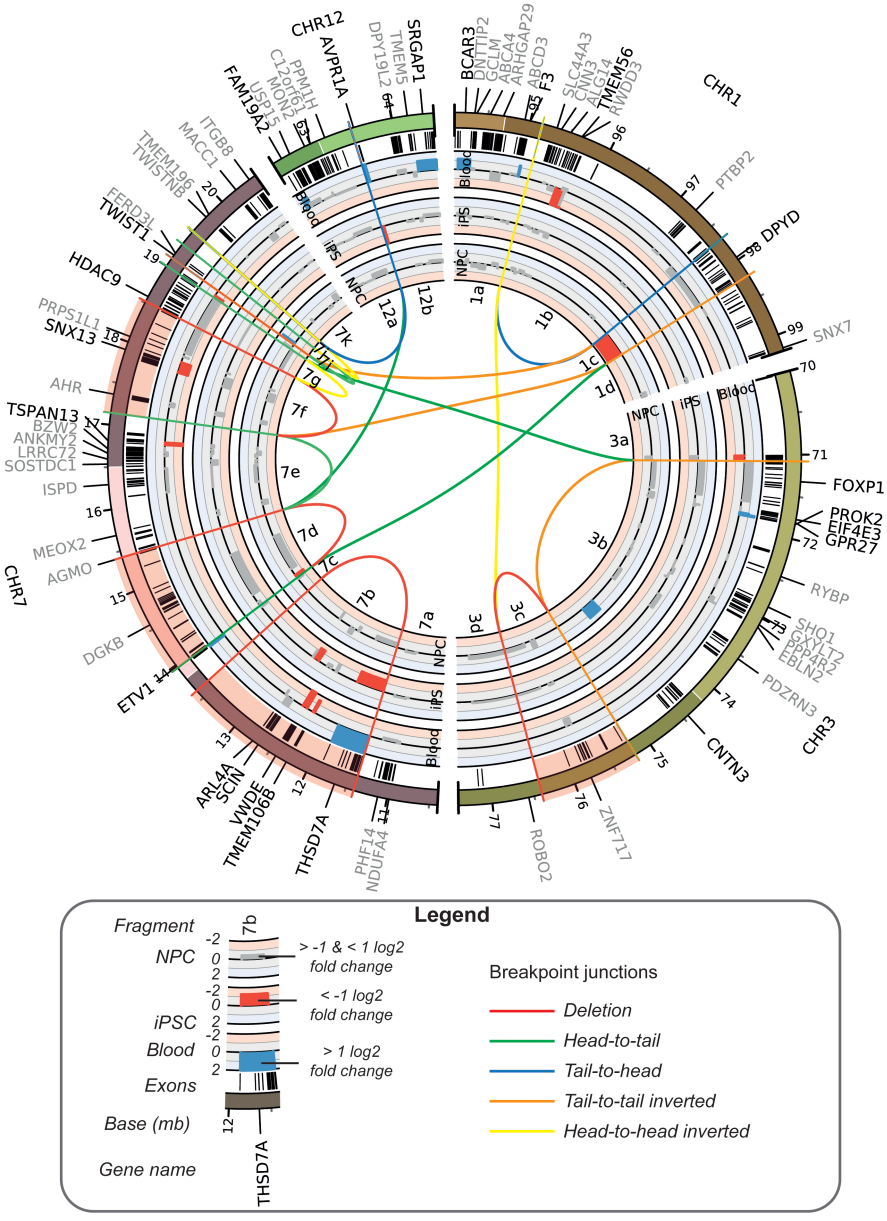


Figure 4.2 | Impact of chromothripsis on expression of genes in proximity to rearrangements. *Circos plot showing the regions affected by chromothripsis on patient chromosomes 1, 3, 7, and 12. The lines in the center of the plot visualize the 17 breakpoint junctions in the patient genome. In total, 67 genes, listed in the outer ring, are located on or within 1 Mb of the rearrangements. Exons are depicted as black bars beneath the chromosome ideograms. The inside, center, and outside bar graphs show the log2 fold expression differences (ranging from 2 to –2) between the patient and the parents in the iPSC-derived neural progenitors, the iPSCs, and the blood cells, respectively. Log2 fold expression differences of at least 1 between the patient and the parents are highlighted with blue (higher expression in patient) and red (lower expression in patient) bars. Grey bars indicate no or small (less than 1 log2 fold) expression differences between the patient and the parent. No bars are shown for genes with less than ten normalized read counts.*

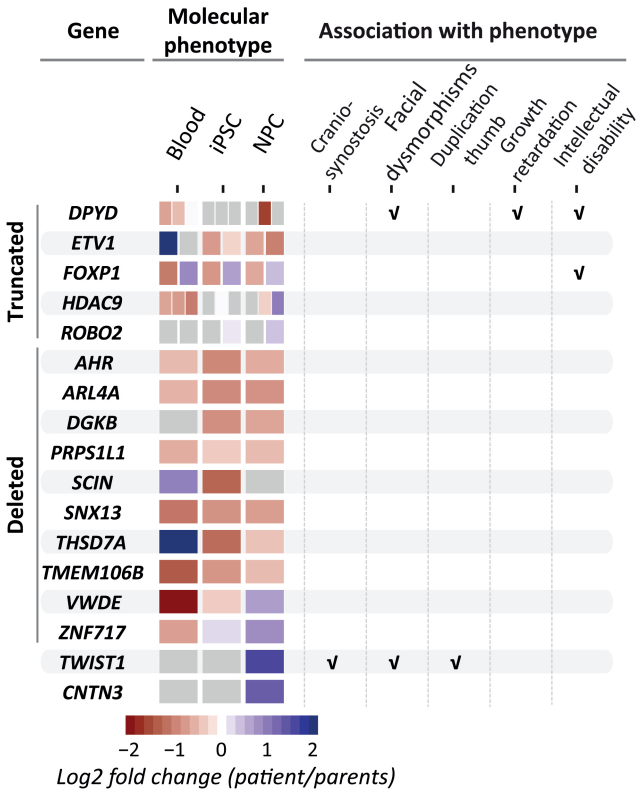


Figure 4.3 | Overview of molecular phenotypes and their association with the patient's phenotype. *Selection of the genes located near the breakpoints with affected coding sequences and/or altered expression. The heatmap indicates the log2 fold expression differences between the patient and the parents in the three different cell types. Expression changes of the truncated genes are split into separate boxes for each gene fragment. Grey boxes are shown for genes with less than ten normalized read counts. More details are provided in Table S4.2 and Table S4.3.*

(Figure 4.4C). Both *DPYD* and *HDAC9* are disrupted by two breakpoints, but these breakpoints only have a minor impact on the expression of these genes in the assessed cell types (van Heesch et al., 2014) (Figure 4.4B,D).

Of these six disrupted genes, *FOXP1* (OMIM:605515) and *DPYD* (OMIM:612779) are associated with (neuro-)developmental disorders and may thus be relevant for the patient phenotype (Figure 4.3; Table S4.3). *FOXP1* is an essential transcription factor involved in the development of many tissues, including the brain (Bacon and Rappold, 2012). Heterozygous disruptions of *FOXP1* have been found in several patients with neurodevelopmental disorders, including intellectual disability, autism spectrum disorder, and motor development delay (Bacon and Rappold, 2012). *DPYD* encodes DPD (dihydropyrimidine dehydrogenase), an enzyme involved in the catabolism of pyrimidine bases (Van Kuilenburg et

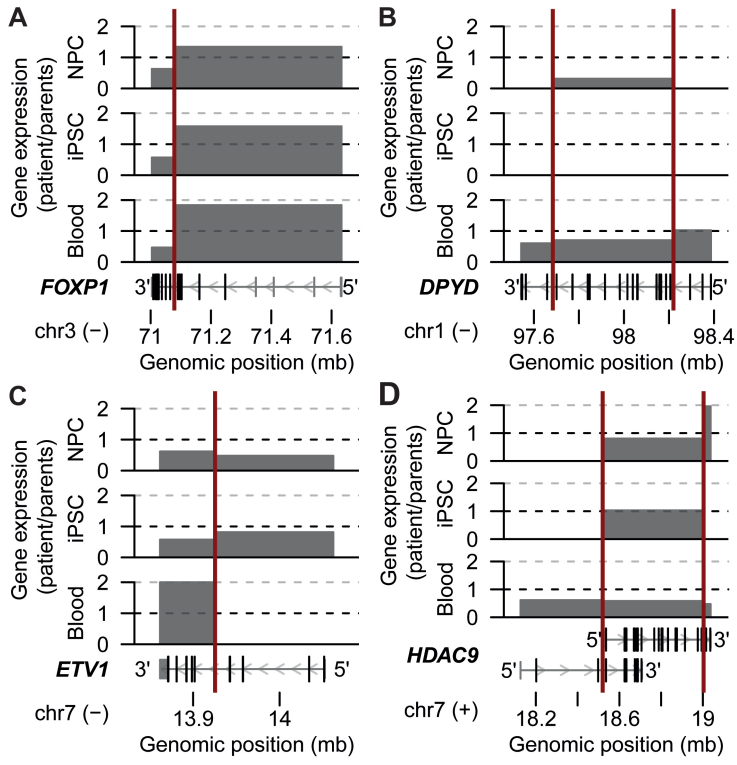


Figure 4.4 | Altered expression patterns of genes with disrupted coding sequences. *Relative expression differences of disrupted genes (a) FOXP1 (NM_032682), (b) DPYD (NM_000110), (c) ETV1 (NM_001163152), and (d) HDAC9 (NM_001204144 and NM_178423) between the patient and parents in the iPSC-derived NPCs, iPSCs, and blood cells. Gene structures of the RefSeq transcripts described above are shown below the graphs. Vertical red lines indicate the breakpoint locations. Minus and plus signs indicate the DNA strand. Expression is not shown for fragments with less than ten normalized read counts in the patient or the parents.*

al., 1999). Most carriers of heterozygous *DPYD* mutations are healthy, but some patients with hemizygous deletions affecting *DPYD* have neurodevelopmental disorders, including autism spectrum disorders (Carter et al., 2011; Pinto et al., 2014; Prasad et al., 2012), schizophrenia (Xu et al., 2012), epilepsy (Lal et al., 2015), and intellectual disability (D'Angelo et al., 2015; Van Kuilenburg et al., 1999; Willemsen et al., 2011). The disrupted coding sequences, altered expression, and association with congenital disease make it likely that the disruptions of *FOXP1* and possibly *DPYD* contributed to the developmental delay and intellectual disability of the patient. However, none of the broken or deleted genes have been associated with craniosynostosis, one of the major phenotypic appearances of the patient (Figure 4.3; Table S4.3).

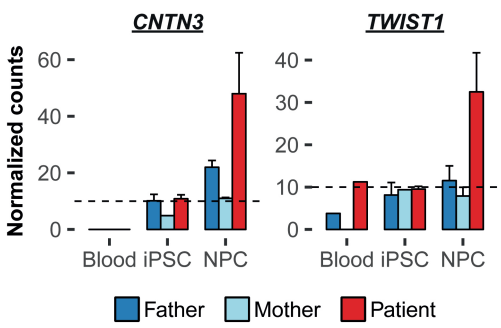


Figure 4.5 | Overexpression of TWIST1 and CNTN3 exclusively detectable in the iPSC-derived neural progenitors. *Bar graphs of CNTN3 and TWIST1 normalized gene expression in the blood cells, iPSCs, and iPSC-derived neural progenitors of the chromothripsis patient and the parents. The dashed horizontal line indicates the expression threshold of ten normalized read counts. Error bars indicate the standard error.*

4.3.4 Overexpression of *TWIST1* and *CNTN3* in the patient's iPSC-derived NPCs

Two genes that are located on inverted regions, but are not deleted or truncated, *TWIST1* and *CNTN3*, show a more than twofold difference in RNA expression in the NPCs derived from the patient in comparison to the parental cells (Figure 4.5). Both genes are hardly expressed in blood cells and the coding sequences of these genes are not disrupted by the rearrangements, indicating that positional effects rather than gene dosage cause their misexpression. *CNTN3* (also known as *contactin-3*, *PANG*, or *BIG-1*) is a member of the contactin family of neural cell adhesion molecules, but little is known about the specific functions of *CNTN3* (Mohebiany et al., 2014; Shimoda and Watanabe, 2009; Zuko et al., 2011). *CNTN3* is mainly expressed postnatally in specific subsets of neurons and promotes neurite outgrowth in isolated rat neurons (Mohebiany et al., 2014; Yoshihara et al., 1994). Copy number changes of close family members *CNTN4*, *CNTN5*, and *CNTN6* have been associated with autism spectrum disorders (Morrow et al., 2008; Zuko et al., 2013). We hypothesized that misexpression of *CNTN3* in neural progenitor cells may have affected the proper differentiation and migration of the patient's cortical neurons. To test this hypothesis we performed in utero electroporations of *CNTN3* overexpression plasmids in neural progenitors of the developing mouse cortices. In this experiment we did not detect any change in the migration of neurons in the cortical layers (Figure S4.6). We therefore consider it unlikely that misexpression of *CNTN3* has interfered with this developmental process in the patient.

4.3.5 Deregulation of *TWIST1* associated with patient's phenotype

The other overexpressed gene located near the breakpoints in the patient NPCs is *TWIST1*, a basic helix-loop-helix (bHLH) factor essential for mesoderm and neural crest development, including the morphology and migration of head mesenchyme cells (Qin et al., 2012). *TWIST1* mutations and deletions (OMIM: 601622) are the main cause of Saethre–Chotzen syndrome (OMIM: 101400), characterized by craniosynostosis and limb abnormalities, including polydactyly,

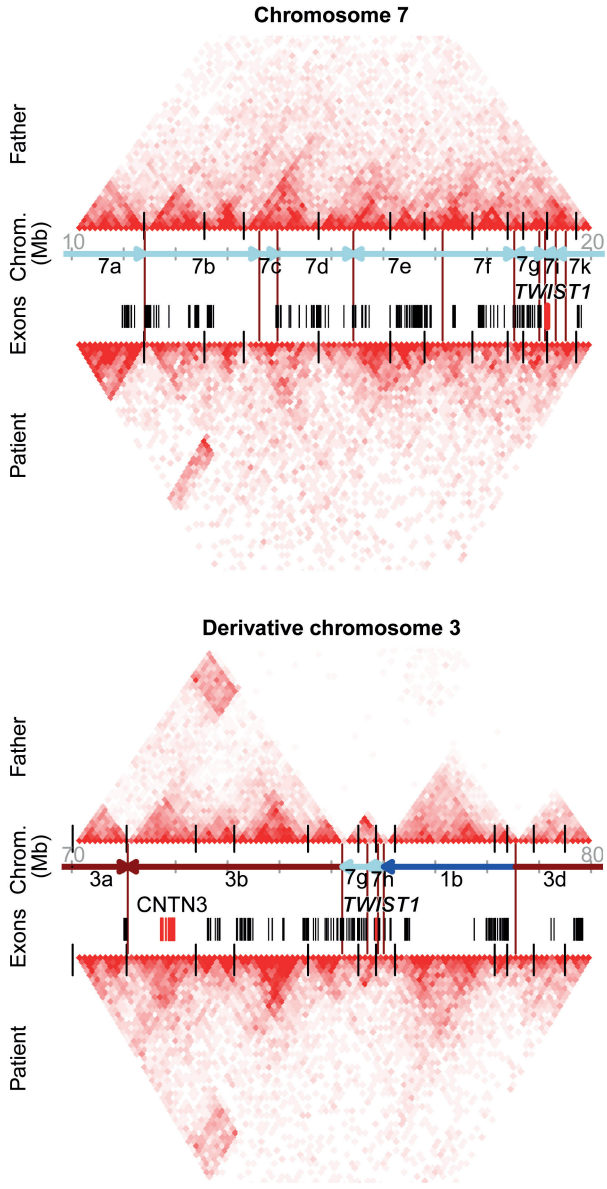


Figure 4.6 | Gains of genomic interactions on the derivative chromosomes of the patient. *Hi-C chromatin interaction maps of the patient's (cell line UMCU15, bottom panels) and father's (cell line UMCU23, top panels) chromosome 7 (left panels) and derivative chromosome 3 (right panels). Interactions are shown at 100-kb resolution. The vertical black lines at the bases of the heatmaps depict the predicted TAD boundaries in hESCs as determined by (Dixon et al., 2012). Vertical red lines between the interaction maps indicate the breakpoint locations in the patient.*

brachydactyly, and syndactyly (Cunningham et al., 2007; Twigg and Wilkie, 2015). Several craniosynostosis patients with translocation breakpoints near *TWIST1* have been described (Cai et al., 2003; Krebs et al., 1997; Rose et al., 1997). The phenotypes of these patients largely resemble the phenotype of the patient described in this study. Overexpression of *TWIST1* has been shown to inhibit osteoblast differentiation in vitro and overexpression of *Twist1* in mouse embryonic limbs lead to reduced limb size (Firulli et al., 2007; Funato et al., 2001; Lee et al., 1999). Ectopic *TWIST1* expression may disturb the balance between *TWIST1*, its dimerization partners such as *HAND2* and *TCF12*, and its competitors for binding partners (Connerney et al., 2006; Firulli et al., 2005; Sharma et al., 2013). In general, however, the phenotypes of patients with *TWIST1* mutations and deletions are linked to *TWIST1* haploinsufficiency (Twigg and Wilkie, 2015). In addition, trisomy of the 7p15.3pter locus including the *TWIST1* gene has been associated with delayed closure of the fontanels, the opposite phenotype of the patient described in this study and patients with *TWIST1* haploinsuffiency (Aswini et al., 2011; Stankiewicz et al., 2001).

The overexpression of *TWIST1* in the NPCs derived from the patient indicates a disturbed transcription regulation. We hypothesized that this deregulation may have led to decreased *TWIST1* expression in neural crest and mesodermal cell types, resulting in a phenotype parallel to that of patients who have haploinsufficiency of this gene. To test this hypothesis, we investigated the regulatory landscape surrounding the *TWIST1* gene. First we performed Hi-C to determine the genomic interactions on the derivative chromosomes in the patient. The topologically associated domain (TAD) structures of the unaffected chromosomes of the patient and father are similar to the previously published TAD structures by Dixon and colleagues (Dixon et al., 2012) (Figure 4.6, Figure S4.7). Disruption of TAD boundaries can cause ectopic interactions between gene promoters and enhancers and this may lead to disease (Lupiáñez et al., 2016). Thirteen TADs are directly affected by the breakpoints in the patient and five TAD boundaries are deleted (Figure 4.6; Figure S4.7). Many ectopic genomic interactions cross the breakpoint junctions on the derivative chromosomes of the patient. For example, many interactions between the genomic regions of chromosome 1, 3, and 7 that form derivative chromosome 3 in the patient are not present in the father (Figure 4.6). We could not precisely discern between reads of the unaffected maternal and affected paternal alleles and therefore could not specifically determine the genomic architecture of the derivative chromosomes.

Secondly, we performed 4C-seq on the NPCs of the patient and the father using *TWIST1* as bait to determine potential gains and losses of genomic interactions of *TWIST1* in the patient. *TWIST1* mostly interacts with a region encompassing three putative TADs in the NPCs of the father (Figure 4.7A). These

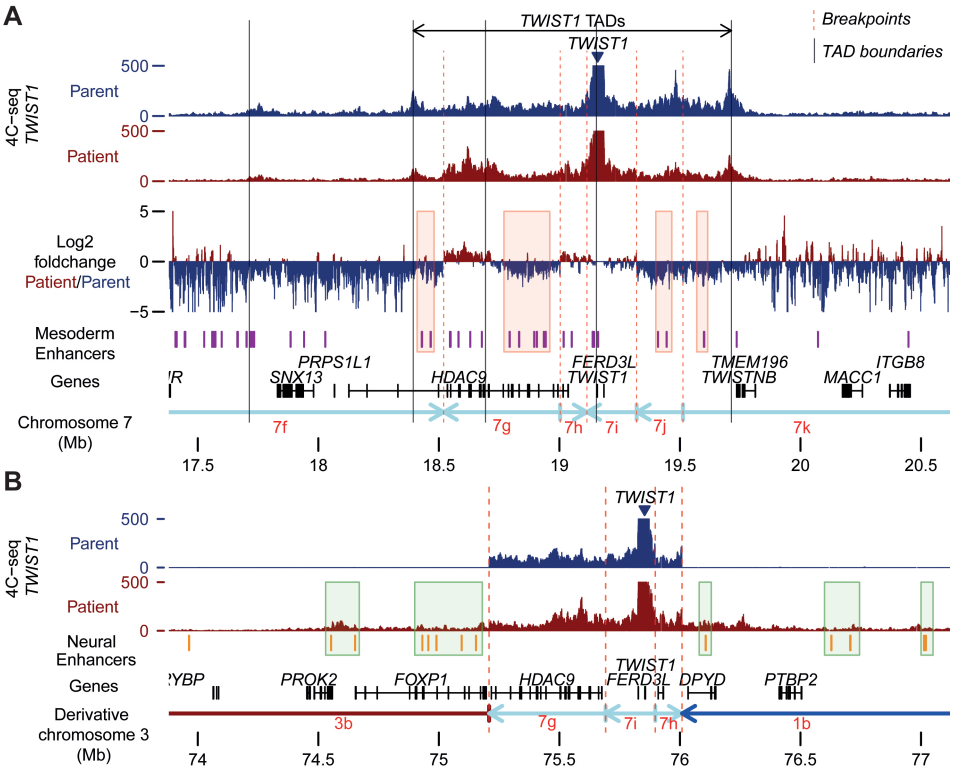


Figure 4.7 | Gains and losses of enhancer interactions with the TWIST1 locus in the patient. (A) *4C-seq data show that TWIST1 mainly contacts a region encompassing three TADs (termed TWIST1 TADs) in the NPCs of the father (cell line UMCU23). The y-axis indicates the number of normalized 4C-seq reads cutoff at 500 normalized reads. TAD boundaries in H1-ESCs were determined by Hi-C analysis by (Dixon et al., 2012). ChromHMM analysis of Roadmap ChIP-seq data of primary fibroblasts with high TWIST1 expression indicates that these TWIST1 TADs contain multiple enhancers active in mesodermal cells (shown in purple). The TWIST1 4C-seq data of the patient's NPCs (UMCU15) shows that TWIST1 has reduced interactions with several of these enhancers (red highlights), which likely had an impact on TWIST1 expression in the patient.* ***(B)*** *The 4C-seq data, depicted on the derivative chromosome 3 in the patient, shows that TWIST1 gained several ectopic contacts with enhancers active in neural cells in the patient. Enhancer activity was obtained from ChromHMM analysis of Roadmap ChIP-seq data of NPCs derived from differentiation of hESCs. 4C-seq using two of these enhancers as baits confirms the ectopic interactions between the enhancers and TWIST1 (Figure S4.8). These ectopic interactions may explain the overexpression of TWIST1 in the patient's NPCs.*

three TADs are disrupted by five breakpoints in the patient and parts of these TADs are inverted or translocated away from *TWIST1*. These disrupted *TWIST1* TADs contain several mesodermal enhancers active in cells with high *TWIST1* expression and known *TWIST1* enhancers (Figure 4.7A) (Birnbaum et al., 2012; Ernst and Kellis, 2012; Siekmann et al., 2015). The *TWIST1* 4C-seq shows that there are losses of interactions between these enhancers and *TWIST1* in the patient (Figure 4.7A, red highlights). These losses of contacts with several of its enhancers could lead to

reduced *TWIST1* expression in neural crest-derived cells involved in craniosynostosis and possibly contribute to the craniosynostosis phenotype (Twigg and Wilkie, 2015).

In addition, the 4C-seq experiments show that *TWIST1* gained aberrant interactions with several enhancers active in neural progenitor cells (Figure 4.7B, green highlights; Figure S4.8). It is likely that these ectopic enhancer interactions drive the overexpression of *TWIST1* in the NPCs of the patient. Thus, chromosome conformation capture data suggest that *TWIST1* has lost interactions with mesodermal enhancers and has gained new interactions with enhancers that are active in neurons, which may explain deregulation of *TWIST1* expression in the patient. The resemblance with phenotypes of patients with *TWIST1* mutations, deletions, and translocations strongly suggests a causative role of the *TWIST1* deregulation in the development of the phenotype of our patient. This important molecular phenotype with a likely impact on the phenotype of the patient is only detectable in the patient iPSC-derived NPCs.

4.4 Discussion

We determined the molecular effects of complex chromosomal rearrangements by transcriptome analyses on blood cells, iPSCs, and iPSC-derived neural progenitors from an MCA/MR patient with chromothripsis. In addition, we performed chromosome conformation capture analyses on the iPSC-derived neural progenitors to study the genomic architecture of the derivative chromosomes. We confirmed several previously identified direct effects of the breakpoints on gene expression, such as reduced expression of several hemizygously deleted genes and misexpression of fused (*DPYD-ETV1*) and truncated genes (*FOXP1* and *ETV1*) (van Heesch et al., 2014). In addition, some genes that are located near the breakpoints but are not directly affected by the breakpoints (*TWIST1* and *CNTN3*) were differentially regulated in the patient, indicating effects of the rearrangements on the regulatory DNA landscape. The altered expression of *TWIST1*, loss of genomic interactions with several of its enhancers, and the resemblance of the patient's phenotype with *TWIST1*$^{+/-}$ patients indicate that the *TWIST1* deregulation is a major cause of the patient's phenotype. The effect on *TWIST1* expression was not detectable in the blood cells of the patient, highlighting the importance of using disease-relevant cell types for the interpretation of the consequences of genomic rearrangements.

Although genomic rearrangements caused by chromothripsis are non-recurrent, the effects of complex rearrangements on the phenotype of a patient may be inferred from patients with similar phenotypes caused by less complex genomic rearrangements. In this study, especially the detected deregulation of *TWIST1* expression, which was only detected in the patient iPSC-derived NPCs, may explain a large part of the patient phenotype (the craniosynostosis and

the doubling of the thumbs). The coding sequence of *TWIST1* is not affected by the rearrangements, but translocations near *TWIST1* have been found before in patients with similar phenotypes (Cai et al., 2003; Krebs et al., 1997; Rose et al., 1997). Effects on *TWIST1* expression would have been difficult to predict by only studying the genomic variation of the patient, which demonstrates the importance of transcriptome analysis by RNA-seq to detect such effects in disease-relevant cell types. 4C-seq analyses showed that *TWIST1* gained and lost interactions with several enhancers, which could have led to the deregulation of the normal gene expression in different cell types. This example of *TWIST1* misexpression due to positional effects highlights the importance of not focusing solely on copy number changes or truncated and fused genes when studying the effects of chromosomal rearrangements (Spielmann and Mundlos, 2013). This is further underscored by our finding that only half of the deleted genes in this patient show a consistent reduced expression, suggesting dosage compensation at the RNA level for the other half of the deleted genes. With our approach, we narrowed down a list of 67 candidate genes within 1 Mb of the breakpoints to a list of three genes that likely contribute to the patient's phenotype.

Only a minority of the *TWIST1*$^{+/-}$ patients show signs of developmental delay and intellectual disability like those observed for the patient described in this study. It is very well possible that a combination of molecular effects led to the complex phenotype of the patient. For example, the disrupted *FOXP1* and *DPYD* genes are known MCA/MR genes that may have contributed to the intellectual disability and developmental delay in our patient. We cannot exclude that there are additional molecular effects in other cell types that also have contributed to the phenotype.

4.5 Conclusions

By analyzing the transcriptomes of blood cells, iPSCs, and iPSC-derived neuronal cells of a chromothripsis patient and both parents we identified the functional effects of the rearrangements that likely have contributed to the patient's phenotype. In particular we observed a cell type-specific effect of the rearrangements on the expression of *TWIST1*, even though the coding sequence of this gene was not disrupted by the rearrangements. This study shows the power of transcriptome and chromosome conformation capture analyses to detect effects of structural rearrangements on both coding sequences and regulatory elements. We identified clinically relevant molecular effects specific to the iPSC-derived neuronal cells. These findings underscore the importance of using disease-relevant cell types to better understand the molecular effects of chromosomal rearrangements.

4.6 Supplements

4.6.1 Abbreviations

FBS: Fetal bovine serum; hESC: Human embryonic stem cell; IL: Interleukin; iPSC: Induced pluripotent stem cell; Mb: megabase; MCA/MR: Multiple congenital abnormalities and/or mental retardation; NPC: Neural progenitor cell; PBMC: Peripheral blood mononuclear cell; PBS: Phosphate-buffered saline; RT: Room temperature; TAD: Topologically associated domain; TPO: Thrombopoietin.

4.6.2 Acknowledgements

We are grateful to Anko de Graaff from the Hubrecht Imaging Center and Livio Kleij for their support with imaging. We thank Carlo Vermeulen and Geert Geeven from the Hubrecht Institute for their help with the 4C-seq experiments. We are also grateful to Elzo de Wit from the Netherlands Cancer Institute (NKI) for assistance with Hi-C data analysis. We would also like to thank the Utrecht Sequencing Facility (USF) for sequencing.

4.6.3 Funding

This work was financially supported by a Vici grant (865.13.004) from the Netherlands Science Foundation (NWO) to EC.

4.6.4 Availability of data and materials

The RNA-seq, Hi-C, and 4C-seq datasets supporting the conclusions of this article have been deposited in the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) under the accession number EGAS00001001896.

4.6.5 Authors' contributions

EWK, SvH, WPK, and EC designed the study. KB and EK generated and cultured the iPSC lines. EWK and MK performed iPSC culturing, in vitro differentiation, and immunofluorescence stainings. EWK performed confocal microscopy, RNA extractions, preparation of RNA-seq libraries, immunofluorescent staining and analysis of brain sections, and cloning of overexpression constructs. SM, JdL, MvR, MS, and MvI analyzed sequencing data. SvH performed Hi-C and SM performed 4C-seq. RH performed karyotyping. NV and EI provided patient information. YA and JP performed in utero electroporations. SM, EWK, and EC wrote the manuscript. All authors contributed to the final version of the manuscript. All authors read and approved the final manuscript.

4.6.7 Competing interests

MS is an employee of Cergentis. The remaining authors declare that they have no competing interests.

4.6.8 Consent for publication

Not applicable.

4.6.9 Ethics approval and consent to participate

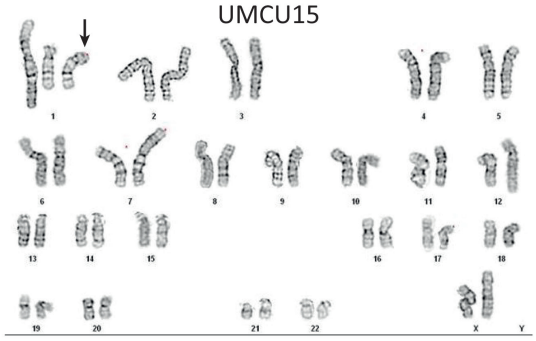
Written informed consent to participate in this study was obtained from the parents of the pediatric patient. Genetic analyses were performed according to the guidelines of the Medical Ethics Committee of the University Medical Center Utrecht. The study was performed in accordance with the Declaration of Helsinki. Animal use and care were in accordance with institutional and national guidelines and approved by the Animal Ethic Committee (DEC) of the Utrecht University.

4.6.10 Supplemental figures

A

Cell line		Number of metaphases investigated	Banding quality		Karyotype
			Best	Average	
UMCU14	Patient	20	500	300	46,XX,der(1)(7pter→7p21.3::7p21.3→7p21.2::1p21.3→1qter), der(3)(3pter→3p13::3p12.3→3p13::7p21.1→7p21.1::1p21.3→1p21.3::3p12.3→3qter), der(7)(12qter→12q14.2::7p21.1→7p21.1::7p21.1→7qter), der(12)(12pter→12q14.2::7p21.2→7p21.1::1p21.3→1pter)
UMCU15	Patient	20	500	400	47,XX,der(1)(7pter→7p21.3::7p21.3→7p21.2::1p21.3→1qter)x2, der(3)(3pter→3p13::3p12.3→3p13::7p21.1→7p21.1::1p21.3→1p21.3::3p12.3→3qter), der(7)(12qter→12q14.2::7p21.1→7p21.1::7p21.1→7qter), der(12)(12pter→12q14.2::7p21.2→7p21.1::1p21.3→1pter)
UMCU16	Patient	20	500	400	46,XX,der(1)(7pter→7p21.3::7p21.3→7p21.2::1p21.3→1qter), der(3)(3pter→3p13::3p12.3→3p13::7p21.1→7p21.1::1p21.3→1p21.3::3p12.3→3qter), der(7)(12qter→12q14.2::7p21.1→7p21.1::7p21.1→7qter), der(12)(12pter→12q14.2::7p21.2→7p21.1::1p21.3→1pter)
UMCU23	Father	20	400	400	46,XY
UMCU32	Father	20	400	400	46,XY
UMCU30	Mother	21	400	400	46,XX,der(20)t(1;20)(q25;p11.2 or q11.2)

B



C

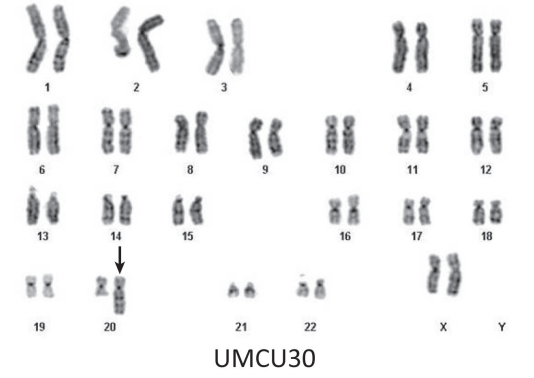
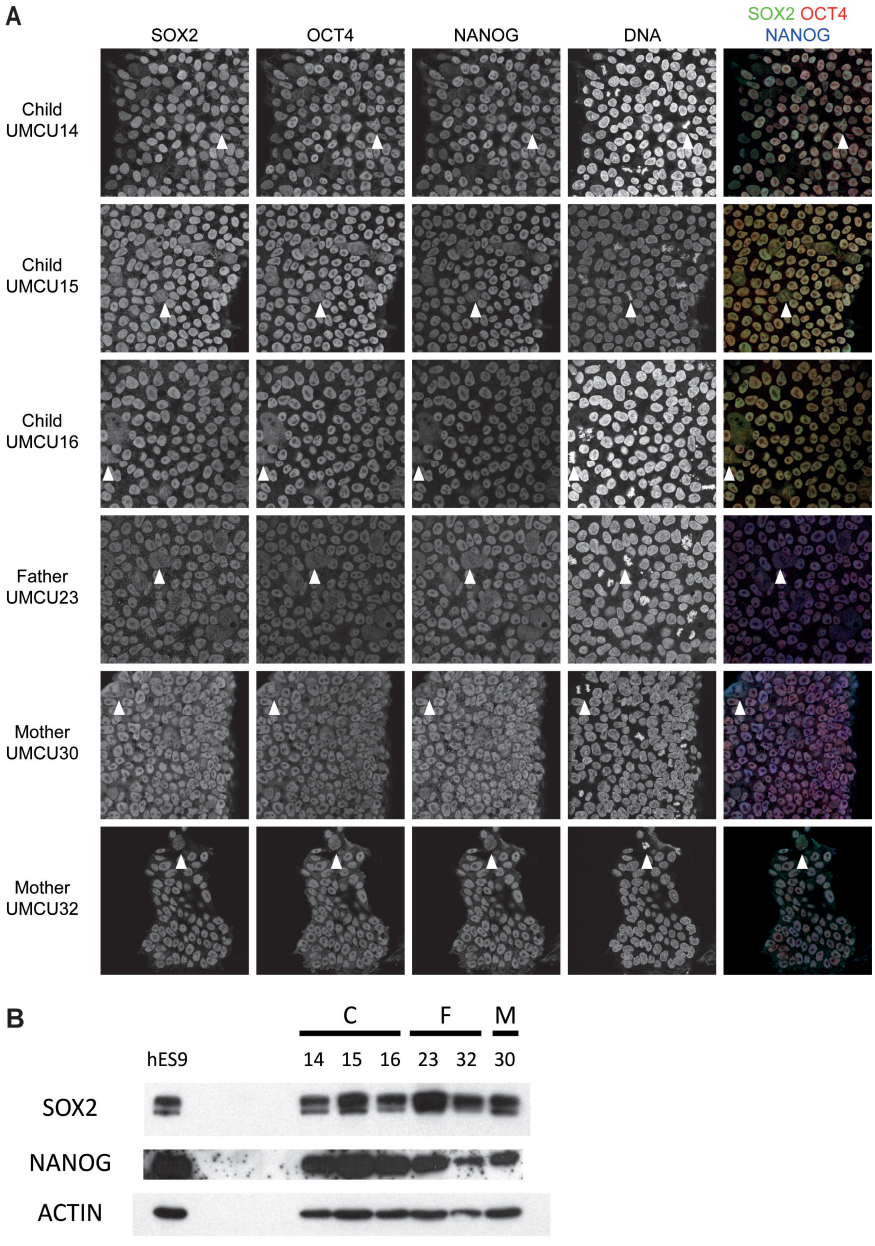
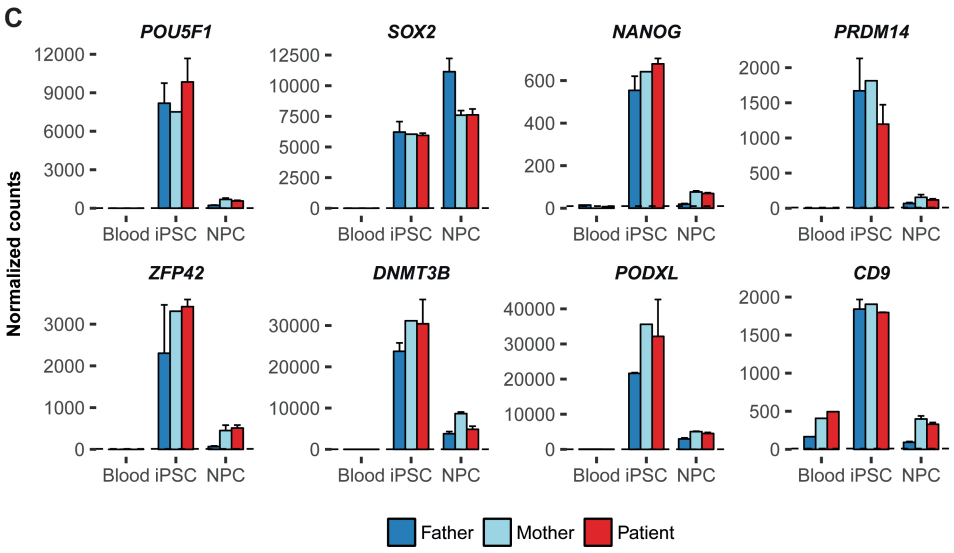


Figure S4.1 | Karyotypes of the patient's and parent's iPSC lines. (A) *Overview of karyotyping results. Most iPSC lines contained the expected karyotypes. One iPSC line derived from the patient and one derived from the mother acquired an additional genomic rearrangement during cultivation.* ***(B)*** *Karyogram of one of the patient iPSC lines (UMCU15) containing a duplication of derivative chromosome 1.* ***(C)*** *Karyogram of the iPSC line derived from the mother (UMCU30) showing a translocation between chromosome 20 and a fragment of chromosome 1. This rearrangement is located more than 70 Mb away from the locations of the rearrangements in the patient.*

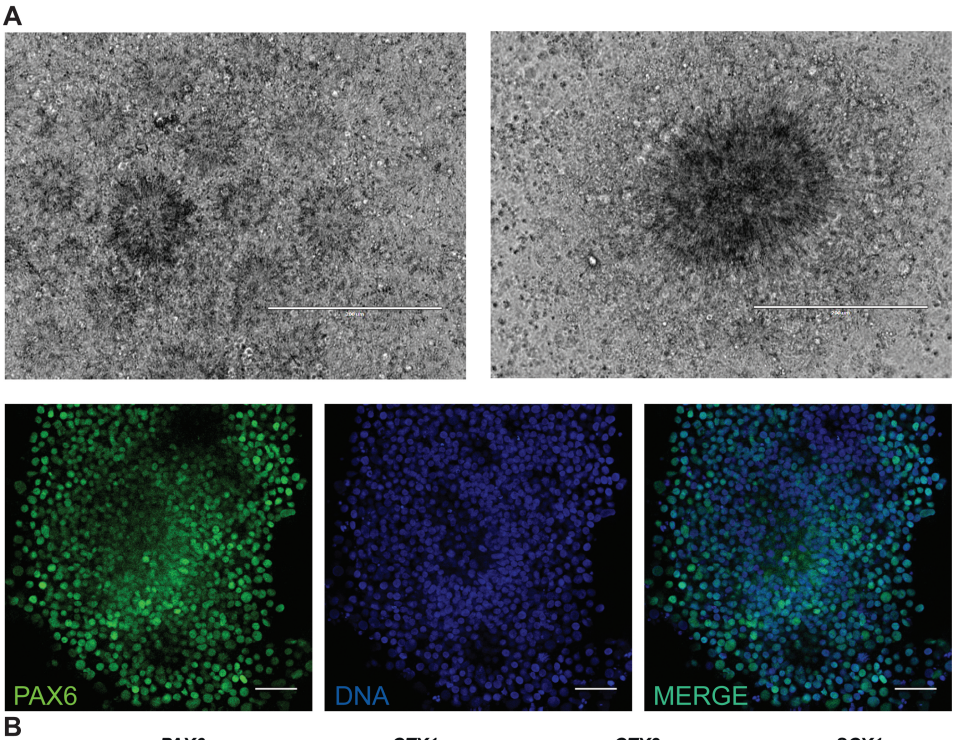

A
SOX2
OCT4
NANOG
DNA
SOX2 OCT4
NANOG
Child
UMCU14
Child
UMCU15
Child
UMCU16
Father
UMCU23
Mother
UMCU30
Mother
UMCU32
B
C
F
M
hES9
14 15 16 23 32 30
SOX2
NANOG
ACTIN

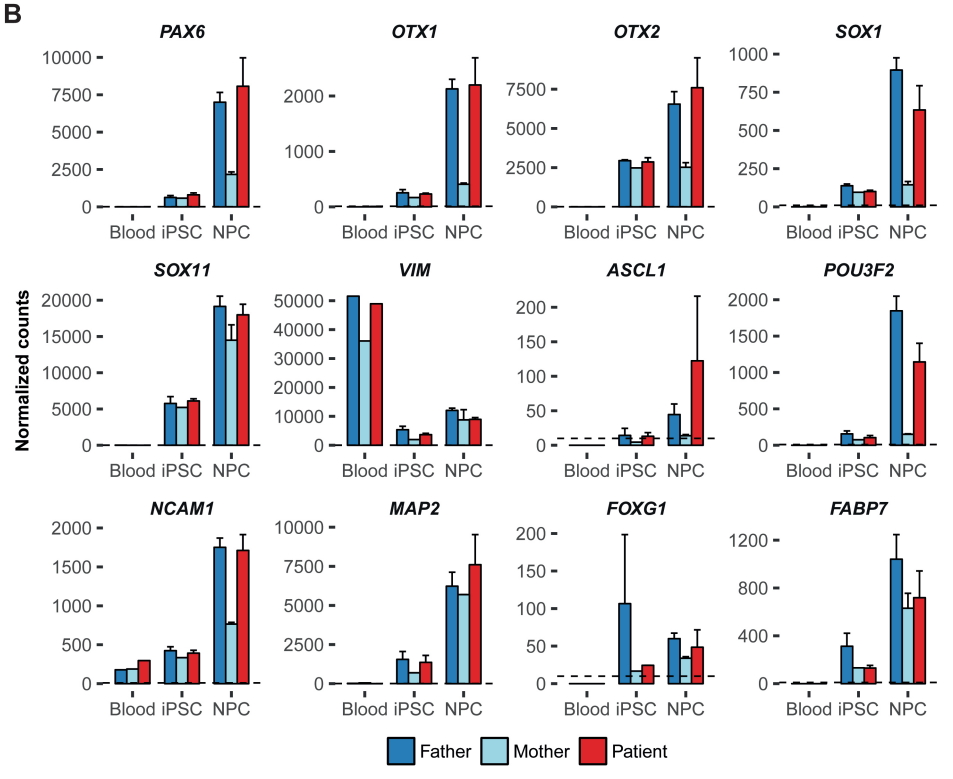
>>>



<<<

Figure S4.2 | Patient's and parental iPSC lines express high levels of pluripotency markers. ***(A)*** *Immunostainings showing the expression of the pluripotency markers SOX2, OCT4 and NANOG (column 1 to 3). DNA is stained with DAPI (fourth column).* ***(B)*** *Western blot showing expression of SOX2, NANOG and ACTIN in hES9 cells (positive control) and the patient/child (14, 15, 16), father (23 and 32) and mother (30). The iPSC lines show expression levels of SOX2 and NANOG comparable to the hES9 cell line.* ***(C)*** *Bar graphs showing the normalized RNA expression of eight genes associated with pluripotency. Patient's RNA expression data was generated for cell line UMCU14 and UMCU15. SOX2 is a marker for both pluripotent stem cells and early neural cells. The dashed horizontal line indicates the expression threshold of 10 normalized read counts. Error bars indicate standard error.*


A
PAX6
DNA
MERGE


B
PAX6
OTX1
OTX2
SOX1
SOX11
VIM
ASCL1
POU3F2
NCAM1
MAP2
FOXG1
FABP7
Normalized counts
Blood iPSC NPC
Father
Mother
Patient

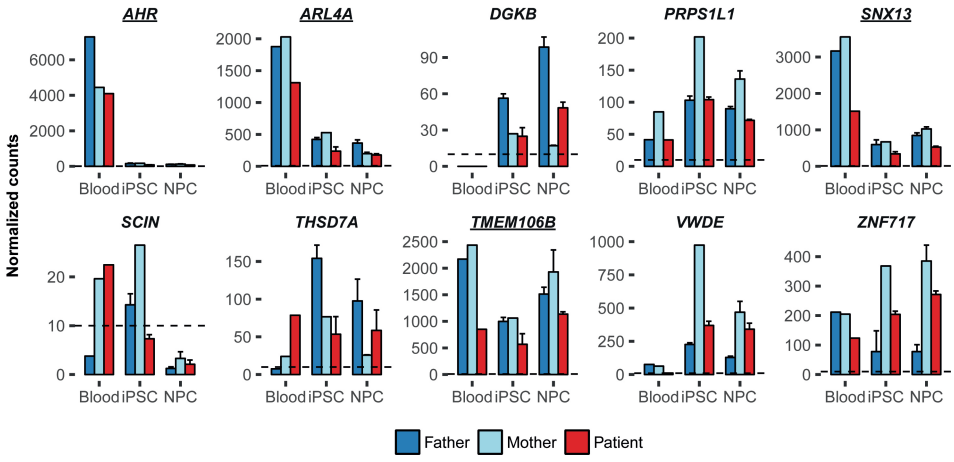


Figure S4.4 | Decreased RNA expression of four of the ten deleted genes in the patient. *Bar graphs showing the normalized RNA expression for the ten deleted genes located on four deleted fragments in the patient. The four genes with underscored names show a decreased expression in all three cell types. The dashed horizontal line indicates the expression threshold of 10 normalized read counts. Error bars indicate standard error.*

<<<

Figure S4.3 | iPSC-derived NPCs form neural rosettes and express high levels of early neural markers. (A) *Examples of brightfield (top) and immunofluorescent images (bottom) of neural rosettes formed by iPSC-derived NPCs 15 days after the start of differentiation. PAX6 is a marker for NPCs (bottom, left). DNA is stained with DAPI (bottom, center). The scale bars in the upper images indicate 200 µm.* ***(B)*** *Bar graphs showing the normalized expression of eight genes associated with early neural cells in patient's and parental blood cells, iPSCs and iPSC-derived NPCs. The dashed horizontal line indicates the expression threshold of 10 normalized read counts. Error bars indicate standard error.*

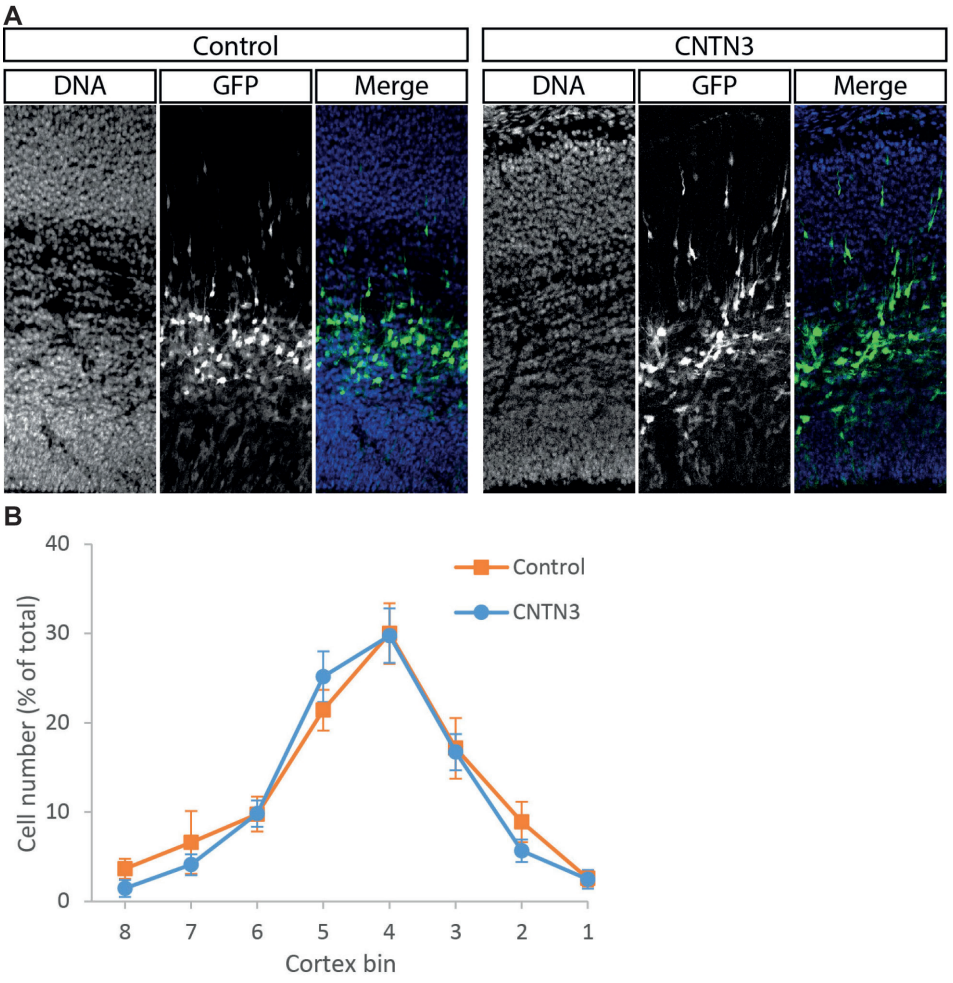


Figure S4.5 | Migration of E16.5 mouse embryonic cortical neurons not affected by CNTN3 overexpression. ***(A)*** *Immunofluorescent stainings of brain sections of a E16.5 mouse embryo treated with control (left) or pCAG-CNTN3 vectors (right) on day E14.5. A pCAG-GFP vector was co-injected with the pCAG-CNTN3 or control constructs to identify successfully targeted cells. DNA was stained with Hoechst.* ***(B)*** *Quantification of the number of GFP-positive cells in equally sized bins covering the cortical layers from the ventricle border (bin 1) to the pial surface (bin 8).*

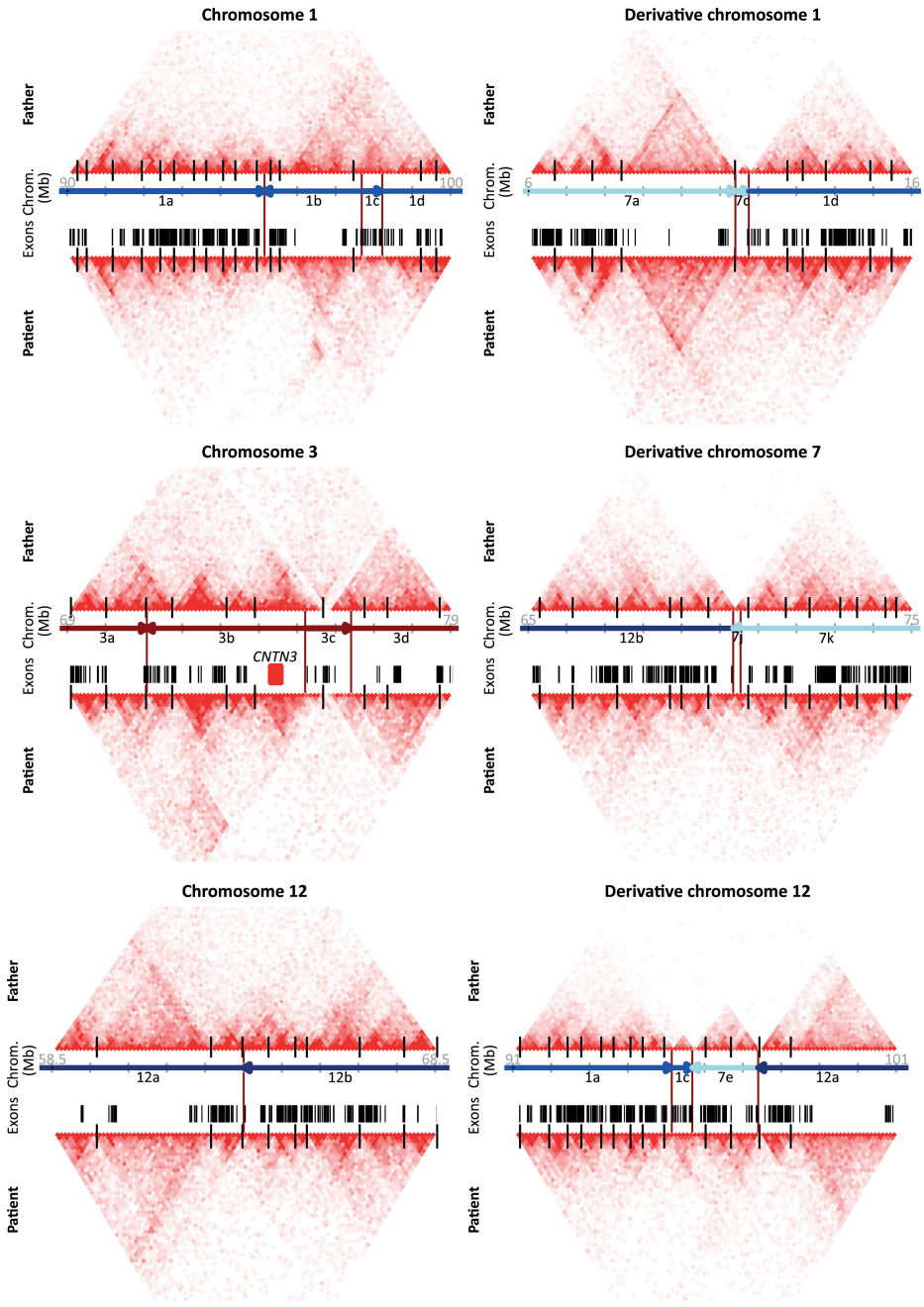


Figure S4.6 | Changes in genomic interactions on derivative chromosomes of the patient. *Hi-C chromatin interactions maps of the patient's and father's chromosomes (left panels) and derivative chromosomes (right panels). Interactions are shown at a 100 kb resolution. The vertical black lines at the bases of the heatmaps depict the predicted TAD boundaries in hESCs as determined by (Dixon et al., 2012). Vertical red lines between the interaction maps indicate the breakpoints locations in the patient.*

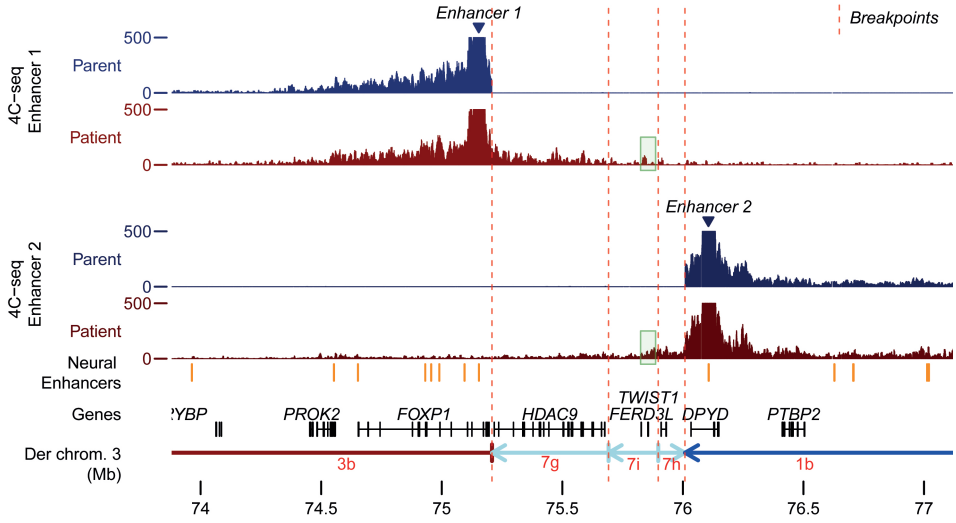


Figure S4.7 | Enhancers active in NPCs gained ectopic interactions with TWIST1 (highlighted in green) in the patient. *Two enhancers active in NPCs (determined by ChromHMM analysis of Roadmap ChIP-seq data of hESC-derived NPCs) were used as bait for 4C-seq. These aberrant interactions with neural enhancers may be the cause of TWIST1 overexpression in the patient's NPCs. The y-axes indicate the number of normalized 4C-seq reads cut-off at 500 reads.*

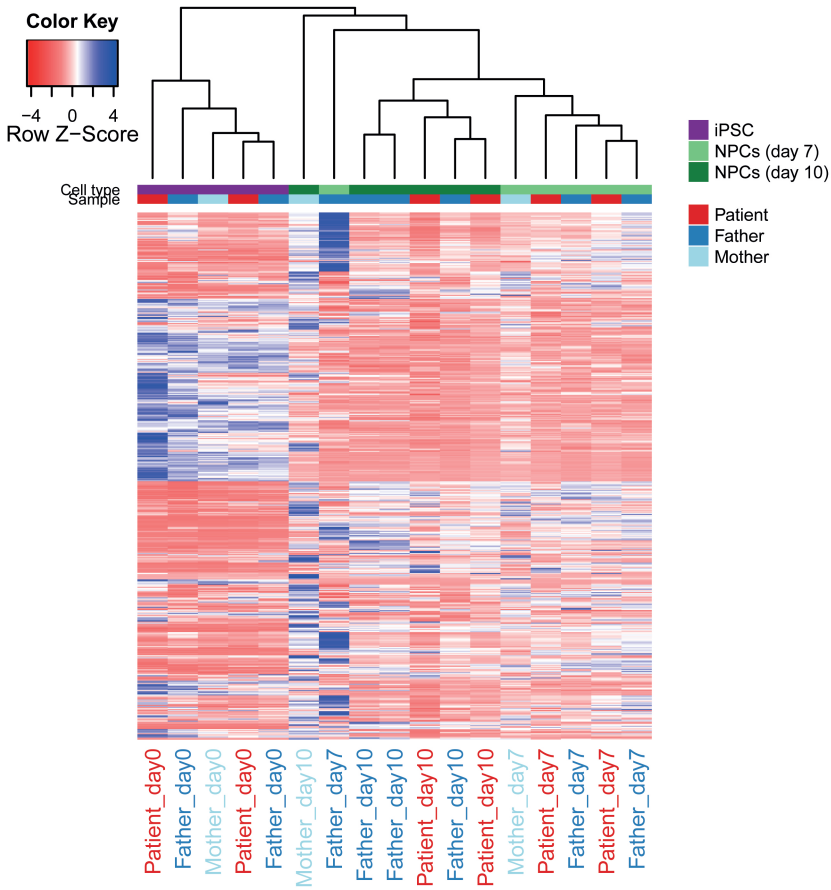


Figure S4.8 | RNA expression profiles of day 7 and day 10 NPCs cluster together. *Heatmap showing the results of an Euclidean hierarchical clustering analysis of the 500 genes with highest variance between the 17 samples.*

Chapter 5

Biallelic variants in *POLR3GL* cause endosteal hyperostosis and oligodontia

Paulien A. Terhal*, Judith M. Vlaar*, Sjors Middelkamp*, Rutger A. J. Nievelstein, Peter G. J. Nikkels, Jamila Ross, Marijn Créton, Jeroen W. Bos, Elsbeth S. M. Voskuil-Kerkhof, Edwin Cuppen, Nine Knoers, Koen L.I. van Gassen**

** Equal contribution*

*** Corresponding author*

Adapted from:

European Journal of Human Genetics (2019)

Abstract

RNA polymerase III (Pol III) is an essential 17-subunit complex responsible for the transcription of small housekeeping RNAs such as transfer RNAs and 5S ribosomal RNA. Biallelic variants in four genes (*POLR3A*, *POLR3B*, and *POLR1C* and *POLR3K*) encoding Pol III subunits have previously been found in individuals with (neuro-) developmental disorders. In this report, we describe three individuals with biallelic variants in *POLR3GL*, a gene encoding a Pol III subunit that has not been associated with disease before. Using whole exome sequencing in a monozygotic twin and an unrelated individual, we detected homozygous and compound heterozygous *POLR3GL* splice acceptor site variants. RNA sequencing confirmed the loss of full-length *POLR3GL* RNA transcripts in blood samples of the individuals. The phenotypes of the described individuals are mainly characterized by axial endosteal hyperostosis, oligodontia, short stature, and mild facial dysmorphisms. These features largely fit within the spectrum of phenotypes caused by previously described biallelic variants in *POLR3A*, *POLR3B*, *POLR1C*, and *POLR3K*. These findings further expand the spectrum of POLR3-related disorders and implicate that *POLR3GL* should be included in genetic testing if such disorders are suspected.

5.1 Introduction

RNA polymerase III (Pol III) is one of the three polymerase complexes involved in eukaryotic RNA synthesis. RNAs synthesized by Pol III include the small transfer RNAs (tRNA), 5S ribosomal RNA (rRNA), U6 small nuclear RNA, U5 rRNA, and several other non-coding RNAs (Arimbasseri and Maraia, 2016; White, 2011). Pol III is a highly conserved essential enzyme complex consisting of 17 subunits (Figure 5.1A). Twelve of these subunits are shared with RNA polymerase I and/or RNA polymerase II (Vannini and Cramer, 2012). Germline variants in four genes encoding for Pol III subunits POLR3A, POLR3B, POLR1C, and POLR3K have been shown to cause a spectrum of phenotypes termed POLR3-related disorders, which were previously appointed as tremor-ataxia with central hypomyelinisation (TACH), leukodystrophy with oligodontia (LO, MIM: 607694), 4H (hypomyelination, hypodontia, hypogonadotropic hypogonadism) syndrome (MIM: 612440) or isolated hypogonadotropic hypogonadism (Bernard et al., 2012; Cayami et al., 2015; Daoud et al., 2013; Dorboz et al., 2018; Ghoumid et al., 2017; Gutierrez et al., 2015; Minnerop et al., 2017; Ozgen et al., 2005; La Piana et al., 2016; Potic et al., 2012; Richards et al., 2017; Saitsu et al., 2011; Shimojima et al., 2014; Takanashi et al., 2014; Terao et al., 2012; Tétreault et al., 2011; Thiffault et al., 2015; Wolf et al., 2014). Frequently occurring phenotypic features in individuals with POLR3-related disorders are white matter abnormalities, cerebellar signs, motor delay and/or regression, dental abnormalities, myopia, short stature, and hypogonadotropic hypogonadism (Wolf et al., 2014). Specific combinations of biallelic splicing and truncating variants in *POLR3A* can also lead to the distinct neonatal progeroid (Wiedemann-Rautenstrauch) syndrome (MIM: 264090) (Jay et al., 2016; Paolacci et al., 2018; Wambach et al., 2018).

The vertebrate Pol III complex contains either the widely expressed POLR3GL (also known as RPC7L or RPC32β) or its isoform POLR3G (also known as RPC32α) (Haurie et al., 2010; Renaud et al., 2014). The two paralogous genes likely originated from a gene duplication event in a common ancestor of vertebrates and their protein sequences share 46% amino acid identity (Renaud et al., 2014). POLR3GL is part of a detachable Pol III subcomplex important for transcription initiation together with POLR3C and POLR3F (Vannini and Cramer, 2012). Here we report biallelic variants in the *POLR3GL* gene in three individuals with syndromic forms of endosteal hyperostosis in combination with oligodontia.

5.2 Materials and methods

5.2.1 Whole exome sequencing

Written informed consent was obtained from all included individuals and all procedures were performed in accordance with the guidelines of the Medical Ethics Committee (METC) of the University Medical Center Utrecht. Research has been performed in

accordance with the Declaration of Helsinki. After referral for routine diagnostic whole exome sequencing (WES), exomes of individuals P1 and P3 and their parents were enriched using the Agilent SureSelect XT Human All Exon kit V5 and sequenced on the HiSeq2500 sequencing system (Illumina, San Diego, CA, USA) in rapid 2 × 100 bp run mode with a mean target depth of 100×. Reads were aligned to hg19 using BWA (BWA-MEM v0.7.5a) and variants were called using GATK haplotype caller (V2.7-2).

5.2.2 Variant filtering and reporting

Detected variants were annotated, filtered and prioritized using the Bench NGS Lab platform (Agilent-Cartagenia, Santa Clara, CA, USA). Only variants that fitted a de novo or recessive inheritance model were analyzed. Variants dominantly inherited from one of the parents were excluded from the analysis. Reporting of de novo variants in candidate genes (genes of uncertain clinical significance) was restricted to putative protein changing variants in genes that are intolerant to missense and loss-of-function variants (Lek et al., 2016). For the recessive inheritance hypothesis, homozygous and compound heterozygous putative protein changing variants were filtered using a population allele frequency cutoff of 0.5% (ExAC database (Lek et al., 2016)). Variants in candidate recessive genes were only reported if at least one allele carried a putative loss-of-function variant. Larger deletions/duplications, missense, synonymous, and intronic variants affecting protein function of other genes cannot be excluded. No putative protein changing de novo variants were identified in individuals P1 and P3. The only variants that fulfilled the stringent diagnostic filtering and reporting criteria were homozygous and compound heterozygous variants in the *POLR3GL* gene in P1 and P3. No putative causative variants were identified in the other POLR3-related genes. The presence of the *POLR3GL* variants was confirmed in all three individuals and their parents by Sanger sequencing on an ABI 3730 analyzer with BigDye chemistry V.3.1. Monozygosity between individuals P1 and P2 was confirmed by Short Tandem Repeat (STR) analysis. *POLR3GL* variants are annotated according to reference NC_000001.10 (NM_032305.1). *POLR3GL* exons are numbered according to ENST00000369314.1 (GRCh37. p13, exon 2 is named ENSE00003603498 and exon 5 is named ENSE00003501352).

5.2.3 RNA sequencing

Peripheral blood mononuclear cells (PBMCs) were isolated from fresh blood samples using Ficoll-Paque PLUS (GE Healthcare Life Sciences, Cleveland, Ohio) and SepMate tubes (STEMCELL Technologies, Köln, Germany) according to the manufacturer's protocols. Total RNA was isolated from the PBMCs using the QIAsymphony (Qiagen, Venlo, The Netherlands) or the RNeasy Kit (Qiagen). RNA-sequencing (RNA-seq) libraries were prepared using TruSeq Stranded Total RNA Library Prep Kit (Illumina) according to the manufacturer's protocol. RNA-seq libraries were pooled and

sequenced on a NextSeq500 (Illumina) in 2 × 75 bp paired-end mode. RNA sequencing data analysis was performed using a custom in-house pipeline (https://github.com/UMCUGenetics/RNASeq). Briefly, the reads were mapped against the human reference genome (CRCh37/hg19) using STAR (Dobin et al., 2013). Mapped reads were quantified using HTseq-count (Anders et al., 2015) and read counts were normalized using the R-package DESeq2 (Love et al., 2014). DESeq2 was also used to perform differential gene expression analysis. Differential *POLR3GL* (ENST00000369314.1, CRCh37) exon expression analysis of individual P2, father of individuals P1/P2, mother of individuals P1/P2, individual P3 and 20 control subjectsfrom an in-house database was performed using the R-package DEXSeq (Anders et al., 2012; Reyes et al., 2013).

5.2.4 PCR validation of exon skipping

cDNA was synthesized with OligoDT and SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, California) following manufacturer's protocol. PCR reactions were prepared in 20 µl volumes containing 2× Phusion High-Fidelity master mix (New England Biolabs, Ipswich, MA, USA), 3% DMSO, 0.5 µM forward primer, 0.5 µM reverse primer and 10% template cDNA and were performed using the following PCR conditions: one step of 98 °C for 30 s; 16 thermal cycles (including touchdown steps of –0.5 °C each annealing step) of: 98 °C for 10 s, 61 °C (–0.5 °C per cycle) for 30 s and 72 °C for 30 s; followed by 20 thermal cycles of: 98 °C for 10 s, 58 °C for 30 s, and 72 °C for 30 s; followed by one step of 72 °C for 10 min. PCR products were analyzed on 1% agarose gel with 1:15000 SYBR safe DNA gel stain (ThermoFisher Scientific, Waltham, MA, USA). The following primer sequences have been used (Figure S5B): Pr1: GCCCAGTACATTTCAAGTTGG, Pr2: GCAGCA GGTTTATTCACTGG, Pr5: TGTAATCCGTTCCTTCTG TAGC, Pr6: CCTTATCTTCTGTGGTCTTAGGG.

5.3 Results

5.3.1 Detection of biallelic *POLR3GL* variants by WES

Diagnostic trio-based WES identified biallelic *POLR3GL* variants in a female of 19 years old (individual P1) and one unrelated female of 36 years old (individual P3). Individual P1 has a monozygotic twin sister (individual P2, Figure 5.1B) and the presence of the same *POLR3GL* variants in her sister was validated by Sanger sequencing. The siblings and the unrelated individual show syndromic forms of endosteal hyperostosis in combination with oligodontia, but they were only matched after the identification of the *POLR3GL* variants. Individuals P1 and P2 have a homozygous c.326-1G>A p.? splice acceptor site variant upstream of exon 5 of *POLR3GL*. They inherited these variants from their healthy parents who are heterozygous carriers of this variant (Figure 5.1B, C). Genealogical analysis indicates that the father and mother have a shared ancestor seven generations ago and are therefore distantly related (Figure 5.1B). Individual P3

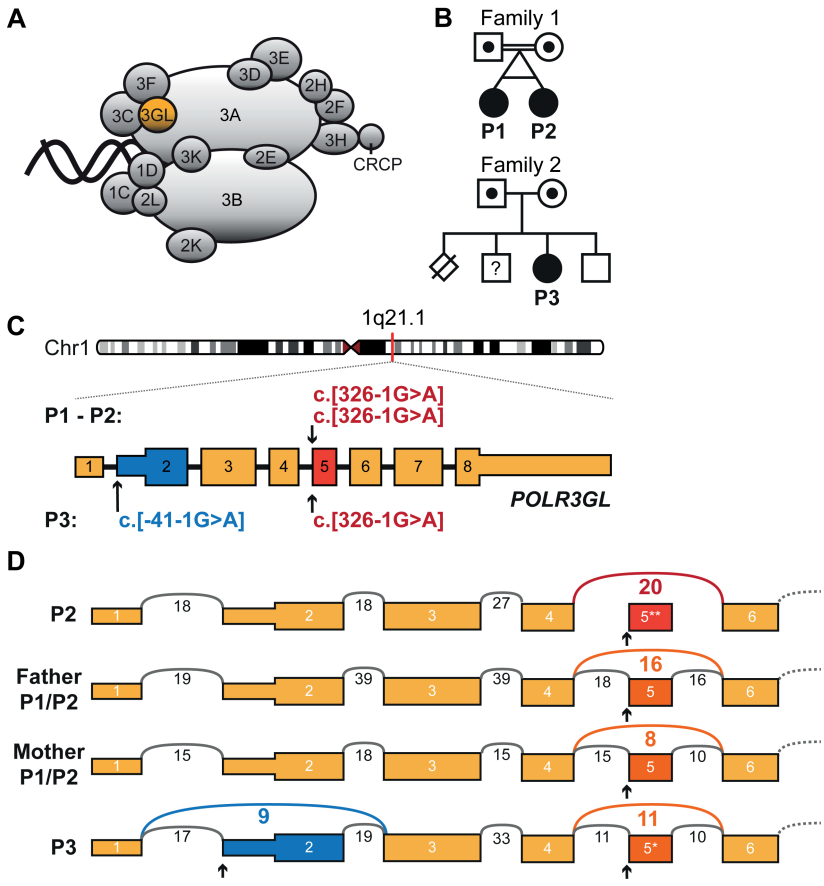


Figure 5.1 | Biallelic POLR3GL splice site variants in three individuals with endosteal hyperostosis and oligodontia. (A) *Schematic representation of Pol-III protein complex with the highlighted subunit POLR3GL which forms a trimer subcomplex with POLR3C and POLR3F. Adapted from (Flores et al., 1999).* ***(B)*** *Pedigrees of the two families included in this study. Filled shapes denote affected individuals and dotted shapes denote heterozygous carriers. The affected monozygotic twin of family 1 inherited the POLR3GL c.326-1G > A homozygous variant from their healthy, distantly related carrier parents. Individual P3 from family 2 inherited a heterozygous c.-41-1G > A variant from her mother and a c.326-1G > A variant from her father. She has one healthy brother and one possibly affected brother who was not available for genetic counseling (square with question mark). One perinatal death of a sib due to a neural tube defect was reported. Nomenclature of variants is according to HGVS nomenclature.* ***(C)*** *Schematic representation of the POLR3GL coding sequence (ENSG00000121851, mRNA NM_032305.1, GRCh37/hg19) showing the positions of the homozygous splice site variants in individuals P1 and P2 and the compound heterozygous variants in individual P3. Variant c.326-1G > A (red) is predicted to cause an skip of exon 5 (ENSE00003501352) in the transcript and variant c.-41-1G > A (blue) may cause aberrant splicing of exon 2 (ENSE00003603498) containing the POLR3GL translation initiation site.* ***(D)*** *Schematic sashimi plot showing the number of RNA-seq reads crossing the junctions of POLR3GL exons 1 to 6 in blood cells of individual P2, father and mother of individuals P1/P2 and individual P3. The reads that skip either exon 2 or exon 5 are displayed in color. The arrows denote the location of the splice site variants. *p < 0.05, **p < 0.01 Benjamini-Hochberg adjusted p-values for differential exon usage of the exons of all transcripts of POLR3GL calculated with the R-package DEXseq.*

carries compound heterozygous variants (c.[-41-1G > A];[326-1G > A] p.[?]; [?]) in the splice acceptor sites of *POLR3GL* exons 2 and 5 (Figure 5.1C). The variant c.326-1G > A p.?, which was also found in P1 and P2, was inherited from her father and the variant upstream of exon 2 c.-41-1G > A p.? (rs782661984) was inherited from her mother (Figure 5.1B,C). The genealogical study did not find indications for a relationship between the two families. Both *POLR3GL* variants are reported in the GnomAD database (v2.1), but only in a heterozygous state at very low allele frequencies (c.-41-1G > A: 0.00002475 and c.326-1G > A: 0.000007955) (Lek et al., 2016). Variants were submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/).

5.3.2 Individuals with biallelic *POLR3GL* variants show endosteal hyperostosis and oligodontia

The three affected individuals with *POLR3GL* variants show overlapping phenotypes (Table 5.1, Supplementary Case Reports). Individuals P1 and P2 are monozygotic twin sisters who were born by vacuum extraction because of umbilical cord prolapse at 36 weeks. Individual P1 showed a delayed speech and motor development with hypotonia. She has a disharmonic IQ profile with an average verbal IQ and a low performance IQ (40 points below her verbal IQ). She is diagnosed with a pervasive developmental disorder and currently attends a school for physically handicapped children. Ophthalmological examination at the age of 18 years showed mild hypermetropia with good vision. She developed normal secondary sexual characteristics, but puberty was delayed and periods started at age 17 years. She is frequently using a wheelchair because of pains in her back and feet since the age of 16 years. Neurological investigations showed no cerebellar signs like ataxia or intention tremor, but there was hypotonia in arms and legs, with relatively low reflexes and proximal weakness of leg muscles. In addition, she has neurogenic bladder dysfunction with recurrent cystitis and an MRI scan of the lumbar spine showed bulging discs of L2-L3 and L3-L4 affecting the cauda equina.

Her sister, individual P2, has a non-progressive congenital spastic diplegia. She has severe motor problems and has never walked independently. Periventricular localized white matter abnormalities including a focal thinning of the corpus callosum were detected on an MRI scan of the brain at the age of 3.5 years (Figure S5.1). These abnormalities were consistent with brain lesions caused by perinatal asphyxia and there were no signs of a diffuse hypomyelination typically observed in individuals with POLR3-related leukodystrophy (Piana et al., 2014; La Piana et al., 2016; Wolf et al., 2014). Like her sister, she has a mean verbal IQ and a significant lower performance IQ. She attends a special school. Ophthalmological examination at the age of 18 years showed cerebral visual impairment in combination with mild hypermetropia. Puberty began at the age of 13 years and her periods started spontaneously at 16 years. Neurological examination at the age of 18 years revealed mild pseudobulbar dysarthria, bilateral spastic paresis on the right side more pronounced than on the left side and bilateral

Table 5.1 | Clinical features of individuals with biallelic POLR3GL variants

	P1	P2	P3
Age at follow-up	19 years	19 years	36 years
Gender	Female	Female	Female
POLR3GL variants	c.[326-1G > A]; [326-1G > A]	c.[326-1G > A]; [326-1G > A]	c.[-41-1G > A]; [326-1G > A]
Neurological features			
Intellectual disability	–, Mean verbal IQ, low performance IQ, PDD-NOS	–, Mean verbal IQ, low performance IQ	–, Mild learning problems
Motor retardation	+	+	+
Cerebellar signs (ataxia, dysmetria)	–	–	–
Microcephaly	–	+	–
Upper motor signs (pyramidal)	–	+, non-progressive spastic paraparesis	–
Seizures	–	–	–
Dysarthria	–	+, mild	–
Ophthalmological			
Myopia	–	–, astigmatism, cerebral visual impairment	–
Hearing			
Hearing loss	–	–	+, mainly conductive
Dental			
Oligodontia	+	+	+
Orthopedic			
Club feet	+	+	+, very mild
Growth impairment (<p5)	+	+	+
Endocrine			
Growth hormone deficiency	–	–	–
Delayed puberty	+	–	+
Radiological			
Endosteal sclerosis	+	+	+
MRI			
Cerebellar hypoplasia	U	–	U
Diffuse hypomyelination	U	– (at age 3.5 years)	U
Hypoplasia of corpus callosum	U	+, localized	U
Other			
Motor regression	+, frequent wheelchair use because of back pain	stable, wheelchair dependent, spasticity, hip problems	+, due to coxarthrosis
Dysmorphic facial features	+	+	+

Babinski reflexes. No ataxia or intention tremor was noted.

Both sisters were born with club feet and mild syndactyly of the second and third toe (Figure S5.2A–D). In addition they have oligodontia with only 8 (P1) and 14 teeth (P2) in their permanent dentition (Figure 5.2B,C,E,F). Radiological examination showed that both sisters have a mainly axial localized form of endosteal sclerosis (Figure 5.3A-D and Figure S5.3A–C). They also have a short stature (the height of P1 was 140 cm (–4,7 standard deviations (SD)) at age 18 years and the height of P2 was 136 cm (–4,59SD) at age 15 years), but endocrinological analyses of IGF-1, IGFBP3, free T4 and TSH levels at age of 18 years did not indicate growth hormone deficiency (Supplementary Case Reports). In addition, normal levels of bone metabolism markers were detected at the age of 18 years (Supplementary Case Reports). The twins have mild facial dysmorphisms including upslanting palpebral fissures, thin lips with downturned corners of the mouth, a flat philtrum and a beaked nose with relatively long columella (Figure 5.2A,D).

Individual P3 is a 36-year old female who was born after an uneventful pregnancy. She had a growth retardation in her childhood and she has a short stature (142.8 cm, –4.5SD at adulthood), but growth hormone levels measured at ages 7, 18, and 34 were

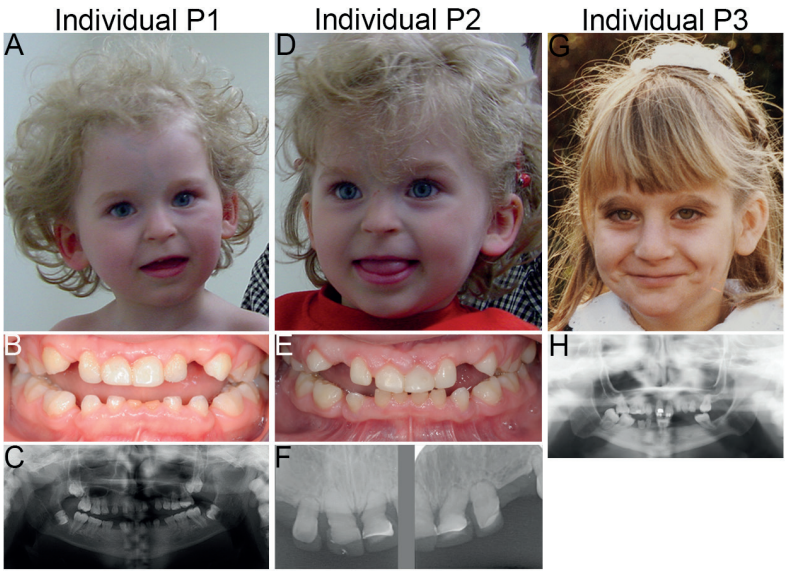


Figure 5.2 | Facial appearances and dental abnormalities of the three individuals carrying biallelic POLR3GL variants. (A) *Photograph showing the facial characteristics of individual P1 at the age of 3 years including upslanting palpebral fissures, thin lips, a flat philtrum and relatively long columella.* ***(B)*** *Intra-oral photograph of dental abnormalities of individual P1 showing an anterior open bite extending to the buccal teeth whereby only the most distal molars occlude. The central lower deciduous teeth are worn because of prolonged use.* ***(C)*** *Orthopantomographic radiograph of the full dentition of individual P1. The orthopantomograph shows that the permanent dentition consists of the first and second molars, all other teeth are congenitally absent. In addition, it shows that the upper- and lower molar pulp chamber have a taurodontic shape.* ***(D)*** *Photograph of the face of individual P2 at the age of 3 years showing upslanting palpebral fissures, a flat philtrum and thin lips with relatively long columella.* ***(E)*** *Intraoral photograph of dental abnormalities of individual P2. Only the distal molars occlude. There is spacing in between the teeth due to the growth of the jaws and absence of permanent teeth.* ***(F)*** *Radiograph of deciduous upper frontal teeth of individual P2 showing short erratic shape of the teeth and obliterated root canals. The original shape of the crowns and the covering with composite can be seen.* ***(G)*** *Photograph showing the facial characteristics of individual P3 at the age of 8 years which include a high nasal bridge with a relatively long columella and thin lips.* ***(H)*** *Orthopantomographic radiograph of the full dentition of individual P3. Upper medial incisors and 5 molars are present in the permanent dentition with taurodontic shape of the upper- and lower molars.*

normal (Supplementary Case Reports). She had a normal speech development and a slightly delayed motor development and attended regular primary and secondary school with some additional support. She had amblyopia of the left eye for which she wastreated with a patch. Ophthalmological examination at the age of 34 years revealed suboptimal vision with mild subcapsular posterior cataract and a bilateral refraction abnormality (Supplementary Case Reports). She had a delayed puberty with breast development starting around 15 years and a menarche at 17 years and 9 months, but she developed normal secondary sexual characteristics. She has oligodontia with only seven teeth in her permanent dentition (Figure 5.2H). Like the twins, she has several dysmorphic features including mild proptosis, a high nasal bridge with a relatively long

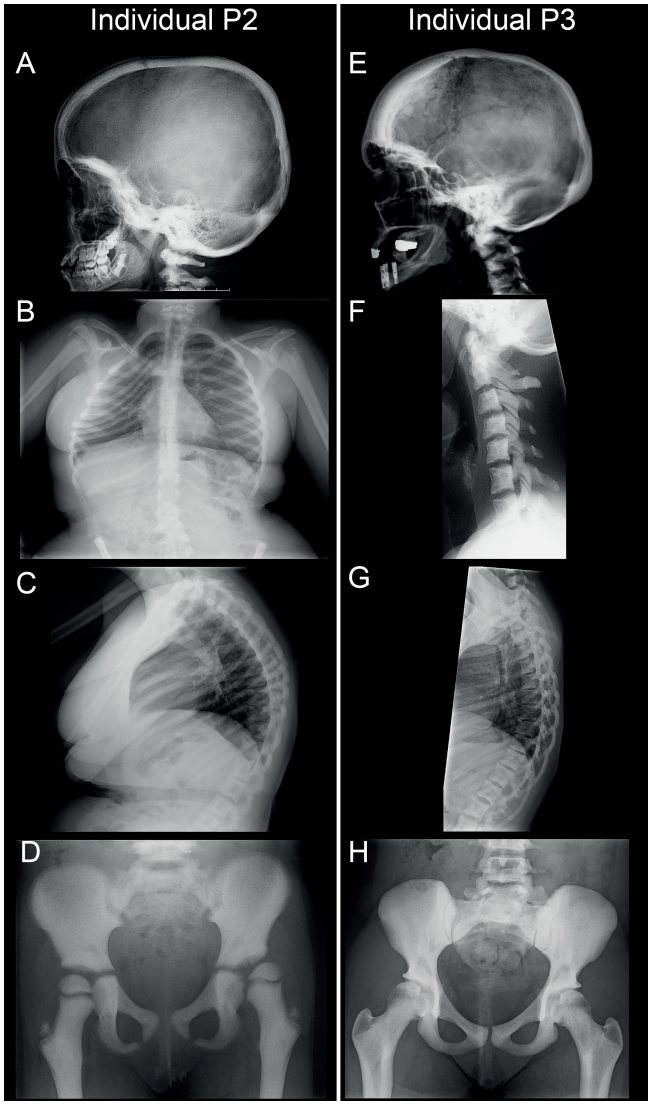


Figure 5.3 | Skeletal X-ray scans showing endosteal hyperostosis in individuals P2 and P3. (A-D) *X-rays of individual P2 at different ages.* ***(A)*** *X-ray at the age of 9 years showing sclerotic thickening of the neurocranium and the cranial base with sclerosis of C1 and C2.* ***(B)*** *Sclerotic ribs with scoliosis at the age of 18 years.* ***(C)*** *Sclerotic margins of the vertebrae with increased kyphosis of the thoracic spine at the age of 18 years.* ***(D)*** *X-ray of the pelvis at the age of 6 years showing diffuse hyperostosis, coxa valga and lateral displacement of the left femoral head.* ***(E-H)*** *X-rays of individual P3 at the age of 26 years.* ***(E)*** *Adult lateral skull showing sclerotic thickening, especially in the frontal and occipital area of the neurocranium.* ***(F)*** *Diffuse hyperostosis of the vertebral bodies and arches of the cervical spine.* ***(G)*** *Sclerotic margins of the vertebral bodies and arches of the thoracic spine.* ***(H)*** *Diffuse hyperostosis of the iliac and public bones and to a lesser extent of the proximal femora.*

columella, a broad nasal tip, downturned corners of the mouth and retrognathia (Figure 5.2G). She has syndactyly of the second and third toe and brachydactyly with short and stubby toes (Figure S5.2E,F). Skeletal surveys showed that she has a mainly axial localized form of endosteal hyperostosis like the other two individuals (Figure 5.4E-H, Figure S5.3D-I and Figure S5.4). Bone metabolism marker levels were normal at the age of 34 years (Supplementary Case Reports). She has moderate, mainly conductive and progressive hearing loss since the age of 18 years possibly due to tympanosclerosis. She is having increasing pain in the hips due to bilateral coxarthrosis since the age of 33 years and she recently underwent arthroplasty of the right hip at the age of 34 years. Neurological examination at the age of 34 years showed bilateral hypotonia and

proximal muscle weakness (MRC4) in the arms, but normal distal muscle strength. No signs of cerebellar involvement like ataxia, nystagmus or coordination difficulties were present at that age.

Thus, the three individuals with biallelic *POLR3GL* variants show overlapping phenotypic features, including axial endosteal hyperostosis, oligodontia, short stature and mild facial dysmorphisms.

5.3.3 Splice site variants in *POLR3GL* cause exon skipping

The detected biallelic *POLR3GL* variants are predicted to disrupt splice acceptor sites which could cause aberrant splicing of *POLR3GL* exon 5 and/or exon 2. We performed RNA-seq on blood samples of individual P2, the parents of individuals P1/P2 and individual P3 to determine the effects of the variants on splicing of *POLR3GL* RNA transcripts. Exon 5 is skipped in the *POLR3GL* RNA transcripts of individual P2 containing homozygous c. [326-1G > A]; [326-1G > A] variants and this exon is only partially included in the heterozygous carrier parents (Figure 5.1D). The compound heterozygous *POLR3GL* splice variants c. [-41-1G > A] and c. [326-1G > A] in individual P3 are predicted to cause either skipping of exon 2 or exon 5. Indeed, RNA-seq shows that both exons are only partially included in the *POLR3GL* transcripts of this individual (Figure 5.1D). PCR analysis confirms that this individual has no expression of full-length *POLR3GL* transcripts containing both exon 2 and exon 5 (Figure S5.5). The abundance of *POLR3GL* transcripts is not significantly altered in the individuals and their parents compared to the expression levels in unaffected controls (Figure S5.6). There is no indication for partially compensating cryptic *POLR3GL* splice acceptor sites in the RNA-seq data. Overall, these results confirm that the splice acceptor variants disrupt *POLR3GL* RNA transcripts in the affected individuals.

5.4 Discussion

Here we describe the presence of biallelic *POLR3GL* splice site variants in three individuals with axial endosteal hyperostosis, oligodontia, short stature, and mild facial dysmorphisms. Biallelic variants in *POLR3A*, *POLR3B*, *POLR3K* and *POLR1C*, which encode other Pol III subunits, have previously been associated with a spectrum of phenotypes (Table 5.2, Table S5.1). Patients with a severe POLR3-related disorder show a progressive disease with leukodystrophy, motor regression, oligodontia, hypogonadotropic hypogonadism, myopia, and intellectual disability. In contrast, some individuals at the mild end of the spectrum present only with learning difficulties and a delay in motor development (Wolf et al., 2014) or with isolated hypogonadotropic hypogonadism (Richards et al., 2017). The oligodontia, short stature and delayed puberty in the individuals described here are present in more than half of the individuals with POLR3-related disorders (Table 5.2). Most, but not all (Minnerop et al., 2017; Richards et al., 2017), described individuals with biallelic variants in Pol

Table 5.2 | Summary of the main clinical features in individuals with POLR3-related disorders

	POLR3A+ POLR3B	***POLR1C***	***POLR3K***	***POLR3GL***
Number of individuals	n=147	n=8	n=2	n=3
Intellectual disability	13/42	6/8	2/2	0/3
Motor delay	58/116	7/8	2/2	3/3
Cerebellar signs (ataxia, dysmetria)	142/147	8/8	8/8	0/3
Endosteal hyperostosis	U	U	U	3/3
Myopia	87/121	3/8	1/1	0/3
Oligo- / hypodontia	86/131	3/8	1/2	3/3
Hypogonadotropic hypogonadism / Delayed puberty	53/69	0/2	1/2	2/3
Short stature	54/126	U	2/2	3/3
Diffuse white matter abnormalities / Hypomyelination	100/126	8/8	2/2	U

U = unknown
Not all phenotypic characteristics have been explicitly specified or examined for each of the individuals described in literature and this may lead to an under- or over appreciation of some phenotypic characteristics in the spectrum of POLR3-related disorders. The clinical presentation of cohorts of individuals with POLR3A and POLR3B variants are frequently discussed together in literature and therefore the data for these two genes are merged in the table. More detailed information can be found in Table S5.1. References: POLR3A+POLR3B: (La Piana et al., 2016; Richards et al., 2017; Saitsu et al., 2011; Tétreault et al., 2011; Wolf et al., 2014). POLR1C: (Thiffault et al., 2015). POLR3K: (Dorboz et al., 2018).

III genes have hypomyelination and cerebellar signs (Wolf et al., 2014). It is uncertain whether the individuals described here have developed white matter abnormalities during follow-up, because MRI scans of the brain have not been performed at later ages due to ethical and practical considerations. The three individuals do not suffer from a progressive neurological disorder with cerebellar, pyramidal or extrapyramidal signs. In addition, they do not have progressive myopia, which has been described for the majority of individuals with POLR3-related disorders. The endosteal hyperostosis in the individuals with *POLR3GL* variants is remarkable and has only been reported in two individuals with *POLR3B* variants (Ghoumid et al., 2017; Ozgen et al., 2005). However, skeletal X-rays are not always performed and the presence of endosteal hyperostosis in other individuals with POLR3-related disorders might therefore be underestimated (Ghoumid et al., 2017). We therefore recommend to perform skeletal surveys in all individuals with POLR3-related disorders. Taken together, the extraneurologic phenotypic features of the three individuals with *POLR3GL* variants fit in the spectrum of POLR3-related disorders, but the absence of progressive cerebellar, pyramidal or extrapyramidal features is remarkable.

It remains unclear how the newly identified *POLR3GL* variants lead to the observed phenotypes. The *POLR3GL* transcripts of the individuals described here lack exon 2, which contains the translation initiation site, or exon 5, which is part of the core domain essential for interacting with other Pol III subunits (Boissier et al., 2015). Nevertheless, it is likely that these variants do not lead to a full loss-of-function of POLR3GL due to the essential functions of the Pol III complex. In addition, biallelic nonsense or full loss of function variants of the other Pol III subunits have not been reported. The skip of exon 5 does not cause a frameshift in the *POLR3GL* RNA sequence and likely results in disrupted POLR3GL protein missing 19 of the 218 amino acids (p.(Asp109_Arg128delinsGly)). In contrast, loss of exon 2 causes the loss of the canonical translation initiation site and the next potential downstream translation initiation site is in exon 3 located at p.Met71. Therefore, the *POLR3GL* allele missing exon 2 is either not functional or misses the first 70 amino acids (p.(Met1_Ala70del)). In theory, the isoform POLR3G could compensate for the reduced POLR3GL function. It seems that POLR3G or POLR3GL largely bind to the same target genes, but that their activity is controlled by different mechanisms (Renaud et al., 2014). However, it has been shown that, at least in HeLa and Huh7 human cancer cell lines, POLR3G cannot fully compensate for the loss of POLR3GL (Haurie et al., 2010). Homozygous knockout variants of *Polr3c* and *Polr3f*, which form a subcomplex of Pol III together with POLR3GL, as well as homozygous knockout variants of *Polr3a* and *Polr3b* are lethal in mice (Koscielny et al., 2014). *POLR3GL* knockout animal models have not been described to our knowledge. These observations suggest that the variants are hypomorphic and that the affected POLR3GL retains some form of essential functioning.

The phenotypic overlap between individuals carrying variants affecting the functions of Pol III subunits suggests that a common function of this complex is affected. Deficiency of tRNAs normally generated by Pol III could be an important cause of the phenotypes (Arimbasseri and Maraia, 2016). However, it is unknown why such a deficiency would lead to the observed cell-type specific phenotypes. Neuronal tissues seem to be highly sensitive to pathogenic variants in tRNA genes such as *n-Tr20* in mice or genes coding for tRNA processing enzymes such as *CLP1* (Kirchner and Ignatova, 2014). Deficiencies of other Pol III transcribed RNAs such as U5 rRNA cannot be excluded, although ribosomopathies mainly lead to other phenotypes (Yelick and Trainor, 2015). The teeth and bone phenotypes in the individuals with biallelic variants in *POLR3GL* and other Pol III-subunit genes suggest that osteoblasts or osteoclasts are affected (Wolf et al., 2014), but it is unclear why these specific cell types would be sensitive to reduced Pol III activity.

In conclusion, biallelic splice site variants in *POLR3GL* can cause endosteal hyperostosis, oligodontia and short stature. These phenotypes fit within the spectrum of phenotypes observed in individuals carrying variants in other Pol III subunits. These findings show that it is important to include *POLR3GL* in genetic testing if a POLR3-

related disorder is suspected, especially if endosteal hyperostosis and oligodontia are observed.

5.5 Supplementary material

5.5.1 Acknowledgements

We thank the individuals and their parents for their participation in this study. We thank Utrecht Sequencing Facility (USEQ) for providing RNA-sequencing service and data. USEQ is subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University. We thank Jacques Giltay for his contribution to the inclusion of one of the families. We thank F.A.M. Hennekam for performing the genealogical studies in the families. This work was financially supported by a Vici grant (865.13.004) from the Netherlands Science Foundation (NWO) to Edwin Cuppen.

5.5.2 Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5.5.3 Case reports

Family 1, individual P1

Individual P1 (age 19) is the first born of a monozygotic female twin. Intrauterine growth retardation was present. She was born by vacuum extraction because of umbilical cord prolapse after 36 weeks and two days of gestation with a birth weight of 1555 gram. Apgar scores were 8 and 10 at respectively one and five minutes. After birth the presence of club feet was noted, which were treated with splints and plaster. In addition, she had an enlarged fontanel and an umbilical hernia for which she was operated on at the age of 5 years. There was a delayed development of speech and a delayed motor development with hypotonia. She could walk independently when she was two years old. She had difficulty learning to swim or ride a bicycle and reported balance problems. Club feet were initially treated with splints, but at the age of 5 years a bilateral Turco posteromedial release operation was performed. During the operation the left m. tibialis posterior and both m. peronei could not be identified. She has completed secondary school with support and obtained a degree. She has a disharmonic IQ, verbal IQ being average and performance IQ about 40 points below her verbal IQ. She has been diagnosed with a pervasive developmental disorder. At this moment she is attending a school for physically handicapped children. Eruption of primary dentition was on time, but teeth were small. However, eruption of permanent teeth was delayed and oligodontia was detected. She only has 8 teeth of her permanent

dentition (16, 17, 26, 27, 36, 37, 46, 47). Ophthalmological examination showed mild hypermetropia with good vision (right eye S plan=C-0.50 ax 90, left eye S +0.50=C-1.00 ax 90). She is using a wheelchair for long distances since the age of sixteen years because of pain in her feet and fatigue. At the age of seventeen years she also started to suffer from back pain and difficulty with walking and therefore she has been using the wheelchair more often since then, also for short distances. She experienced recurrent urinary tract infections from the first year after birth and became incontinent at the age of 17 years for which catheterization was started. Neurological examination showed gibbus deformity, hypotonia in arms and legs with relatively low reflexes and proximal weakness of leg muscles. Urological examination showed hyperreflexia and dyssynergic miction with small capacity of the bladder. Puberty was delayed, but she developed normal secondary sexual characteristics with somewhat small breasts. Her periods started spontaneously at age 17 years, after that time she started using a contraceptive pill which stimulated her breast development.

Both twin sisters had severe pneumonia when they were three years old and radiological examination at that time showed a mainly axial localized form of endosteal sclerosis. Bone densitometry showed an increased bone density which has stabilized during the years (at age 7 years hip 1,36 g/cm^2 Z-score 10,5 and lumbar spine 0.89 g/cm^2 Z-score 5,0). Physical examination at the age of 14 years showed a height of 140,4 cm (-3,94 standard deviations, SD), skull circumference of 52,5 cm (-1,43 SD), sitting height/height ratio 0.5 (-1.43 SD). She has facial dysmorphisms including upslanting palpebral fissures, thin lips with downturned corners of the mouth and a flat philtrum and a beaked nose with relatively long columella. In addition, she has bilateral camptodactyly of the fifth fingers, hyperextensible distal finger joints, relatively long middle finger compared to the index finger, relative large hallux, with a sandle gap, syndactyly of the second and third toe, mild varus deformity of the feet and prominent heels. Neurological examination showed hypotonia in arms and legs with relatively low reflexes. MRI of the lower spine at the age of 18 years showed lumbar kyphosis with bulging of all lumbar discs, narrowing of the spinal canal L2-L3 with pressure on the cauda equina. Endocrinological investigations showed no signs for growth hormone deficiency, hypothyroidism or abnormalities in calcium phosphate metabolism. At the age of 18 years IGF-1 was 36 nmol/L (mean), IGF-BP3 170 nmol/L (mean), TSH 3,8 mU/L (normal, ref. 0.4-4.3), T4 140 nmol/l (normal, ref. 60-150), Ca 2.50 mmol/L (normal, ref. 2.10-2.55), Phosphate 1.30 mmol/L (normal, ref. 0.90-1.50), Alkaline phosphatase 91 U/L (normal, ref. < 98).

Family 1, individual P2

Individual P2 (age 19) is the second born of the monozygotic twin sisters. She had a birth weight of 1395 gram. Apgar scores were 7 and 9 at one and five minutes respectively. After birth bilateral club feet were noted, for which she was treated with splints and

plaster. She soon developed non-progressive spastic paresis, the right side being more affected than left side, possibly due to prenatal asphyxia. She has severe motor problems; rolling over and sitting are still very difficult for her. She has pseudobulbar dysarthria, which is probably also related to the asphyxia. She has never walked independently and is wheelchair dependent. She underwent several hip operations because of hip luxation due to her spasticity, twice on the left side (derotational femoral osteotomy) and once on the right side (adductor tenotomia and derotational femoral osteotomy) and received multiple botulin injections. At the left side a stable hip subluxation is present. Similar to her twin sister she has a mean verbal IQ and a significant lower performance IQ. She is attending a school for physically handicapped children. Because of sleeping problems polysomnography was performed which showed no signs of hypoventilation or desaturations. A cerebral visual impairment in combination with mild hypermetropia (right eye S +1.00=C-2.50 ax 180, left eye S +1.00, C-4.00 ax 90) was diagnosed by ophthalmological examination. Eruption of the primary and permanent dentition was delayed and oligodontia is present in her permanent dentition (element 15-25 and 35-45 are missing).

Physical examination at the age of 14 years showed a skull circumference of 52 cm (-2.5-2 SD) and a relatively short trunk. She has similar facial dysmorphism as her sister including upslanting palpebral fissures, a flat philtrum, thin lips with a long nose, a relatively long columella and mild protruding ears. Bilateral club feet with syndactyly between the second and third toe with a long hallux were noted. Pubertal development started at age 13 and periods starting spontaneously at 16 years. Neurological examination at the age of 18 years revealed mild pseudobulbar dysarthria, bilateral spastic paresis, on the right side more pronounced than on the left side, and bilateral Babinski reflexes. MRI of the brain and thoracic and cervical spine at the age of 3.5 years showed subependymal white matter abnormalities, and white matter anomalies located in the posterior part of the capsula interna, possibly due to perinatal asphyxia. Like her twin sister, she was diagnosed with a mainly axial localized form of endosteal hyperostosis at the age of three years. Bone densitometry showed an increased bone density which, similar to her sister, has stabilized during the years (at age 7 years hip 1,19 g/cm^2 Z-score 8,3 and lumbar spine 0.81 g/cm^2 Z-score 3,8). Endocrinological investigations showed no signs for growth hormone deficiency, hypothyroidism or abnormalities in calcium phosphate metabolism. At the age of 18 years IGF-1 was 53 nmol/L (between +1 and +2 SD on reference curve), IGF-BP3 218 nmol/L (+1 SD on reference curve), TSH 1,5 mU/L (normal, ref. 0.4-4.3), FT4 15,9 pmol/l (normal, ref. 8.0-18.0), total Ca 2.53 mmol/L (normal, ref. 2.10-2.55), Phosphate 1,55 mmol/L (normal, ref. 0.90-1.50), Alkaline phosphatase 143 U/L (mildly elevated, ref. < 98).

Family 2, individual P3

Individual P3 (age 36) was born as second child of healthy, non-consanguineous parents. She had a birth weight of 2800 gram and a birth length of 45 cm. Apgar scores were 8 and 9 at one and five minutes, respectively. An enlarged posterior fontanel was noted after birth. Walking was slightly delayed. She followed regular primary school with some extra support in math and reading. Thereafter, she followed several courses and currently she has a job in a social workplace. Growth hormone tests were performed because of growth retardation in childhood, but no growth hormone deficiency was detected. Pubertal development was delayed and started at age 15 years, followed by her periods at age 17 years and 9 months. She developed normal secondary sexual characteristics. Currently she is using a contraceptive pill because of irregular menses. In her permanent dentition she only has 7 teeth, namely the upper medial incisors and 5 molars. The teeth are relatively small and have an abnormal shape. Since the age of 18 years, she has hearing aids due to at first progressive hearing loss. At the age of 34 years Fletcher index is 53 dB AD, 65 dB AS. Hearing has been stable in recent years. CT scan of the os petrosum showed normal inner ear structures with retracted eardrum of the left ear. She had amblyopia of the left eye for which she was treated with a patch. Ophthalmological examination at the age of 34 years revealed suboptimal vision with mild subcapsular posterior cataract and a bilateral refraction abnormality of +1.5 Dpt, right eye C-2.00 ax 145, left eye C-0.75 ax 25. She is increasingly complaining about fatigue problems and has an essential hypertension.

Individual P3 has a short stature with a height of 142,8 cm (-4,46 SD), skull circumference of 52,7 cm (-1,55 SD) and sitting height/height ratio of 0.51 (-1.09 SD) at the age of 26 years. Radiographs were taken at this age because of her short stature and endosteal hyperostosis was detected. She has several dysmorphic features like mild proptosis, high nasal bridge with a relatively long columella, broad nasal tip, downturned corners of the mouth and retrognathia. In addition, she has brachydactyly with short, stubby toes and syndactyly between the second and third toe. She has a soft skin, clear hyperlaxity of the elbow joints, fingers and knees. There is an increased lumbar lordosis and thoracic kyphosis.

Since the age of 33 years she is having increasing pain in the hips due to bilateral coxarthrosis, with the right hip being more severely affected than the left hip. She recently underwent arthroplasty of the right hip at the age of 34 years (Figure S5). Neurological examination at the age of 34 years showed bilateral hypotonia and proximal muscle weakness (MRC4) in the arms, but normal distal muscle strength. Due to hip problems, proximal muscles of the leg could not be tested, there was a normal muscle strength distally. She has mild varus deformity of feet. She also has hyperactive, but not pathological, reflexes. At the age of 34 years IGF-1 and IGFBP-3 are normal (respectively 13,8 nmol/L and 2,08 mg/L, -1.60 SD and -0.60 SD). Thyroid function and calcium phosphate metabolism is normal. FT4 is 15 pmol/l (ref. 10-22), total calcium

2,49 mol/L (ref. 2.20-2.60), phosphate 0.94 mmol/L (ref. 0.80-1.50), PTH 1,7 pmol/L (ref. 1.0-7.0).

Female P3 has two brothers and one of them also is affected with oligodontia with 11 teeth missing from his permanent dentition. The affected brother has had treatment with growth hormone because of growth hormone deficiency as a child. He is living abroad and unfortunately no further details could be retrieved. The other brother and the parents are healthy and have a normal height.

5.5.3 Supplemental figures

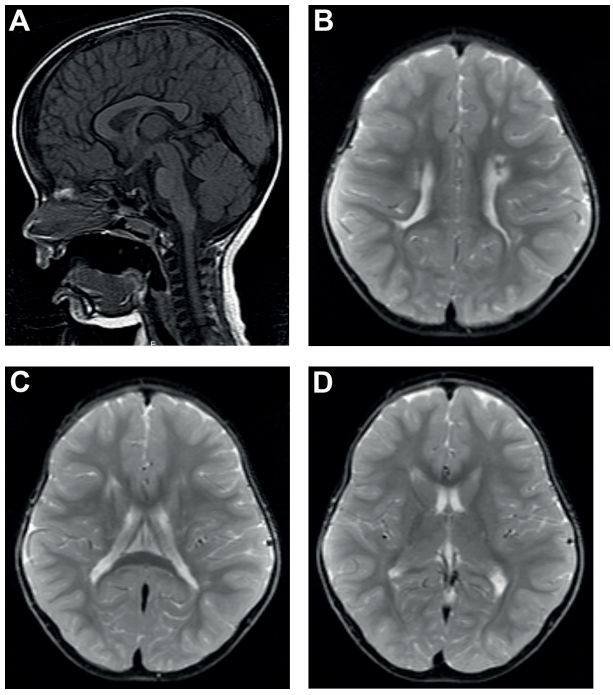


Figure S5.1 | MRI of the brain of individual P2 at age 3.5 showing white matter abnormalities most likely due to perinatal asphyxia. (A) *Sagittal T1-weighted image showing a focal thinning of the corpus callosum at the transition between corpus and splenium, and a normal aspect of the vermis of the cerebellum.* ***(B-D)*** *Axial T2-weighted images showing T2 hyperintense periventricular white matter lesions with signs of traction at the ventricular wall and diffuse white matter loss without signs of generalized hypomyelination, which are therefore most likely related to a (perinatal) hypoxic-ischemic incident.*

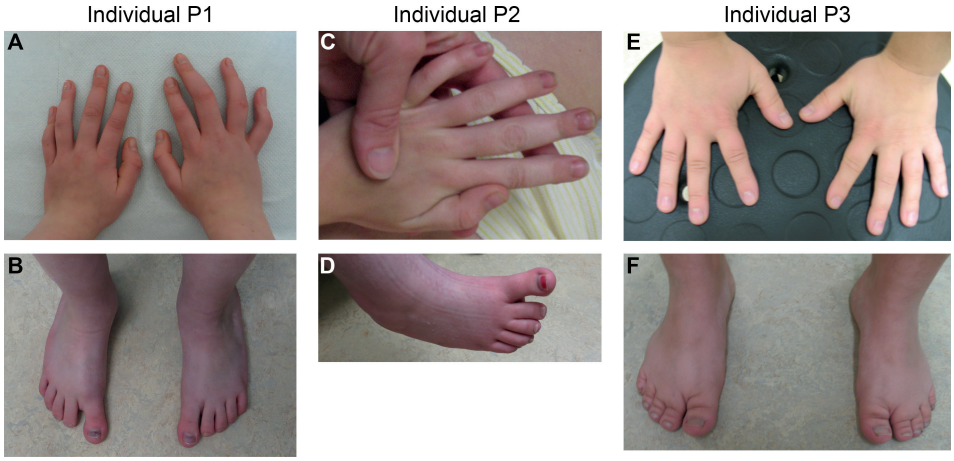


Figure S5.2 | Photographs showing hand and foot abnormalities. ***(A)*** *Hand of P1 at age 15 years showing bilateral camptodactyly of the fifth fingers, hyperextensible distal finger joints and relatively long middle finger compared with the index finger.* ***(B)*** *The feet of P1 at age 15 years showing sandle gap, long hallux, syndactyly of the second and third toe of both feet and mild varus position of feet.* ***(C)*** *Hand of P2 at age 9 years shows relatively long middle finger.* ***(D)*** *Picture of one of the feet of P2 at age 15 years shows a club foot with syndactyly between the second and third toe with a long hallux. (E) Picture of the hand of P3 at age 26 years shows brachydactyly.* ***(F)*** *Picture of the feet of P3 at age 26 years with short, stubby toes and syndactyly between the second and third toe.*

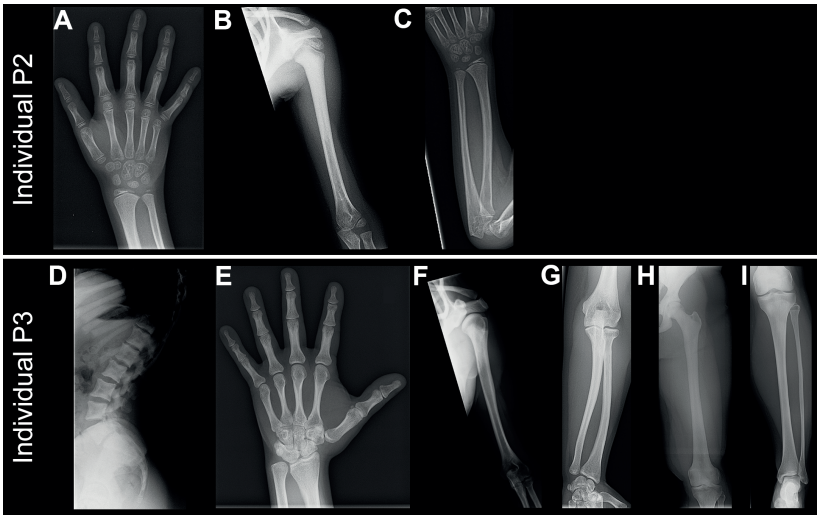


Figure S5.3 | Detection of endosteal sclerosis in individuals P2 and P3 by skeletal X-ray scans. ***(A-C)*** *X-rays of individual P2 at different ages. (A) X-ray at 9 years showing only mild endosteal sclerosis of several phalanges.* ***(B)*** *X-ray at 7 years showing endosteal sclerosis and diffuse metaphyseal sclerosis of the proximal humerus.* ***(C)*** *X-ray at 9 years showing no abnormalities in the lower arm.* ***(D-I)*** *X-rays of individual P3 at age 26 years.* ***(D)*** *X-ray showing sclerotic margins of the vertebral bodies and arches in the lumbar spine.* ***(E)*** *X-ray showing endosteal sclerosis of several bones in the hand, especially the metacarpal bones and proximal phalanges.* ***(F)*** *X-ray showing endosteal sclerosis of the proximal part of the humerus.* ***(G)*** *X-ray showing only mild endosteal thickening of the lower arm. (H) X-ray showing endosteal thickening and sclerosis of the upper leg.* ***(I)*** *X-ray showing mild patchy sclerosis of the proximal part of the tibia.*

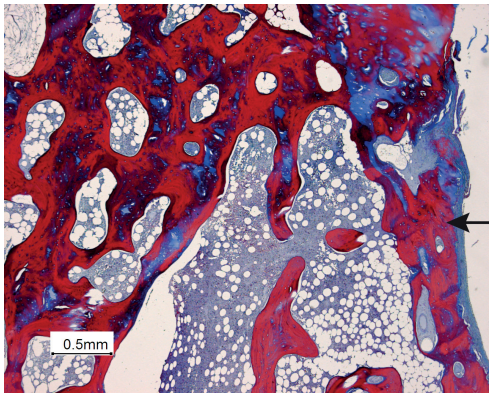


Figure S5.4 | Histological examination (trichrome staining) of the right femoral head after hip replacement of individual P3 at age 34. *The femoral head showed aspecific degenerative changes of joint cartilage. The bony trabeculae were prominent with signs of delayed remodelling and in the center of trabeculae woven bone. The arrow indicates pronounced increase of cortical bone consistent with osteosclerosis.*

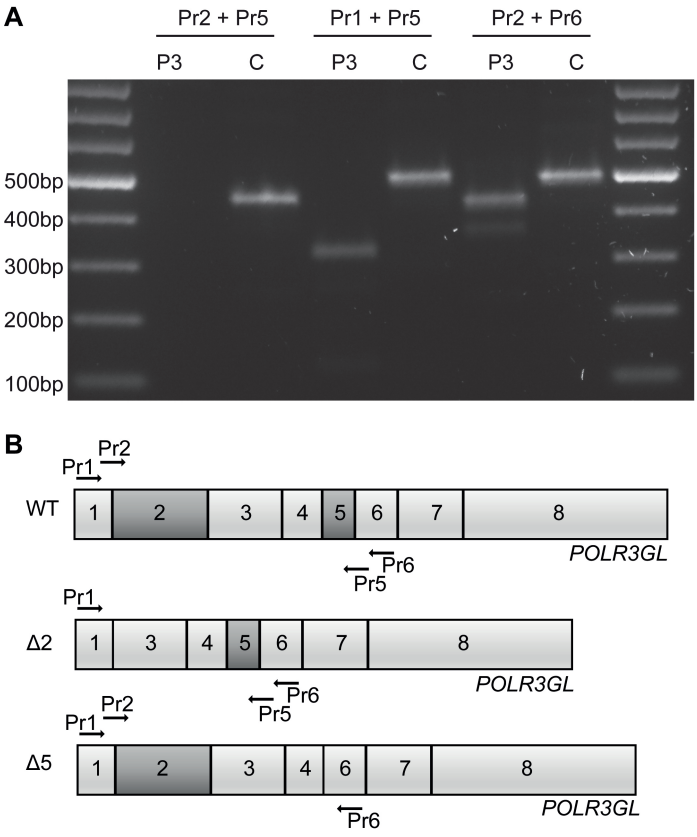


Figure S5.5 | PCR analysis showing absence of full length POLR3GL RNA transcripts in blood cells of individual P3. (A) *PCR products amplified from POLR3GL cDNA of individual P3 and a control sample separated on agarose gel. There is no PCR product detectable after amplification with primers complementary to exon 2 and exon 5 in individual P3 indicating that no full length POLR3GL transcript is present (lane 2). PCR amplification with primer complementary to exon 1 and 5 results in a PCR product with the size of transcripts missing exon 2 (321 bp, lane 4). PCR amplification with primer complementary to exon 2 and 6 results in a PCR product with the size of transcripts missing exon 5 (424 bp, lane 6).* ***(B)*** *Schematic overview of the POLR3GL transcripts in unaffected and affected individuals (missing exon 2 (Δ2) and/or exon 5 (Δ5)) and the locations of the PCR primers. PCR primers Pr1 and P2 are designed on respectively POLR3GL exon 1 and exon 6. Primers Pr2 and Pr5 are designed on respectively affected exon 2 and exon 5 and are not able to bind the POLR3GL cDNA if the respective exon is absent.*

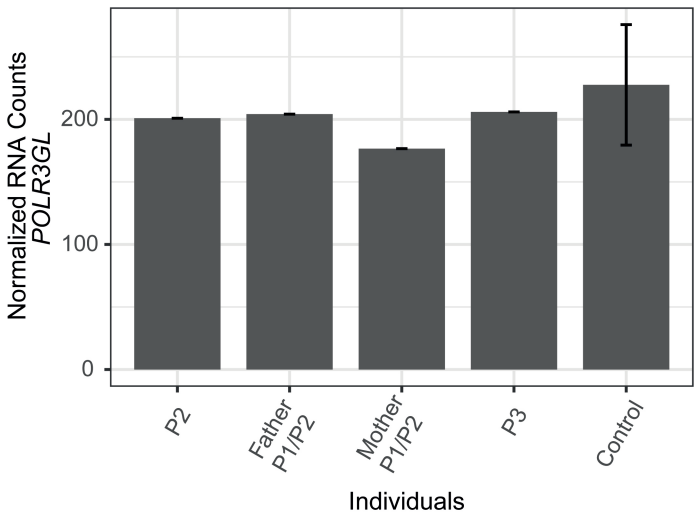
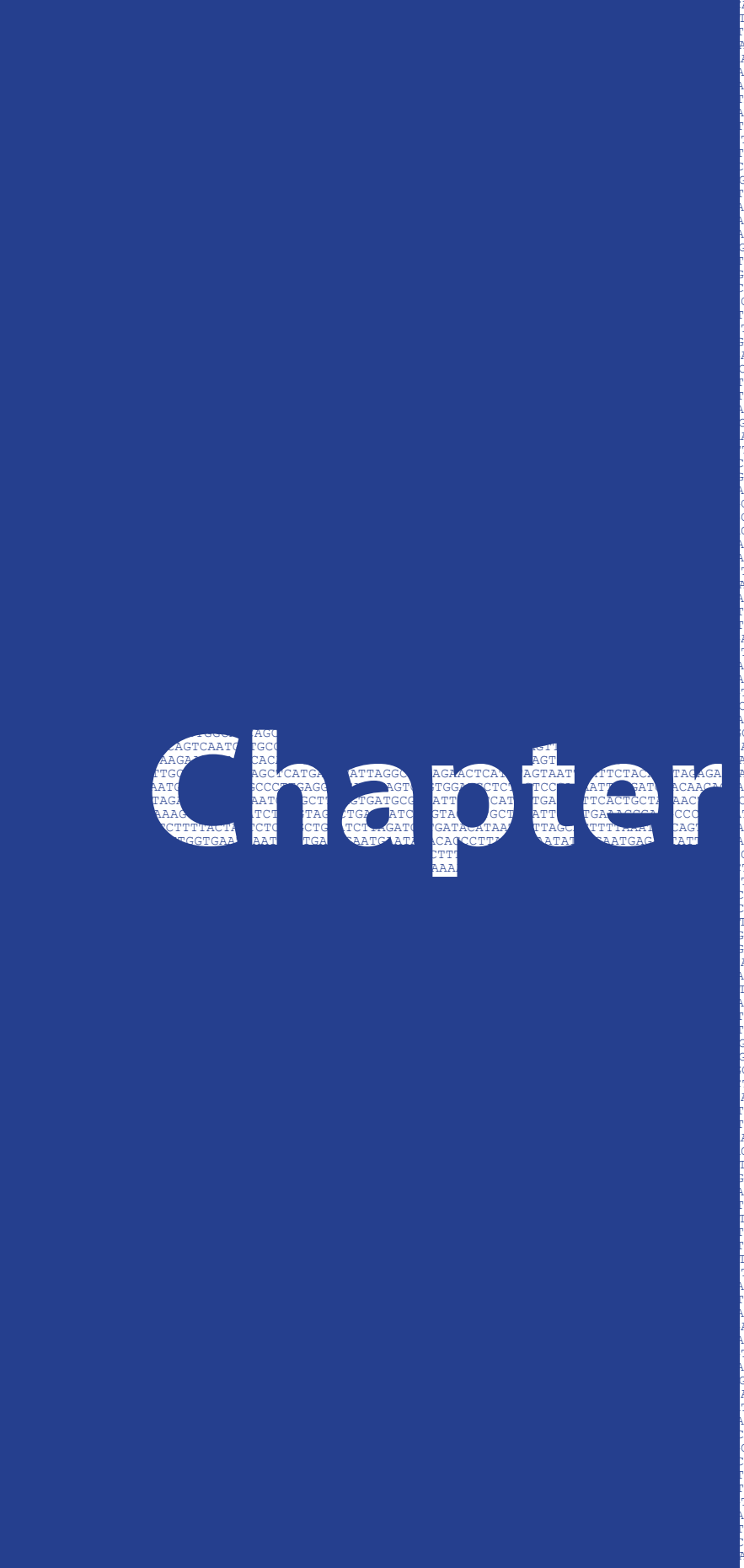


Figure S5.6 | POLR3GL transcript abundance is not affected in the individuals with biallelic POLR3GL variants. *POLR3GL mRNA expression levels in individual P2, her parents, individual P3 and 20 unaffected controls determined by RNA sequencing of peripheral blood mononuclear cells (PBMCs). Differential expression analysis did not show signficant differences in transcript abundances between affected individuals and control group.*

Chapter 6



Discussion

De novo structural variants (SVs) are an important cause of neurodevelopmental disorders. Rapid technological advances in the last two decades have made it feasible to routinely detect *de novo* SVs in individual patients. New opportunities for personalized medicine (or precision medicine) in which diagnosis and treatment are tailored to individual patients based on their genome can now be realized (Ashley, 2016). Large-scale sequencing studies have greatly improved our knowledge about the complexity and variability of the human genome. Despite the impressive technological advances, still half of the patients with neurodevelopmental disorders do not receive a genetic diagnosis. New disease-causing genetic variants, such as the novel biallelic variants in the *POLR3GL* gene in patients with endosteal hyperostosis and oligodontia described in chapter 5, are still frequently discovered. However, even if a pathogenic variant is identified in a patient, the precise molecular mechanisms leading to the disorder are frequently unknown. This is especially the case for SVs, which can affect many genes and regulatory elements in many different ways. In this thesis we applied multi-omics and computational approaches to study the molecular consequences of *de novo* structural variation in patients with neurodevelopmental disorders. In addition, we used single cell sequencing techniques to study the role of sperm DNA damage in the aetiology of *de novo* SVs, mosaicism and genomic instability in early mammalian embryos. Clinical implementation of whole genome sequencing and other sequencing techniques offer opportunities to improve the diagnostic yield in the near future. Large consortia, such as the International Rare Diseases Research Consortium (IRDiRC), have set ambitious goals to enable a rapid genetic diagnosis for every rare disease patient by 2027 (Austin et al., 2018). However, many remaining technical, computational, logistical and ethical challenges have to be faced in the coming years to achieve these goals.

6.1 Technical challenges in variant detection

6.1.1 Clinical implementation of whole genome sequencing improves diagnostic yield

A genetic diagnosis will be missed or incomplete if the causing variant(s) remains undetected after genetic testing. Detection of structural variants in the clinic has greatly improved since the introduction of genomic microarrays (Miller et al., 2010). However, most arrays can only detect large copy number variants (deletions and duplications) generally over 10 kilobases (Kb) in size (Alkan et al., 2011; Pinto et al., 2011). Arrays can be supplemented with karyotyping, which can be used to detect balanced rearrangements such as translocations, and whole exome sequencing (WES) to detect single nucleotide variants in genes. Despite their great usefulness, it is known that these techniques miss many relevant variants that can be detected by whole genome sequencing. This has been shown by several studies (Belkadi et al., 2015; Gilissen et al., 2014; Lelieveld et al., 2015; Meienberg et al., 2016; Stavropoulos et al., 2016; Trost et al., 2018), including our own study described in chapter 3, in which we showed that WGS detected additional *de novo* SVs and complexities of *de novo* SVs in 7 out of 18 cases (38%) that were previously studied by arrays. It is likely that WGS will replace many of the currently used genetic tests in the near future as a one-test-fits-all in clinical genetics.

Currently there is much discussion if, how and when WGS will be implemented in a clinical setting. Clinical implementation of WGS will require time, financial, staff training, logistical and computational investments, while many diagnostic labs have recently already invested great efforts in the implementation of WES. Scepticism about the performance of WGS in variant detection may play a role in holding back from implementing WGS. Currently used short read (2x 150 basepairs) Illumina-based WGS can capture most, but not all, types of genetic variation. Especially the detection of SVs remains a challenge, because there are no golden standard software tools for SV calling and filtering yet. In addition, many SVs are located around repetitive regions in the genome and these regions are still difficult to cover properly with short read sequencing. Long read sequencing has shown that there are more than 20,000 SVs in a human genome compared to the reference genome, whereas less than half of these can be detected by short read WGS (Chaisson et al., 2019; Collins et al., 2019; Nelson et al., 2019). However, most missed SVs are relatively small SVs ranging from 50 to 2000 basepairs, including many tandem repeat and retro-transposon insertions (Chaisson et al., 2019). Most, but not all, SVs affecting genes and regulatory elements can be detected by short-read sequencing (Chaisson et al., 2019). The vast majority of these missed SVs would have also been missed by current standard clinical tests. A recent study of the gnomAD consortium showed their WGS approach could detect 97.8% of the SVs (>40Kb) previously detected by microarrays in 1,893 individuals (Collins et

al., 2019). Despite remaining technical challenges, current short-read WGS already outperforms WES and microarrays in the detection of genetic variants. Technical considerations are therefore not one of the main reasons not to implement WGS in clinical genetics anymore. As will be discussed next, improvements in software will likely lead to further improvements in variant detection and filtering from WGS data. A major advantage of sequencing whole genomes is that the data generated now can be reanalysed later by new algorithms, without the need for resequencing (which may be required if only the exome is sequenced). Initially, WGS may serve as a second-tier genetic test especially for patients without a conclusive diagnosis from current standard genetic tests (like we used it in chapter 3). This may allow genetic diagnostics labs to gain experience with the challenges, but also the opportunities provided by WGS.

6.1.2 Development of bioinformatic tools to improve structural variant calling

Detection and filtering of SVs in WGS data remains a challenge due to the lack of a comprehensive gold-standard computational workflow. The genome of a patient contains millions of SNVs and thousands of SVs compared to the reference genome and most of these variants can be detected by WGS (1000 Genomes Project Consortium et al., 2015; Chaisson et al., 2019). However, only one or a few of these variants contributed to the disease phenotype of the patient. Stringent filtering of variants is essential to reduce the number of variants that need to be manually curated (Cooper and Shendure, 2011; Eilbeck et al., 2017). Too stringent filtering may lead to false negatives (true variants that are not reported) (Goldfeder et al., 2016) and therefore a balanced variant prioritization strategy is required. Currently it is still necessary to apply multiple SV calling and filtering algorithms to capture as many true SVs from WGS data as possible as each tool has its specific strengths for specific types of SVs or specific size range (Chaisson et al., 2019; Guan and Sung, 2016). Over 50 different SV calling methods with varying specificity, sensitivity and processing speeds have been described (Guan and Sung, 2016; Trost et al., 2018). In addition, several workflows or pipelines that integrate multiple SV callers (such as SVMerge, MetaSV, FusorSV and SURVIVOR) have been described. Initially it was difficult to compare the performances of different tools because of the limited availability of genome-wide, validated truth sets of SVs. In recent years, more high quality genomes have become available that can be used to benchmark variant callers (such as the genome sequences generated by the Genome In A Bottle (GIAB) consortium (Zook et al., 2018, 2014)) and this has led to more widely adoption of several best performing callers. It has been shown that combining two (such as Pindel and VariationHunter) or preferably three callers (such as Manta, VariationHunter and Lumpy) is optimal (Chaisson et al., 2019). Although there is no gold-standard SV calling workflow yet, combinations of SV callers can detect most

relevant SVs. The challenge of developing a gold-standard SV calling workflow is widely recognized and, with the enormous investments in WGS, it is likely that SV callers will further evolve and uniform workflows will be developed (Birney et al., 2017). The costs for processing and storing WGS data form a considerable portion of the total costs of performing WGS, especially with the declining costs of sequencing itself. Efficiency, speed and scalability of analysis software if therefore important to reduce the costs of data processing. Cloud-based solutions such as svtools that can take advantage of massive and efficient compute power are being developed (Larson et al., 2018) and may further streamline the processing of WGS data. This is especially important for large scale studies, where even small cost reductions per analysed genome can lead to large cost savings.

In addition to integration of multiple SV callers, it may also be possible to improve SV calling and filtering by using machine learning approaches. The rapid advancement of artificial intelligence, driven by the developments of faster algorithms and increases in computational capacities, is one of the most exciting developments in science and medicine. The explosion in the amount of generated sequencing data requires smart computational approaches to process and interpret the data (as will be discussed later). Machine learning approaches can be used to automatically detect patterns in complex genomic datasets, learn from these patterns and generate models describing these patterns. The models can be used to make predictions from other datasets and generate new hypotheses that could not have been anticipated by researchers (Eraslan et al., 2019; Libbrecht and Noble, 2015; Zou et al., 2019). Deep learning approaches have been applied to many areas within biomedical science in recent years (Wainberg et al., 2018). Deep neural networks have for example been applied to improve SNV and indel calling from both short and long reads (for example: (Luo et al., 2019; Poplin et al., 2018). Recently a tool to detect SVs in targeted sequencing data (not whole genome) based on deep learning has been published, showing that such an approach can be used to detect SVs (Park et al., 2019). Further refined machine learning algorithms to detect SVs are being developed and it will be interesting to see to what extent they can improve SV detection and filtering.

6.1.3 Dependence of next-generation sequencing on the quality of reference genomes

Currently whole genome sequencing is mostly based on resequencing in which the genomic positions of reads are determined by mapping to a reference genome. Although the sequence of the human reference genome is of very high quality, there are still gaps and errors in the genome also in some clinically relevant regions (Schneider et al., 2017). Additionally, the sequence of the reference genome was derived from several individuals (although most of it was obtained from one individual) (Schneider et al., 2017) and it is not a good representation of the population variation, which

can lead to biases (Paten et al., 2017). These shortcomings of the reference genome can lead to mapping errors and misinterpretation of genetic variants. Relevant reads overlapping SVs may be filtered out during data processing if they do not properly align with the reference genome. The performance of resequencing is dependent on the quality of the reference genome and therefore it is important that the reference genome is continuously being improved (Schneider et al., 2017).

Much of the knowledge about genomes and pathogenic variation is based on variation in the genomes of individuals from European ancestry. This has a negative impact on the diagnosis of patients of non-European descent and this creates inequalities in the healthcare system (Petrovski and Goldstein, 2016). It is important to study genomes from multiple ancestries at a large scale to improve diagnostics. To better deal with population variation, it may be more appropriate to select one of multiple reference sequences that better fit with the variation in an individual instead of using a single reference genome for everyone. This can be achieved by using graph-based representations of genomes (Paten et al., 2017). The most recent human reference genomes (GRCh38 and to a lesser extent GRCh37) contain blocks of alternative sequences representing genomic sequences (haplotypes) of different individuals for several highly variable genomic regions (Schneider et al., 2017). Reads not mapping properly to sequence of the major locus in the reference genome may better map to one of the alternative loci, which may further improve mapping and variant detection (Paten et al., 2017). However, this also requires software that can handle these graph-based representations of genomes. Genome graphs are still in early development, but, although it may still take many years, may have a large impact on the way genomic data is analysed (Paten et al., 2017). Reference genomes are used as a coordinate system and updates to reference genomes sometimes involve changes to the coordinates (for example during the update from GRCh37/hg19 to GRCh38/hg38). Although coordinates used in genomic datasets can be translated to another coordinate system, this regularly introduces errors (Pan et al., 2019) and it requires substantial efforts, especially if large datasets or complex computational tools are involved. Therefore, it can take years before improved versions of reference genomes are widely adopted, which can lead to data duplication and fragmentation. Thus, representations of reference genomes are continuously evolving, but researchers will have to co-evolve to optimally benefit from improvements in reference genome.

In addition to improving the sequence of the reference genome, it is also important to improve the annotation of the genome. A large fraction of the genome still has an unknown function and even for genes there is no full consensus about their locations and structures. Several databases containing gene models (exon-intron structures) exist (such as Genbank and Ensembl) and they contain different numbers of genes (Mudge and Harrow, 2016). In addition, genes can have alternative transcripts containing different exonic sequences coding for different protein isoforms (Mudge

and Harrow, 2016). As a consequence, a detected variant may overlap with a region that is considered as a gene by one database, but as a non-coding region in another database. Thus, further refining gene annotation is important to improve variant annotation (Eilbeck et al., 2017). There are still many opportunities to improve human reference genomes and genome annotations, which may contribute to improve variant detection and annotation. The precise impact of these improvements on the detection and interpretation of structural variants remains to be demonstrated.

6.1.4 Long read sequencing opens new possibilities in SV detection

Even though new developments in SV detection and filtering will further improve our ability to study SVs, the short length of the reads routinely generated by Illumina sequencers will remain a hurdle for comprehensive SV detection. Many SVs are associated with repetitive regions in the genome, but these regions are difficult to sequence with short reads. Longer reads can contain more unique sequences and therefore may be better aligned to a reference genome. In addition, long reads may span entire repetitive elements and large parts of structural variants and are therefore better suited to detect SVs and reconstruct complex SVs (Chaisson et al., 2015a, 2019). The length of reads typically generated by Illumina sequencing has improved in recent years (the first Illumina sequencers generated reads of only 36 basepairs (Bentley et al., 2008), but nowadays reads of 2x150 basepairs are common for WGS), but it may be difficult to increase the length of reads much further based on the currently used Illumina sequencing chemistry. In recent years fundamentally different technologies have evolved to routinely sequence very long reads. Reads of over 10Kb are generated by single-molecule real time (SMRT) sequencing from Pacific Biosciences (PacBio) (Eid et al., 2009) and nanopore-based sequencing from Oxford Nanopore Technologies (ONT) (Deamer et al., 2016) and even reads longer than 100Kb can be produced. In addition, linked reads technology, which does not generate "real" long reads, but makes use of microfluidics and barcoding to combine short-reads to cover regions of ~100Kb, has been developed by 10X Genomics (Marks et al., 2019). These technologies are still rapidly developing, but they have already realized their potential by improving reference genomes and by detecting many SVs that could not be detected by short-read sequencing (for example: (Chaisson et al., 2015a, 2019; Cretu Stancu et al., 2017; Nelson et al., 2019)). Long read sequencing also has the potential to improve other sequencing applications, such as transcriptomics, epigenomics and chromatin conformation capture methods (Sedlazeck et al., 2018). Disadvantages of current long read sequencing technologies are relatively high error rates and difficulties sequencing homopolymer runs. However, error rates are declining with technological improvements, deeper sequencing coverage and improved error correction software (Sedlazeck et al., 2018). The sequencing costs per base and the SNV error rates are still much higher than short-read sequencing and therefore long-read sequencing

will likely not be broadly adopted in the clinic soon. Long-read sequencing can be beneficial for the diagnosis of specific disorders involving repeats. In addition, capture of specific relevant regions before sequencing may reduce sequencing costs. The long-read sequencing technologies are relatively new and still rapidly evolving and it will take some more years for the technology to mature into clinical applications with lower error rates and reduced sequencing costs. The precise impact of long-read sequencing on future clinical genetics is hard to guess, like it was unimaginable how fast NGS techniques would develop 15 years ago. Nevertheless, especially detection of SVs in diagnostics can be improved by further maturation of these technologies.

6.1.5 Detecting SVs in de novo genome assemblies

Ideally genomes are assembled *de novo* directly from the sequencing reads to detect all variants without first mapping the reads to a (biased) reference genome. However, currently it is not practical yet to generate a high-quality *de novo* human genome assembly without large gaps from short reads (without performing Hi-C) due to the repetitive elements in the genome (Chaisson et al., 2015b). Long-read sequencing makes high-quality *de novo* assembly more feasible, although generation of an assembly is still computationally intensive and is not perfect yet, especially for diploid genomes with heterozygous regions (Sedlazeck et al., 2018). *De novo* assemblies have the potential to improve variant detection, but further developments in technology and algorithms are required for routine application of *de novo* assemblies (Chaisson et al., 2015b; Sedlazeck et al., 2018). An interesting solution to take advantage of the power of *de novo* assembly is to perform such assemblies only for specific regions in the genome such as regions with indications for the presence of an SV. Various tools such as Manta, novoBreak and SvABA can perform such a local (or micro) assemblies by only assembling a selection of specific reads (for example reads not mapping properly to the reference) (Chen et al., 2015; Chong et al., 2016; Wala et al., 2018). These hybrid approaches combine high sensitivity with high efficiency and have become popular tools to detect SVs from short-read WGS data.

6.2 Challenges in interpretation of structural variation

6.2.1 Large scale sequencing projects to improve the interpretation of genomes

Most genetic variants can already be captured by current WGS, but for many of these variants we do not know if and how they contribute to disease. Further technical advances in sequencing technology and software will help to improve the yield of genetic diagnostics, but the largest gains can likely be achieved by improving our abilities to interpret the genomic data. A large number of disease-causing variants have been identified in the last decade due to the large-scale introduction of whole

exome sequencing. Still many new associations between genes and disease are being discovered, although the rate of discovery appears to be declining. An essential factor underlying the success of gene discovery is the integration of all these genomic datasets in large databases such as ExAC, GnomAD, 1000 Genomes, Database of Genomic Variants, ClinVar or DECIPHER. Especially for SNVs these databases have been extremely useful so far to calculate the frequencies of the variants in populations, which allows filtering of SNVs based on low allele frequencies (which are more likely to be pathogenic compared to common variants). The Database of Genomic Variants (DGV) is the largest database for structural variants containing a collection of more than 500,000 unique SVs determined in over 20,000 genomes by many (>70) different arrays or low-coverage WGS studies in healthy individuals (MacDonald et al., 2014). Recently two large SV datasets containing hundreds of thousands unique germline SVs detected in high-quality WGS data of respectively 17,795 and 14,891 individuals have been released (Abel et al., 2018; Collins et al., 2019). These databases contain SV population allele frequencies which will be of great value to filter for SVs that are rare in healthy individuals.

6.2.2 Unlocking the knowledge potential of clinical genomes by data sharing

In addition to integrating SVs from healthy individuals, it is of course also essential to collect and share genomic and phenotypic information from patients. DECIPHER is an important database from the Deciphering Developmental Disorders (DDD) consortium which currently contains information of over 29,000 patients (Firth et al., 2009). This data source has been important for the discovery of many new pathogenic variants shared by multiple patients. However, the database contains mostly pre-filtered variants and the quality of the variant detection and classification and the descriptions of the phenotypes is very variable, making interpretation of the data sometimes difficult. Continuous improvements, maintenance and curation of such databases are important, not in the last place because they frequently contain errors that can have far-reaching consequences (Wright et al., 2018a). However, the amount of available patient data in open databases is very low compared to the thousands of clinical WES and array datasets that are generated weekly by genetic centres worldwide (Boycott et al., 2019). Crude estimates suggest that more than 60 million patients (including 20 million patients with rare diseases) will have their genomes sequenced by 2025 (Birney et al., 2017). This clinically generated data contains a wealth of potential new knowledge about genetic variation, but currently the data is usually not widely shared, stored in in-house databases and unavailable for large-scale integrative research. The variants in the *POLR3GL* gene we described in chapter 5 for example were discovered because the exomes of the patients were sequenced in the same hospital by chance and therefore the patients could be matched. It is likely that the exomes of other

patients with similar variants in the *POLR3GL* have already been sequenced somewhere else, but that these variants could not be not classified as (likely) pathogenic. Currently, there are very useful platforms available to share variants of unknown significance, such as GeneMatcher (Sobreira et al., 2015). However, sharing through these networks is not (fully) automated and requires manual interpretation, selection and uploading of variants or genes by a clinical geneticist or researcher.

Sharing of clinical data will give patients the opportunity to contribute to genetic research. However, upscaling of data sharing will require streamlining of data analysis pipelines, systematic variant interpretation and standardized phenotype descriptions. Additionally, a cultural shift is necessary, because the traditional competitive attitude between research labs frequently leads to a reluctance to openly share data. Finally, and importantly, several ethical challenges have to be overcome (Kaye, 2011). The genome sequence contains very privacy-sensitive information and it is nearly impossible to fully anonymize a genome which opens a risk for re-identification of genomic data. Therefore, there has to be a balance between responsible data sharing and the privacy of an individual. It is possible to only share specific, potentially interesting variants or to only give restrictive access to genomic and phenotypic information (The Global Alliance for Genomics and Health, 2016; Wright et al., 2016). Updated legislation and international agreements about the use of genomic data, but also practical guidelines for clinicians and researchers, are necessary to ensure the privacy of the patients who have their genomes sequenced. Additionally, implementation and sharing of whole genome sequencing data will undoubtedly lead to the detection of more variants of unknown significance and unsolicited findings. It is challenging and frequently undesirable to explain such findings to patients and their families and clear protocols are required to guide the clinicians in decision-making (Pollard et al., 2019). Thus, large scale integration of genomic data has been crucial for the discovery of many new pathogenic genetic variants and upscaled sharing of clinically generated genomic data will be a rich source for new variant discovery. Many national and global initiatives are actively dealing with the logistical, security and ethical challenges coming with large-scale sharing of privacy-sensitive genomic data (Stark et al., 2019).

6.2.3 Multifactorial causes of congenital disorders

In addition to discovery of new pathogenic genetic variants, it is also important to gain more insight in currently not well-understood molecular mechanisms leading to disease. For example, it was recently suggested that the phenotypes of around 10% of patients with genetic disorders may be caused by *de novo* SNVs causing cryptic splice sites in introns (Batzoglou et al., 2019). The authors developed a deep learning approach to detect such variants that can have a large impact on gene function. Such a disease mechanism is underappreciated, because it was previously difficult to detect, not in the last place because these variants are located in introns that are not covered

by whole exome sequencing. Identifying these intronic variants causing cryptic splice sites with WGS therefore has the potential to considerably improve diagnostic yield. Conversely, it has been suggested that variants causing nonsense mediated decay of mRNA (such as frameshift and nonsense mutations) may not always be as pathogenic as generally thought, because of compensation by not well-understood feedback mechanisms leading to transcriptional adaptation (El-Brolosy et al., 2019). Pathogenicity of variants is currently mostly based on recurrence of the variant in patients with similar phenotypes and the absence of the variants in unaffected individuals. However, the precise molecular consequences of variants leading to disease are frequently unknown, especially for structural variants that can affect many genes. A molecular diagnosis is usually based on a single genetic event, but it may well be possible that multiple variants together lead to the phenotype. Most phenotypic traits are determined by multiple genes, but the importance for multi-genic effects in congenital disease is not clear, largely because it is difficult to determine such effects. In chapter 3 and 4 we showed several examples in which combinations of multiple genes affected by (complex) SVs may have contributed to the complex phenotypes of the patients. Additionally, in chapter 5 we found compound heterozygous single nucleotide variants leading to biallelic disruption in the *POLR3GL* gene of a patient with endosteal hyperostosis and oligodontia. Routine screening for such relatively simple compound heterozygous variants within a single gene is feasible if both parents are sequenced (Eilbeck et al., 2017). However, it may well be possible that many disorders are caused by combinations of different types of variants affecting a gene or multiple genes in the same pathway. It may for example be possible that regulation of a gene is disrupted on one allele by an SV and the function of the gene affected by a SNV on the other allele. It is challenging to detect such effects, partly because SVs and SNVs are frequently treated separately in variant interpretation workflows. However, multiple separate variants or genes may each explain part of a phenotype and therefore it can be important not to discard variants with only a potential minor impact and not to stop looking for other pathogenic variants if already one such variant is found (as is frequently done with gene panels).

6.2.4 Deciphering the role of non-coding variation in developmental disorders

The vast majority of currently known pathogenic variation is located within genes. Non-coding variation forms a potential source to further improve the diagnostic yield, but the significance of this variation in developmental disorders is unclear. A large part of the non-coding genome shows biochemical activity and is involved in regulation of gene expression (Kellis et al., 2014). Although much has been learned about the organization of the genome in recent years, still much remains to be explored about the precise roles of all involved proteins (Bonev and Cavalli, 2016; Rowley and

Corces, 2018). The expression of many developmental genes is controlled by multiple enhancers and loss of a single enhancer usually only has a mild impact on gene expression (Gasperini et al., 2019; Osterwalder et al., 2018). Single nucleotide variants in enhancers therefore likely only cause disease in some cases (Short et al., 2018). Structural variants in contrast can affect many regulatory elements and therefore can have a devastating effect on the regulatory landscape of genes important in embryonic development (Spielmann et al., 2018; Weischenfeldt et al., 2013). Indirect, positional effects of SVs on gene expression are challenging to study mainly due to the cell-type specificity of these effects, which, in the case of developmental disorders, take place during embryonic development and may not be detectable after birth. The regulatory landscape of genes may also be less conserved as gene sequences, which complicates study of positional effects in model organisms. Therefore, we generated and differentiated patient-specific induced pluripotent stem (iPS) cells to study positional effects in disease relevant cell types in chapter 4. As also has been shown by other studies, iPS cells are very useful to model positional effects of SVs (Laugsch et al., 2019), but they also have some limitations. Additional genomic rearrangements may arise during derivation and propagation of the cells, as we also found in one of the clones we cultured, highlighting the importance to cultivate multiple clonal lines. In addition, differentiation of the iPS cells can be quite heterogeneous between clones and even within clones, resulting in the presence of multiple distinct cell types with differences in gene expression and genomic interactions. Single-cell transcriptomics may be useful to correct for heterogeneity in differentiation status between cells (provided the genes of interest have detectable gene expression) (Nguyen et al., 2018).

Despite the usefulness of iPS cells to study molecular consequences of SVs, it remains relatively time and labour intensive to derive these cells from many patients. Therefore, we developed a computational tool in chapter 3 to predict effects of SVs on genes driving the developmental disorder phenotypes based on WGS data, phenomatching and publicly available chromatin conformation data. SVs can affect hundreds of genes and our approach helps to reduce the number of genes requiring manual validation by prioritizing candidate driver genes. Disadvantage of such prioritization tools are the risk for overinterpretation and overdiagnosis and the dependency on public databases, which may be incomplete or contain errors. Therefore, experimental validation of putative pathogenic variants will remain essential. Validation of clinical genetic variants can also yield important insights in fundamental biological processes. For example, studies of variants in *RSPO2*, which are involved in extreme human limb phenotypes, lead to new insights in unexpected functions of this gene in the Wnt pathway (Szenker-Ravi et al., 2018). In chapter 5 we found likely pathogenic variants in the RNA polymerase III subunit POLR3GL. It is quite surprisingly that there are hardly any functional studies on variants in RNA polymerase III genes, despite the importance of these genes in fundamental transcriptional

processes and the interesting tissue-specific phenotypes described in hundreds of patients. Developments in genome editing technology will likely help to make experimental validation more efficient.

6.2.5 Mosaicism is likely an underappreciated cause of developmental disorders

Mosaicism is another phenomenon of which the relevance in developmental disorders is still not well understood. Ongoing mutagenesis after fertilization can lead to formation of multiple cell populations with different genotypes within an individual (Biesecker and Spinner, 2013). The impact of mosaic variants depends on the timing when they arise and the cell types they affect. As we have also shown in chapter 2, mosaicism is very common in early mammalian embryos (McCoy, 2017) and it is even more prevalent and extreme in embryos derived from fertilization with damaged sperm. Although many of the embryonic cells affected by mosaic genetic variants will not contribute to further embryonic development, some cells may be at the basis of important cell lineages if the embryo survives all selective barriers. Mosaic variants may for example end up in specific cells of the developing brain and thereby disturb normal embryonic development. It is known that genetic mosaicism is very common in the brain and several developmental disorders have been shown to be caused by mosaic variants (Acuna-Hidalgo et al., 2016; Rohrback et al., 2018). However, mosaic variants are usually difficult to detect, because they only occur in a subset of cells. Samples for genetic testing are usually obtained from blood, but potential pathogenic mosaic variants may not be present or present at low frequencies in blood and are therefore missed by routine diagnostics. It may therefore be worthwhile to screen multiple tissues or affected tissues originating from different embryonic lineages for genetic variants and/or perform ultra-deep sequencing in patients who could not be diagnosed by regular blood-based genetic tests. Thus, although it is known that mosaicism is a common phenomenon, the relevance of mosaic structural and single nucleotide variants in developmental disorders is likely underappreciated due to the difficulty to detect such variants.

6.2.6 Benefiting from new knowledge by reiteratively analysing genomic data

The continuous discovery of new pathogenic variants makes it worthwhile to perform regular reanalysis of previously generated sequencing data to improve diagnostic yield. Variants classified as variant of unknown significance (VUS) at some point may later turn out to be pathogenic or benign based on new knowledge. Various studies have shown the value of reanalysing old data with new computational pipelines and comparing previously detected variants with updated variant annotation databases (Shuman et al., 2018; Wenger et al., 2017; Wright et al., 2018b). For example, reanalysis

of three-years old whole exome sequencing data generated by the Deciphering Developmental Disorders study increased the diagnostic yield from 27% to 40% (Wright et al., 2018b). In chapter 3 we analysed the whole genomes of patients with previously detected variants of unknown significance and in a few cases we also found new evidence for pathogenicity in literature (such as a duplication of the *TBL1XR1* gene (Riehmer et al., 2017)). It is also possible that variants currently interpreted as pathogenic might be reclassified to benign (it is quite common that variants are interpreted differently by different labs and there are many errors in public databases and older literature). Reanalysis of old data can be challenging and time consuming and therefore it is worthwhile to partially automate this process. Computational phenomatching, based on digitized phenotypes, for example with Human Phenotype Ontology (HPO) terms, can be very useful for this purpose. The HPO database contains a wealth of gene-phenotype associations and it is constantly being updated with the newest knowledge. A disadvantage is that the system currently is purely gene-based and not variant based. The type of variant affecting a gene of course makes a large difference (for example if a gene is affected by a deletion or by a duplication), but current phenomatch approaches cannot make a distinction by this. It would be very valuable if specific variants could be associated with HPO terms instead of just the genes. HPO terms may not only be useful for reanalysis, but also for sharing clinical data. Eventually a clinician or a lab specialist will have to look at the reanalysed data (Salmon et al., 2018), because digital description rarely captures the entire picture of phenotypes that are frequently very complex and also subject to changes in time (especially in the diagnostics of children). Nevertheless, automated prioritization of reanalysed variants can be very useful to reduce the amount of data that needs to manually curated.

6.3 Elucidating the causes of *de novo* SVs in the germline

Technological advancements in sequencing technologies have not only been a great benefit for patient diagnostics and care, but also in fundamental genome research. Genome sequencing has for example lead to many new insights in the mechanisms that lead to the formation of SVs. One of the major advantages of SV detection by WGS is the high resolution, allowing the detections of breakpoint junctions at nucleotide resolution. The nucleotide content (including the presence of (micro-) homology) around breakpoint junctions forms a molecular scar that gives insight in the double-strand break repair mechanisms that created the junction. Studying the scars of *de novo* SVs has yielded some new knowledge in the roles of the various repair mechanisms during gametogenesis and early embryogenesis. Nevertheless, still much remains unknown about double strand break repair in gametes and early embryos. Knowledge about these repair mechanisms in embryos has become increasingly important due to new possibilities in embryonic genome editing by CRISPR/Cas systems, which

make use of DSB repair, for example to generate model organisms. Germline cells are thought to be more protected against mutagenesis, because germline mutation rates are lower than somatic mutation rates. Therefore, it is surprising that early mammalian embryos frequently show genomic instability and chaotic mosaicism, which could be identified by copy number profiling of single embryonic cells (Carbone and Chavez, 2015; McCoy, 2017; Vázquez-Diez and Fitzharris, 2018; Voet et al., 2011). Most *de novo* SVs are located on chromosomes inherited from the father, which suggests a role for sperm DNA damage in the induction of these SVs. In chapter 2 we made use of single cell genome sequencing to show that post-meiotic sperm DNA damage can contribute to formation of *de novo* SVs, embryonic genomic instability and mosaicism. The precise mechanisms are still unclear, but it seems that repair of paternal DSBs by the fertilized oocyte may cause a mitotic delay, leading to chromosomal misalignments, lagging chromosomes and segregation errors during the first cell division (Chavez et al., 2012; Coonen et al., 2004; Vázquez-Diez and Fitzharris, 2018). As a consequence, multipolar cell divisions may occur during the first or second cell division (Kalatova et al., 2015), which lead to random distribution of chromosomal fragments of paternal origin and even to heterogoneic divisions, in which the maternally and paternally-derived genomes are separately distributed to haploid/uniparental daughter cells (Destouni and Vermeesch, 2017). Live-cell time-lapse imaging preferably with labelled DNA combined with single cell sequencing is required to further study the consequences of sperm DNA damage on embryonic genome integrity in more detail.

Additionally, it will be interesting to determine how embryos deal with mosaicism later in development. Most embryos with extensive chromosomal abnormalities will like arrest in development, which is the case for most bovine embryos produced with damaged sperm as well as most human embryos. All sequenced two- and eight-cell embryos derived from fertilization with 10 Gy-irradiated sperm showed genomic abnormalities, but nevertheless some embryos in this treatment group successfully developed into blastocysts. This indicates that some mosaic embryos are able to deal with genomic abnormalities and escape developmental arrest. This is also illustrated by the presence of mixoploid or uniparental lineages in some placentas and even in some live-born humans, which likely originate from early embryonic mosaicism. Some studies suggest that genomic abnormal embryonic cells may selectively undergo apoptosis (Bolton et al., 2016) or are isolated to extraembryonic tissues, allowing rescue of development by euploid cells (McCoy, 2017). Lineage tracing experiments combining imaging with single cell sequencing at later stage embryos may give more insight in how these embryos deal with early-developmental genomic instability. It is important to keep in mind that there are substantial inter-species differences in early embryonic development and that not all findings in model organisms may apply to human development (Carbone and Chavez, 2015).

Increasing our fundamental understanding of early mammalian development

can have a profound impact on healthcare as well. Infertility is a large problem and many infertile couples are dependent on assisted reproductive technology (ART) including in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI). IVF is frequently not successful due to various reasons and sperm DNA damage may be an undervalued cause of failed attempts (Colaco and Sakkas, 2018). Infertile men may have more sperm DNA damage, which may have an impact on the genomic integrity of embryos. IVF lowers the selection barrier for sperm cells and therefore may increase the chance of fertilization with damaged sperm. There are some indications that birth defects are more common in children born via IVF (Colaco and Sakkas, 2018) and it would be interesting to determine if these children have higher levels of *de novo* and mosaic genetic variation in their genomes.

6.4 Conclusions

The field of genomics has advanced tremendously since the finishing of the human reference genome in 2003. Genomes of millions of people will be sequenced in the coming years, which will have an enormous impact on healthcare. While sequencing of genomes has become routine, interpretation of genomes and especially structural variants remains a major challenge. Developments in variant calling and filtering software, computational methods such as deep learning, long-read sequencing, reference genomes and genome assemblies will aid our abilities to detect pathogenic SVs and to determine the consequences of SVs. Each of these developments may have a minor contribution to improving detection and interpretation of SVs, but together they may have a large impact to the way we diagnose patients. Combined with broad scale integration of genomic data, these developments will give allow discovery of new pathogenic variants and they can provide more insights in the least understood disease mechanisms, such as non-coding, multigenic or mosaic mechanisms. The fast developments in genomics make close collaboration between research and the clinic increasingly important for mutual benefit (Birney et al., 2017). The increasing importance of genomics in medicine and the associated ethical considerations also make broader public discussion and education about the role of genomics in society necessary. With the rapid technological progression and broad implementation of WGS, it is not unimaginable that the we will understand the effects of most SVs in the near future. This will contribute to the goal to provide a conclusive genetic diagnosis for every patient with a developmental disorder in the future.

Addendum

References

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68–74.

Aardema, H., van Tol, H.T.A., Wubbolts, R.W., Brouwers, J.F.H.M., Gadella, B.M., and Roelen, B.A.J. (2017). Stearoyl-CoA desaturase activity in bovine cumulus cells protects the oocyte against saturated fatty acid stress. Biol. Reprod. 96, 982–992.

Abel, H.J., Larson, D.E., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., Buyske, S., et al. (2018). Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. BioRxiv 508515.

Acuna-Hidalgo, R., Veltman, J.A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol. 17, 1–19.

Ali, T., Renkawitz, R., and Bartkuhn, M. (2016). Insulators and domains of gene expression. Curr. Opin. Genet. Dev. 37, 17–26.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. 12, 363–376.

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. 17, 487–500.

Allshire, R.C., and Madhani, H.D. (2018). Ten principles of heterochromatin formation and function. Nat. Rev. Mol. Cell Biol. 19, 229–244.

Amarillo, I.E., Dipple, K.M., and Quintero-Rivera, F. (2013). Familial Microdeletion of 17q24.3 Upstream of SOX9 Is Associated With Isolated Pierre Robin Sequence Due to Position Effect. Am. J. Med. Genet. Part A 161, 1167–1172.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169.

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A Switch Between Topological Domains Underlies HoxD Genes Collinearity in Mouse Limbs. Science (80-.). 340, 1234167–1234167.

Antonarakis, S.E., Avramopoulos, D., Blouin, J.-L., Conover Talbot, C., and Schinzel, A.A. (1993). Mitotic errors in somatic cells cause trisomy 21 in about 4.5% of cases and are not associated with advanced maternal age. Nat. Genet. 3, 146–150.

Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. 17, 507–522.

Austin, C.P., Cutillo, C.M., Lau, L.P.L., Jonker, A.H., Rath, A., Julkowska, D., Thomson, D., Terry, S.F., de Montleau, B., Ardigò, D., et al. (2018). Future of Rare Diseases Research 2017–2027: An IRDiRC Perspective. Clin. Transl. Sci. 11, 21–27.

Baart, E.B., Martini, E., Eijkemans, M.J., Van Opstal, D., Beckers, N.G.M., Verhoeff, A., Macklon, N.S., and Fauser, B.C.J.M. (2007). Milder ovarian stimulation for in-vitro fertilization reduces aneuploidy in the human preimplantation embryo: a randomized controlled trial. Hum. Reprod. 22, 980–988.

Bakker, B., Taudt, A., Belderbos, M.E., Porubsky, D., Spierings, D.C., de Jong, T. V, Halsema, N., Kazemier, H.G., Hoekstra-Wakker, K., Bradley, A., et al. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biol. 17, 115.

Batzoglou, S., Arbelaez, J., Kia, A., Sanders, S.J., Li, Y.I., Kanterakis, E., Kyriazopoulou Panagiotopoulou, S., Schwartz, G.B., Cui, W., Farh, K.K.-H., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535-548.e24.

Bean, C.J. (2002). Fertilization in vitro increases non-disjunction during early cleavage divisions in a mouse model system. Hum. Reprod. 17, 2362–2367.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. 112, 5473–5478.

Benko, S., Fantes, J.A., Amiel, J., Kleinjan, D.J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C.T., et al. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat. Genet. 41, 359–364.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., et al. (2013). An estimation of the number of cells in the human body. Ann. Hum. Biol. 40, 463–471.

Biesecker, L.G., and Spinner, N.B. (2013). A genomic view of mosaicism and human disease. Nat. Rev. Genet. 14, 307–320.

Birney, E., Vamathevan, J., and Goodhand, P. (2017). Genomics in healthcare: GA4GH looks to 2022. BioRxiv 203554.

Bolton, H., Graham, S.J.L., Van der Aa, N., Kumar, P., Theunis, K., Fernandez Gallardo, E., Voet, T., and Zernicka-Goetz, M. (2016). Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. Nat. Commun. 7, 11165.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nat. Rev. Genet. 17, 661–678.

van den Bos, H., Bakker, B., Taudt, A., Guryev, V., Colomé-Tatché, M., Lansdorp, P.M., Foijer, F., and Spierings, D.C.J. (2019). Quantification of Aneuploidy in Mammalian Systems. Methods Mol. Biol. 1896, 159–190.

Boycott, K.M., Hartley, T., Biesecker, L.G., Gibbs, R.A., Innes, A.M., Riess, O., Belmont, J., Dunwoodie, S.L., Jojic, N., Lassmann, T., et al. (2019). A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers. Cell 177, 32–37.

Brand, H., Collins, R.L., Hanscom, C., Rosenfeld, J.A., Pillalamarri, V., Stone, M.R., Kelley, F., Mason, T., Margolin, L., Eggert, S., et al. (2015). Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. Am. J. Hum. Genet. 97, 170–176.

Brandler, W.M., Antaki, D., Gujral, M., Kleiber, M.L., Whitney, J., Maile, M.S., Hong, O., Chapman, T.R., Tan, S., Tandon, P., et al. (2018a). Paternally inherited cis-regulatory structural variants are associated with autism. Science (80-.). 360, 327–331.

Braude, P., Bolton, V., and Moore, S. (1988). Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. Nature 332, 459–461.

Brewer, M.H., Chaudhry, R., Qi, J., Kidambi, A., Drew, A.P., Menezes, M.P., Ryan, M.M., Farrar, M.A., Mowat, D., Subramanian, G.M., et al. (2016). Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3. PLoS Genet. 12, 1–16.

Bunyan, D.J., Robinson, D.O., Tyers, A.G., Huang, S., Maloney, V.K., Grand, F.H., Ennis, S., Silva, S.R. De, Crolla, J.A., and McMullan, T.F.W. (2014). X-Linked Dominant Congenital Ptosis Cosegregating with an Interstitial Insertion of a Chromosome 1p21.3 Fragment into a Quasipalindromic Sequence in Xq27.1. Open J. Genet. 04, 415–425.

Busslinger, G.A., Stocsits, R.R., Van Der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature 544, 503–507.

Cai, C., Langfelder, P., Fuller, T.F., Oldham, M.C., Luo, R., van den Berg, L.H., Ophoff, R.A., and Horvath, S. (2010). Is human blood a good surrogate for brain tissue in transcriptional studies? BMC Genomics 11, 589.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M., Osborne, C., et al. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 17, 127.

Carbone, L., and Chavez, S.L. (2015). Mammalian pre-implantation chromosomal instability: species comparison, evolutionary considerations, and pathological correlations. Syst. Biol. Reprod. Med.

Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 17, 224–238.

Chaisson, M.J.P., Wilson, R.K., and Eichler, E.E. (2015a). Genetic variation and the de novo assembly of human genomes. Nat. Rev. Genet. 16, 627–640.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015b). Resolving the complexity of the human genome using single-molecule sequencing. Nature 517, 608–611.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. 10, 1784.

Chang, H.H.Y., Pannunzio, N.R., Adachi, N., and Lieber, M.R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat. Rev. Mol. Cell Biol. 18, 495–506.

Chavez, S.L., Loewke, K.E., Han, J., Moussavi, F., Colls, P., Munne, S., Behr, B., and Reijo Pera, R. a. (2012). Dynamic blastomere behaviour reflects human embryo ploidy by the four-cell stage. Nat. Commun. 3, 1251.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2015). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. 106, 19096–19101.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A.Y., Boutros, P., Chen, J., et al. (2016). NovoBreak: Local assembly for breakpoint detection in cancer genomes. Nat. Methods 14, 65–67.

Colaco, S., and Sakkas, D. (2018). Paternal factors contributing to embryo quality. J. Assist. Reprod. Genet. 35, 1953–1968.

Collins, A.R., Oscoz, A.A., Brunborg, G., Gaivão, I., Giovannelli, L., Kruszewski, M., Smith, C.C., and Stetina, R. (2008). The comet assay: topical issues. Mutagenesis 23, 143–151.

Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Khera, A. V, Francioli, L.C., Gauthier, L.D., Wang, H., Watts, N.A., et al. (2019). An open resource of structural variation for medical and population genetics. BioRxiv 578674.

Conlin, L.K., Thiel, B.D., Bonnemann, C.G., Medne, L., Ernst, L.M., Zackai, E.H., Deardorff, M.A., Krantz, I.D., Hakonarson, H., and Spinner, N.B. (2010). Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Hum. Mol. Genet. 19, 1263–1275.

Coonen, E., Derhaag, J.G., Dumoulin, J.C.M., van Wissen, L.C.P., Bras, M., Janssen, M., Evers, J.L.H., and Geraedts, J.P.M. (2004). Anaphase lagging mainly explains chromosomal mosaicism in human preimplantation embryos. Hum. Reprod. 19, 316–324.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12, 628–640.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J. a, Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., et al. (2011). A copy number variation morbidity map of developmental delay. Nat. Genet. 43, 838–846.

Costes, S.V., Chiolo, I., Pluth, J.M., Barcellos-Hoff, M.H., and Jakob, B. (2010). Spatiotemporal characterization of ionizing radiation induced DNA damage foci and their relation to chromatin organization. Mutat. Res. Mutat. Res. 704, 78–87.

Crasta, K., Ganem, N.J., Dagher, R., Lantermann, A.B., Ivanova, E. V, Pan, Y., Nezi, L., Protopopov, A., Chowdhury, D., and Pellman, D. (2012). DNA breaks and chromosome pulverization from errors in mitosis. Nature 482, 53–58.

Cretu Stancu, M., Van Roosmalen, M.J., Renkens, I., Nieboer, M.M., Middelkamp, S., De Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., et al. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat. Commun. 8, 1–13.

Dali, R., and Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res. 45, 2994–3005.

Dathe, K., Kjaer, K.W., Brehm, A., Meinecke, P., Nürnberg, P., Neto, J.C., Brunoni, D., Tommerup, N., Ott, C.E., Klopocki, E., et al. (2009). Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type A2. Am. J. Hum. Genet. 84, 483–492.

Daughtry, B.L., Rosenkrantz, J.L., Lazar, N.H., Fei, S.S., Redmayne, N., Torkenczy, K.A., Adey, A., Yan, M., Gao, L., Park, B., et al. (2019). Single-cell sequencing of primate preimplantation embryos reveals chromosome elimination via cellular fragmentation and blastomere exclusion. Genome Res. 29, 367–382.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. 34, 518–524.

DECIPHER (2019). Database Statistics. https://decipher.sanger.ac.uk/about#stats (Accessed 31-03-2019)

Dekker, J., Marti-Renom, M. a, and Mirny, L. a (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 14, 390–403.

Delhanty, J.D., Harper, J.C., Ao, A., Handyside, A.H., and Winston, R.M. (1997). Multicolour FISH detects frequent chromosomal mosaicism and chaotic division in normal preimplantation embryos from fertile patients. Hum. Genet. 99, 755–760.

DeStefano, G.M., Fantauzzo, K. a, Petukhova, L., Kurban, M., Tadin-Strapps, M., Levy, B., Warburton, D., Cirulli, E.T., Han, Y., Sun, X., et al. (2013). Position effect on FGF13 associated with X-linked

congenital generalized hypertrichosis. Proc. Natl. Acad. Sci. U. S. A. 110, 7790–7795.

Destouni, A., and Vermeesch, J.R. (2017). How can zygotes segregate entire parental genomes into distinct blastomeres? The zygote metaphase revisited. BioEssays 39, 1–7.

Destouni, A., Esteki, M.Z., Catteeuw, M., Tšuiko, O., Dimitriadou, E., Smits, K., Kurg, A., Salumets, A., Van Soom, A., Voet, T., et al. (2016). Zygotes segregate entire parental genomes in distinct blastomere lineages causing cleavage-stage chimerism and mixoploidy. Genome Res. 26, 567–578.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Dorboz, I., Dumay-Odelot, H., Boussaid, K., Bouyacoub, Y., Barreau, P., Samaan, S., Jmel, H., Eymard-Pierre, E., Cances, C., Bar, C., et al. (2018). Mutation in POLR3K causes hypomyelinating leukodystrophy and abnormal ribosomal RNA regulation. Neurol. Genet. 4, e289.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat. Protoc. 4, 1184–1191.

van Echten-Arends, J., Mastenbroek, S., Sikkema-Raddatz, B., Korevaar, J.C., Heineman, M.J., van der Veen, F., and Repping, S. (2011). Chromosomal mosaicism in human preimplantation embryos: a systematic review. Hum. Reprod. Updat. 17, 620–627.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nat. Rev. Genet. 18, 599–612.

El-Brolosy, M.A., Kontarakis, Z., Rossi, A., Kuenne, C., Günther, S., Fukuda, N., Kikhi, K., Boezio, G.L.M., Takacs, C.M., Lai, S., et al. (2019). Genetic compensation triggered by mutant mRNA degradation. Nature 568, 193–197.

Engwerda, A., Frentz, B., den Ouden, A.L., Flapper, B.C.T., Swertz, M.A., Gerkes, E.H., Plantinga, M., Dijkhuizen, T., and van Ravenswaaij-Arts, C.M.A. (2018). The phenotypic spectrum of proximal 6q deletions based on a large cohort derived from social media and literature reports. Eur. J. Hum. Genet. 26, 1478–1489.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet.

EURORDIS (2019). What is a rare disease? https://www.eurordis.org/content/what-rare-disease (Accessed 30-03-2019)

Evenson, D.P., Darzynkiewicz, Z., and Melamed, M.R. (1980). Relation of mammalian sperm chromatin heterogeneity to fertility. Science (80-.). 210, 1131–1133.

Fabre, P.J., Leleu, M., Mormann, B.H., Lopez-Delisle, L., Noordermeer, D., Beccari, L., and Duboule, D. (2017). Large scale genomic reorganization of topological domains at the HoxD locus. Genome Biol. 18, 1–15.

Falconer, E., and Lansdorp, P.M. (2013). Strand-seq: a unifying tool for studies of chromosome segregation. Semin. Cell Dev. Biol. 24, 643–652.

Falconer, E., Hills, M., Naumann, U., Poon, S.S.S., Chavez, E. a, Sanders, A.D., Zhao, Y., Hirst, M.,

and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods 9, 1107–1112.

Fatehi, A.N. (2006). DNA Damage in Bovine Sperm Does Not Block Fertilization and Early Embryonic Development But Induces Apoptosis After the First Cleavages. J. Androl. 27, 176–188.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. 7, 85–97.

Firth, H. V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. Van, Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. 84, 524–533.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature 538, 265–269.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773.

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S.W., Cairns, J., Collier, A.J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C.S., Fraser, P., et al. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. Elife 6, 1–26.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 15, 2038–2049.

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 377-390.e19.

Gawecka, J.E., Marh, J., Ortega, M., Yamauchi, Y., Ward, M. a, and Ward, W.S. (2013). Mouse zygotes respond to severe sperm DNA damage by delaying paternal DNA replication and embryonic development. PLoS One 8, e56385.

Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. Nature 511, 344–347.

Goldfeder, R.L., Priest, J.R., Zook, J.M., Grove, M.E., Waggott, D., Wheeler, M.T., Salit, M., and Ashley, E.A. (2016). Medical implications of technical accuracy in genome sequencing. Genome Med. 8, 24.

González-Marín, C., Gosálvez, J., and Roy, R. (2012). Types, causes, detection and repair of DNA fragmentation in animal and human sperm cells. Int. J. Mol. Sci. 13, 14026–14052.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.

Graf, A., Krebs, S., Heininen-Brown, M., Zakhartchenko, V., Blum, H., and Wolf, E. (2014). Genome activation in bovine embryos: Review of the literature and new insights from RNA sequencing experiments. Anim. Reprod. Sci. 149, 46–58.

Griffin, J. (2013). Methods of Sperm DNA Extraction for Genetic and Epigenetic Studies. In Methods Mol. Biol., pp. 379–384.

Guan, P., and Sung, W.-K. (2016). Structural variation detection using next-generation sequencing data. Methods 102, 36–49.

Guttmacher, A.E., and Collins, F.S. (2003). Welcome to the genomic era. N. Engl. J. Med. 349, 996–998.

Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yáñez-Cuna, J.O., Amendola, M., van Ruiten,

M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., et al. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 169, 693-707.e14.

Haines, B., Hughes, J., Corbett, M., Shaw, M., Innes, J., Patel, L., Gecz, J., Clayton-Smith, J., and Thomas, P. (2015). Interchromosomal insertional translocation at Xq26.3 alters SOX3 expression in an individual with XX male sex reversal. J. Clin. Endocrinol. Metab. 100, E815–E820.

Hartley, T., Balcı, T.B., Rojas, S.K., Eaton, A., Canada, C., Dyment, D.A., and Boycott, K.M. (2018). The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM®. Am. J. Med. Genet. Part C Semin. Med. Genet. 178, 458–463.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat. Rev. Genet. 10, 551–564.

Hatch, E.M., Fischer, A.H., Deerinck, T.J., and Hetzer, M.W. (2013). Catastrophic nuclear envelope collapse in cancer cell micronuclei. Cell 154, 47–60.

Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8.

Hehir-Kwa, J., Pfundt, R., Veltman, J., and de Leeuw, N. (2013). Pathogenic or not? Assessing the clinical relevance of copy number variants. Clin. Genet. 84, 415–421.

Hehir-Kwa, J.Y., Rodriguez-Santiago, B., Vissers, L.E., de Leeuw, N., Pfundt, R., Buitelaar, J.K., Perez-Jurado, L.A., and Veltman, J.A. (2011). De novo copy number variants associated with intellectual disability have a paternal origin and age bias. J. Med. Genet. 48, 776–778.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. 16, 144–154.

Her, J., and Bunting, S.F. (2018). How cells ensure correct repair of DNA double-strand breaks. J. Biol. Chem. 293, 10502–10511.

Hochstenbach, R., Buizer-Voskamp, J.E., Vorstman, J.A.S., and Ophoff, R.A. (2011). Genome arrays for the detection of copy number variations in idiopathic mental retardation, idiopathic generalized epilepsy and neuropsychiatric disorders: Lessons for diagnostic workflow and research. Cytogenet. Genome Res. 135, 174–202.

Holland, A.J., and Cleveland, D.W. (2012). Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. Nat. Med. 18, 1630–1638.

Ibn-Salem, J., Köhler, S., Love, M.I., Chung, H.-R., Huang, N., Hurles, M.E., Haendel, M., Washington, N.L., Smedley, D., Mungall, C.J., et al. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. 15, 423.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

Iqbal, K., Jin, S.-G., Pfeifer, G.P., and Szabó, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc. Natl. Acad. Sci. U. S. A. 108, 3642–3647.

Janssen, A., van der Burg, M., Szuhai, K., Kops, G.J.P.L., and Medema, R.H. (2011). Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. Science 333, 1895–1898.

Jarvis, G.E. (2016). Estimating limits for natural human embryo mortality. F1000Research 5, 2083.

Javierre, B.M., Sewitz, S., Cairns, J., Wingett, S.W., V??rnai, C., Thiecke, M.J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O.S., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384.e19.

Kaiser-Rogers, K.A. (2005). Androgenetic/biparental mosaicism causes placental mesenchymal dysplasia. J. Med. Genet. 43, 187–192.

Kajii, T., and Ohama, K. (1977). Androgenetic origin of hydatidiform mole. Nature 268, 633–634.

Kalatova, B., Jesenska, R., Hlinka, D., and Dudas, M. (2015). Tripolar mitosis in human cells and embryos: Occurrence, pathophysiology and medical implications. Acta Histochem. 117, 111–125.

Kaminsky, E.B., Kaul, V., Paschall, J., Church, D.M., Bunke, B., Kunig, D., Moreno-De-Luca, D., Moreno-De-Luca, A., Mulle, J.G., Warren, S.T., et al. (2011). An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet. Med. 13, 777–784.

Karczewski, K.J., and Snyder, M.P. (2018). Integrative omics for health and disease. Nat. Rev. Genet. 19, 299–310.

Kaye, J. (2011). The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. Annu. Rev. Genomics Hum. Genet. 13, 415–431.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. 111, 6131–6138.

Kiessling, A.A., Bletsa, R., Desmarais, B., Mara, C., Kallianidis, K., and Loutradis, D. (2009). Evidence that human blastomere cleavage is under unique cell cycle control. J. Assist. Reprod. Genet. 26, 187–195.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature 509, 575–581.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. 20, 207–220.

Kloosterman, W.P., and Cuppen, E. (2013). Chromothripsis in congenital disorders and cancer: similarities and differences. Curr. Opin. Cell Biol. 25, 341–348.

Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-kwa, J.Y., Abdellaoui, A., Lameijer, E., Moed, M.H., Koval, V., Renkens, I., et al. (2015). Characteristics of de novo structural changes in the human genome. Genome Res. 25, 792–801.

Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. Am. J. Hum. Genet. 85, 457–464.

Köhler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., et al. (2017). The human phenotype ontology in 2017. Nucleic Acids Res. 45, D865–D876.

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 47, D1018–D1027.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426.

Kostina, A., Bjork, H., Ignatieva, E., Irtyuga, O., Uspensky, V., Semenova, D., Maleki, S., Tomilin, A., Moiseeva, O., Franco-Cereceda, A., et al. (2018). Notch, BMP and WNT/β-catenin network is impaired in endothelial cells of the patients with thoracic aortic aneurysm. Atheroscler. Suppl. 35, e6–e13.

Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella,

C., Bianco, S., Chiariello, A.M., Jerković, I., et al. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. 50.

Krijger, P.H.L., and de Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. 17, 771–782.

Kurtas, N.E., Xumerle, L., Leonardelli, L., Delledonne, M., Brusco, A., Chrzanowska, K., Schinzel, A., Larizza, D., Guerneri, S., Natacci, F., et al. (2019). Small supernumerary marker chromosomes: A legacy of trisomy rescue? Hum. Mutat. 40, 193–200.

van de Laar, I., Rabelink, G., Hochstenbach, R., Tuerlings, J., Hoogeboom, J., and Giltay, J. (2002). Diploid/triploid mosaicism in dysmorphic patients. Clin. Genet. 62, 376–382.

Lai, W.K.M., and Pugh, B.F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat. Rev. Mol. Cell Biol. 18, 548–562.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell 172, 650–665.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Larson, D.E., Abel, H.J., Chiang, C., Badve, A., Das, I., Eldred, J.M., Layer, R.M., and Hall, I.M. (2018). Svtools : Population-Scale Analysis of Structural Variation. BioRxiv 1000.

Laugsch, M., Bartusel, M., Rehimi, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., Zentis, P., Nikolic, M., Bleckwehl, T., Kolovos, P., et al. (2019). Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs. Cell Stem Cell 24, 736-752. e12.

Lebedev, I.N., Ostroverkhova, N. V, Nikitina, T. V, Sukhanova, N.N., and Nazarenko, S.A. (2004). Features of chromosomal abnormalities in spontaneous abortion cell culture failures detected by interphase FISH analysis. Eur. J. Hum. Genet. 12, 513–520.

Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291.

Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A., and Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum. Mutat. 36, 815–822.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332.

Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu. Rev. Biochem. 79, 181–211.

Lieberman-Aiden, E., Berkum, V., L, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (80-.). 326, 289–293.

Liu, P., Erez, A., Nagamani, S.C.S., Dhar, S.U., Kołodziejska, K.E., Dharmadhikari, A. V, Cooper, M.L., Wiszniewska, J., Zhang, F., Withers, M. a, et al. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 146, 889–903.

Liu, P., Carvalho, C.M.B., Hastings, P.J., and Lupski, J.R. (2012). Mechanisms for recurrent and

complex human genomic rearrangements. Curr. Opin. Genet. Dev. 22, 211–220.

Liu, P., Yuan, B., Carvalho, C.M.B., Wuster, A., Walter, K., Zhang, L., Gambin, T., Chong, Z., Campbell, I.M., Coban Akdemir, Z., et al. (2017). An Organismal CNV Mutator Phenotype Restricted to Early Human Development. Cell 168, 830-842.e7.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell 167, 1170–1187.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21.

Luo, R., Sedlazeck, F.J., Lam, T., and Schatz, M.C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat. Commun. 10, 998.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell 161, 1012–1025.

Ly, P., Teitz, L.S., Kim, D.H., Shoshani, O., Skaletsky, H., Fachinetti, D., Page, D.C., and Cleveland, D.W. (2016). Selective Y centromere inactivation triggers chromosome shattering in micronuclei and repair by non-homologous end joining. Nat. Cell Biol. 1.

Ly, P., Brunner, S.F., Shoshani, O., Kim, D.H., Lan, W., Pyntikova, T., Flanagan, A.M., Behjati, S., Page, D.C., Campbell, P.J., et al. (2019). Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. Nat. Genet. 1.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids Res. 42, 986–992.

Macklon, N.S., Geraedts, J.P.M., and Fauser, B.C.J.M. (2002). Conception to ongoing pregnancy: the "black box" of early pregnancy loss. Hum. Reprod. Update 8, 333–343.

Macklon, N.S., Geraedts, J.P.M., and Fauser, B.C.J.M. (2009). Conception to ongoing pregnancy: the "black box" of early pregnancy loss. Hum. Reprod. Update 8, 333–343.

Maiato, H., and Logarinho, E. (2014). Mitotic spindle multipolarity without centrosome amplification. Nat. Cell Biol. 16, 386–394.

Makrydimas, G. (2002). Complete hydatidiform mole and normal live birth: a novel case of confined placental mosaicism: Case report. Hum. Reprod. 17, 2459–2463.

Makrydimas, G., Sebire, N.J., Thornton, S.E., Zagorianakou, N., Lolis, D., and Fisher, R.A. (2002). Complete hydatidiform mole and normal live birth: a novel case of confined placental mosaicism: case report. Hum. Reprod. 17, 2459–2463.

Mantikou, E., Wong, K.M., Repping, S., and Mastenbroek, S. (2012). Molecular origin of mitotic aneuploidies in preimplantation embryos. Biochim. Biophys. Acta - Mol. Basis Dis. 1822, 1921–1930.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. Genome Res. 230946.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A. 74, 560–564.

McCoy, R.C. (2017). Mosaicism in Preimplantation Human Embryos: When Chromosomal Abnormalities Are the Norm. Trends Genet. 33, 448–463.

McCoy, R.C., Demko, Z., Ryan, A., Banjevic, M., Hill, M., Sigurjonsson, S., Rabinowitz, M., Fraser, H.B., and Petrov, D.A. (2015). Common variants spanning PLK4 are associated with mitotic-origin aneuploidy in human embryos. Science 348, 235–238.

McCoy, R.C., Newnham, L.J., Ottolini, C.S., Hoffmann, E.R., Chatzimeletiou, K., Cornejo, O.E., Zhan, Q., Zaninovic, N., Rosenwaks, Z., Petrov, D.A., et al. (2018). Tripolar chromosome segregation drives the association between maternal genotype at variants spanning PLK4 and aneuploidy in human preimplantation embryos. Hum. Mol. Genet. 27, 2573–2585.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438.

Mehrjouy, M.M., Fonseca, A.C.S., Ehmke, N., Paskulin, G., Novelli, A., Benedicenti, F., Mencarelli, M.A., Renieri, A., Busa, T., Missirian, C., et al. (2017). Regulatory variants of FOXG1 in the context of its topological domain organisation. Eur. J. Hum. Genet.

Mehta, A., and Haber, J.E. (2014). Sources of DNA double-strand breaks and models of recombinational DNA repair. Cold Spring Harb. Perspect. Biol. 6, a016428.

Meienberg, J., Bruggmann, R., Oexle, K., and Matyas, G. (2016). Clinical sequencing: is WGS the better WES? Hum. Genet. 135, 359–362.

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. Am. J. Hum. Genet. 86, 749–764.

Mitter, D., Delle Chiaie, B., Lüdecke, H.J., Gillessen-Kaesbach, G., Bohring, A., Kohlhase, J., Caliebe, A., Siebert, R., Röpke, A., Ramos-Arroyo, M.A., et al. (2010). Genotype-phenotype correlation in eight new patients with a deletion encompassing 2q31.1. Am. J. Med. Genet. Part A 152, 1213–1224.

Montavon, T., Thevenet, L., and Duboule, D. (2012). Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. Proc. Natl. Acad. Sci. U. S. A. 109, 20204–20211.

Mudge, J.M., and Harrow, J. (2016). The state of play in higher eukaryote gene annotation. Nat. Rev. Genet. 17, 758–772.

Munné, S., Blazek, J., Large, M., Martinez-Ortiz, P.A., Nisson, H., Liu, E., Tarozzi, N., Borini, A., Becker, A., Zhang, J., et al. (2017). Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing. Fertil. Steril. 108, 62-71.e8.

Nagaoka, S.I., Hassold, T.J., and Hunt, P. a. (2012). Human aneuploidy: mechanisms and new insights into an age-old problem. Nat. Rev. Genet. 13, 493–504.

Nazaryan-Petersen, L., Eisfeldt, J., Pettersson, M., Lundin, J., Nilsson, D., Wincent, J., Lieden, A., Lovmar, L., Ottosson, J., Gacic, J., et al. (2018). Replicative and non-replicative mechanisms in the formation of clustered CNVs are indicated by whole genome characterization. PLoS Genet. 1–25.

Nelson, B.J., Audano, P.A., Wilson, R.K., Magrini, V., Dougherty, M.L., Welch, A.E., Sorensen, M., Shah, A., Eichler, E.E., McGrath, S.D., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell 176, 663-675.e19.

Newman, S., Hermetz, K.E., Weckselblatt, B., and Rudd, M.K. (2015). Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. Am. J. Hum. Genet. 96, 208–220.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276.

Nguyen, Q.H., Lukowski, S.W., Chiu, H.S., Senabouth, A., Bruxner, T.J.C., Christ, A.N., Palpant, N.J., and Powell, J.E. (2018). Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res. 28, 1053–1066.

Nowakowska, B. (2017). Clinical interpretation of copy number variants in the human genome. J. Appl. Genet. 58, 449–457.

Osterwalder, M., Barozzi, I., Tissiéres, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239–243.

Palmer, N., and Kaldis, P. (2016). Regulation of the Embryonic Cell Cycle During Mammalian Preimplantation Development. Curr. Top. Dev. Biol. 120, 1–53.

Pan, B., Kusko, R., Xiao, W., Zheng, Y., Liu, Z., Xiao, C., Sakkiah, S., Guo, W., Gong, P., Zhang, C., et al. (2019). Similarities and differences between variants called with human reference genome HG19 or HG38. BMC Bioinformatics 20.

Park, H., Chun, S.M., Shim, J., Oh, J.H., Cho, E.J., Hwang, H.S., Lee, J.Y., Kim, D., Jang, S.J., Nam, S.J., et al. (2019). Detection of chromosome structural variation by targeted next-generation sequencing and a deep learning application. Sci. Rep. 9, 3644.

Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res. 27, 665–676.

Pellestor, F. (2014). Chromothripsis: How does such a catastrophic event impact human reproduction? Hum. Reprod. 29, 388–393.

Pellestor, F., Anahory, T., Lefort, G., Puechberty, J., Liehr, T., Hédon, B., and Sarda, P. (2011). Complex chromosomal rearrangements: Origin and meiotic behavior. Hum. Reprod. Update 17, 476–494.

Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 17, 157.

Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. 9.

La Piana, R., Cayami, F.K., Tran, L.T., Guerrero, K., Van Spaendonk, R., Õunap, K., Pajusalu, S., Haack, T., Wassmer, E., Timmann, D., et al. (2016). Diffuse hypomyelination is not obligate for POLR3-related disorders. Neurology 86, 1622–1626.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., MacDonald, J.R., Mills, R., et al. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 29, 512–520.

Pollard, S., Sun, S., and Regier, D.A. (2019). Balancing uncertainty with patient autonomy in precision medicine. Nat. Rev. Genet. 1.

Poplin, R., Chang, P., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987.

Porubský, D., Sanders, A.D., van Wietmarschen, N., Falconer, E., Hills, M., Spierings, D.C.J., Bevova, M.R., Guryev, V., and Lansdorp, P.M. (2016). Direct chromosome-length haplotyping by single-cell sequencing. Genome Res. 26, 1565–1574.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent

chain-terminating dideoxynucleotides. Science 238, 336–341.

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al. (2015). Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 1–11.

Rao, S.S.P.S.P., Huntley, M.H.H., Durand, N.C.C., Stamenova, E.K.K., Bochkov, I.D.D., Robinson, J.T.T., Sanborn, A.L.L., Machol, I., Omer, A.D.D., Lander, E.S.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, 333–339.

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. 49, 36–45.

Reichmann, J., Nijmeijer, B., Hossain, M.J., Eguren, M., Schneider, I., Politi, A.Z., Roberti, M.J., Hufnagel, L., Hiiragi, T., and Ellenberg, J. (2018). Dual-spindle formation in zygotes keeps parental genomes apart in early mammalian embryos. Science (80-.). 361, 189–193.

Richards, M.R., Plummer, L., Chan, Y.-M., Lippincott, M.F., Quinton, R., Kumanov, P., and Seminara, S.B. (2017). Phenotypic spectrum of POLR3B mutations: isolated hypogonadotropic hypogonadism without neurological or dental anomalies. J. Med. Genet. 54, 19–25.

Riehmer, V., Erger, F., Herkenrath, P., Seland, S., Jackels, M., Wiater, A., Heller, R., Beck, B.B., and Netzer, C. (2017). A heritable microduplication encompassing TBL1XR1 causes a genomic sister-disorder for the 3q26.32 microdeletion syndrome. Am. J. Med. Genet. Part A 173, 2132–2138.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317–329.

Robinson, W.P., Lauzon, J.L., Innes, A.M., Lim, K., Arsovska, S., and McFadden, D.E. (2007). Origin and outcome of pregnancies affected by androgenetic/biparental chimerism. Hum. Reprod. 22, 1114–1122.

Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P.J., Beccari, L., El-Idrissi, I., Nguyen Huynh, T.H., Ozadam, H., Dekker, J., and Duboule, D. (2017). The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. Genes Dev. 31, 2264–2281.

Rohrback, S., Siddoway, B., Liu, C.S., and Chun, J. (2018). Genomic mosaicism in the developing and adult brain. Dev. Neurobiol. 78, 1026–1048.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. 13, 1.

Rubin, A.J., Barajas, B.C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M.R., Howard, I., Kim, D.S., Boxer, L.D., Cairns, J., Spivakov, M., et al. (2017). Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nat. Genet. 49, 1522–1528.

Saitsu, H., Osaka, H., Sasaki, M., Takanashi, J.-I., Hamada, K., Yamashita, A., Shibayama, H., Shiina, M., Kondo, Y., Nishiyama, K., et al. (2011). Mutations in POLR3A and POLR3B Encoding RNA Polymerase III Subunits Cause an Autosomal-Recessive Hypomyelinating Leukoencephalopathy. Am. J. Hum. Genet. 89, 644–651.

Sakkas, D., and Alvarez, J.G. (2010). Sperm DNA fragmentation: mechanisms of origin, impact on reproductive outcome, and analysis. Fertil. Steril. 93, 1027–1036.

Salmon, L.B., Orenstein, N., Markus-Bustani, K., Ruhrman-Shahar, N., Kilim, Y., Magal, N., Hubshman, M.W., and Bazak, L. (2018). Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. Genet. Med. 0, 1–9.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.

Schmitt, A.D., Hu, M., and Ren, B. (2016a). Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol. 17, 743–755.

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016b). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 17, 2042–2059.

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864.

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19, 329–346.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728–1732.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: Past, present and future. Nature 550.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. Nat. Rev. Genet. 15, 272–286.

Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H. V, FitzPatrick, D.R., Barrett, J.C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature 555, 611–616.

Shuman, C., Mercimek-Andrews, S., Costain, G., Snell, M., Sondheimer, N., Mendoza-Londono, R., Bowdin, S., Stavropoulos, D.J., Cohn, R.D., Scherer, S.W., et al. (2018). Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. Eur. J. Hum. Genet. 26, 740–744.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679.

Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum. Mutat. 36, 928–930.

Soshnev, A.A., Josefowicz, S.Z., and Allis, C.D. (2016). Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. Mol. Cell 62, 681–694.

Speicher, M.R., and Carter, N.P. (2005). The new cytogenetics: Blurring the boundaries with molecular biology. Nat. Rev. Genet. 6, 782–792.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. 7, 85–97.

Spinella, F., Fiorentino, F., Biricik, A., Bono, S., Ruberti, A., Cotroneo, E., Baldi, M., Cursio, E., Minasi, M.G., and Greco, E. (2018). Extent of chromosomal mosaicism influences the clinical outcome of in vitro fertilization treatments. Fertil. Steril. 109, 77–83.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: From enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626.

Stark, Z., Dolman, L., Manolio, T.A., Ozenberger, B., Hill, S.L., Caulfied, M.J., Levy, Y., Glazer, D., Wilson, J., Lawler, M., et al. (2019). Integrating Genomics into Healthcare : A Global Responsibility. Am. J. Hum. Genet. 104, 13–20.

Stavropoulos, D.J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R.K.C., Szego, M.J., et al. (2016). Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ Genomic Med. 1.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L. a, et al. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell 144, 27–40.

Stessman, H.A., Bernier, R., and Eichler, E.E. (2014). A genotype-first approach to defining the subtypes of a complex disease. Cell 156, 872–877.

Stessman, H.A.F., Turner, T.N., and Eichler, E.E. (2016). Molecular subtyping and improved treatment of neurodevelopmental disease. Genome Med. 8, 1–9.

Strain, L., Warner, J.P., Johnston, T., and Bonthron, D.T. (1995). A human parthenogenetic chimaera. Nat. Genet. 11, 164–169.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81.

Svensson, A.M., Curry, C.J., South, S.T., Whitby, H., Maxwell, T.M., Aston, E., Fisher, J., Carmack, C.E., Scheffer, A., Abu-Shamsieh, A., et al. (2007). Detection of a de novo interstitial 2q microdeletion by CGH microarray analysis in a patient with limb malformations, microcephaly and mental retardation. Am. J. Med. Genet. A 143A, 1348–1353.

Syeda, A.H., Hawkins, M., and McGlynn, P. (2014). Recombination and Replication. Cold Spring Harb. Perspect. Biol. 6, 1–14.

Szenker-Ravi, E., Altunoglu, U., Leushacke, M., Bosso-Lefèvre, C., Khatoo, M., Thi Tran, H., Naert, T., Noelanders, R., Hajamohideen, A., Beneteau, C., et al. (2018). RSPO2 inhibition of RNF43 and ZNRF3 governs limb development independently of LGR4/5/6. Nature 557, 564–569.

Tan, T.Y., Gonzaga-Jauregui, C., Bhoj, E.J., Strauss, K.A., Brigatti, K., Puffenberger, E., Li, D., Xie, L.Q., Das, N., Skubas, I., et al. (2017). Monoallelic BMP2 Variants Predicted to Result in Haploinsufficiency Cause Craniofacial, Skeletal, and Cardiac Features Overlapping Those of 20p12 Deletions. Am. J. Hum. Genet. 101, 985–994.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034.

Taylor, T.H., Gitlin, S.A., Patrick, J.L., Crain, J.L., Wilson, J.M., and Griffin, D.K. (2014). The origin, mechanisms, incidence and clinical consequences of chromosomal mosaicism in humans. Hum. Reprod. Update 20, 571–581.

Tétreault, M., Choquet, K., Orcesi, S., Tonduti, D., Balottin, U., Teichmann, M., Fribourg, S., Schiffmann, R., Brais, B., Vanderver, A., et al. (2011). Recessive Mutations in POLR3B, Encoding the Second Largest Subunit of Pol III, Cause a Rare Hypomyelinating Leukodystrophy. Am. J. Hum. Genet. 89, 652–655.

The Global Alliance for Genomics and Health (2016). A federated ecosystem for sharing genomic, clinical data. Science (80-.). 352, 1278–1280.

Thiffault, I., Wolf, N.I., Forget, D., Guerrero, K., Tran, L.T., Choquet, K., Lavallée-Adam, M., Poitras, C., Brais, B., Yoon, G., et al. (2015). Recessive mutations in POLR1C cause a leukodystrophy by impairing biogenesis of RNA polymerase III. Nat. Commun. 6, 7623.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C.,

Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature 489, 75–82.

Toyoshima, M. (2009). Analysis of p53 dependent damage response in sperm-irradiated mouse embryos. J. Radiat. Res. 50, 11–17.

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W.L., Pereira, S.L., Whitney, J., Chan, A.J.S., Pellecchia, G., et al. (2018). A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. Am. J. Hum. Genet. 102, 142–155.

Tšuiko, O., Catteeuw, M., Zamani Esteki, M., Destouni, A., Bogado Pascottini, O., Besenfelder, U., Havlicek, V., Smits, K., Kurg, A., Salumets, A., et al. (2017). Genome stability of bovine in vivo-conceived cleavage-stage embryos is higher compared to in vitro-produced embryos. Hum. Reprod. 32, 2348–2357.

Tylee, D.S., Kawaguchi, D.M., and Glatt, S.J. (2013). On the outside, looking in: A review and evaluation of the comparability of blood and brain "-omes." Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 162, 595–603.

Vanneste, E., Voet, T., Le Caignec, C., Ampe, M., Konings, P., Melotte, C., Debrock, S., Amyere, M., Vikkula, M., Schuit, F., et al. (2009). Chromosome instability is common in human cleavage-stage embryos. Nat. Med. 15, 577–583.

Vázquez-Diez, C., and Fitzharris, G. (2018). Causes and consequences of chromosome segregation error in preimplantation embryos. Reproduction 155, R63–R76.

Vázquez-Diez, C., Yamagata, K., Trivedi, S., Haverfield, J., and FitzHarris, G. (2016). Micronucleus formation causes perpetual unilateral chromosome inheritance in mouse embryos. Proc. Natl. Acad. Sci. 113, 626–631.

Veltman, J. a, and Brunner, H.G. (2012). De novo mutations in human genetic disease. Nat. Rev. Genet. 13, 565–575.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science 291, 1304–1351.

Verpoest, W., Fauser, B.C., Papanikolaou, E., Staessen, C., Van Landuyt, L., Donoso, P., Tournaye, H., Liebaers, I., and Devroey, P. (2008). Chromosomal aneuploidy in embryos conceived with unstimulated cycle IVF. Hum. Reprod. 23, 2369–2371.

Vissers, L.E.L.M., and Stankiewicz, P. (2012). Microdeletion and microduplication syndromes. Methods Mol. Biol. 838, 29–75.

Vissers, L.E.L.M., Gilissen, C., and Veltman, J.A. (2015). Genetic studies in intellectual disability and related disorders. Nat. Rev. Genet. 1–10.

Voet, T., Vanneste, E., and Vermeesch, J.R. (2011). The human cleavage stage embryo is a cradle of chromosomal rearrangements. Cytogenet. Genome Res. 133, 160–168.

Vorsanova, S.G., Kolotii, A.D., Iourov, I.Y., Monakhov, V. V, Kirillova, E.A., Soloviev, I. V, and Yurov, Y.B. (2005). Evidence for High Frequency of Chromosomal Mosaicism in Spontaneous Abortions Revealed by Interphase FISH Analysis. J. Histochem. Cytochem. 53, 375–380.

Wainberg, M., Merico, D., Delong, A., and Frey, B.J. (2018). Deep learning in biomedicine. Nat. Biotechnol. 36, 829–838.

Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. 28, 581–591.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science (80-.). 362, eaat8464.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Watson, J.D., and Crick, F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 173, 737–738.

Watson, C.T., Tomas, M.-B., Sharp, A.J., and Mefford, H.C. (2014). The Genetics of Microdeletion and Microduplication Syndromes: An Update. Annu. Rev. Genomics Hum. Genet. 15, 215–244.

Weaver, D.T., Fisher, R.A., Newlands, E.S., and Paradinas, F.J. (2000). Amniotic tissue in complete hydatidiform moles can be androgenetic. J. Pathol. 191, 67–70.

Weber, J.L., and Myers, E.W. (1997). Human Whole-Genome Shotgun Sequencing. Genome Res. 7, 401–409.

Webster, E., Cho, M.T., Alexander, N., Desai, S., Naidu, S., Bekheirnia, M.R., Lewis, A., Retterer, K., Juusola, J., and Chung, W.K. (2016). De novo PHIP -predicted deleterious variants are associated with developmental delay, intellectual disability, obesity, and dysmorphic features . Mol. Case Stud. 2, a001172.

Weintraub, A.S., Li, C.H., Zamudio, A. V, Sigova, A.A., Hannet, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2018). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell 1–16.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. 14, 125–138.

Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. Genet. Med. 19, 209–214.

Wetterstrand, K. (2019). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P., and Cotney, J. (2018). High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. Cell Rep. 23, 1581–1597.

Will, A.J., Cova, G., Osterwalder, M., Chan, W.-L., Wittler, L., Brieske, N., Heinrich, V., de Villartay, J.-P., Vingron, M., Klopocki, E., et al. (2017). Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). Nat. Genet. 49, 1539–1545.

Wolf, N.I., Vanderver, A., van Spaendonk, R.M.L., Schiffmann, R., Brais, B., Bugiani, M., Sistermans, E., Catsman-Berrevoets, C., Kros, J.M., Pinto, P.S., et al. (2014). Clinical spectrum of 4H leukodystrophy caused by POLR3A and POLR3B mutations. Neurology 83, 1898–1905.

Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 538, 523–527.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science 318, 1108–1113.

Wright, C.F., Hurles, M.E., and Firth, H. V. (2016). Principle of proportionality in genomic data sharing. Nat. Rev. Genet. 17, 1–2.

Wright, C.F., FitzPatrick, D.R., and Firth, H. V. (2018a). Paediatric genomics: diagnosing rare disease in children. Nat. Rev. Genet. 19, 253–268.

Wright, C.F., McRae, J.F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T.W., Jones, P., Prigmore, E., Rajan, D., Lord, J., et al. (2018b). Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. Genet. Med. 20, 1216–1223.

Yardımcı, G.G., Ozadam, H., Sauria, M.E.G., Ursu, O., Yan, K.K., Yang, T., Chakraborty, A., Kaul, A.,

Lajoie, B.R., Song, F., et al. (2019). Measuring the reproducibility and quality of Hi-C data. Genome Biol. 20, 1–19.

Yauy, K., Gatinois, V., Guignard, T., Sati, S., Puechberty, J., Gaillard, J.B., Schneider, A., and Pellestor, F. (2018). Looking for Broken TAD Boundaries and Changes on DNA Interactions: Clinical Guide to 3D Chromatin Change Analysis in Complex Chromosomal Rearrangements and Chromothripsis BT - Chromothripsis: Methods and Protocols. F. Pellestor, ed. (New York, NY: Springer New York), pp. 353–361.

Zappala, Z., and Montgomery, S.B. (2016). Non-Coding Loss-of-Function Variation in Human Genomes. Hum. Hered. 81, 78–87.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. 16, 172–183.

Zentner, G.E., and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. Nat. Rev. Genet. 15, 814–827.

Zepeda-Mendoza, C.J., and Morton, C.C. (2019). The Iceberg under Water: Unexplored Complexity of Chromoanagenesis in Congenital Disorders. Am. J. Hum. Genet. 104, 565–577.

Zepeda-Mendoza, C.J., Ibn-Salem, J., Kammin, T., Harris, D.J., Rita, D., Gripp, K.W., MacKenzie, J.J., Gropman, A., Graham, B., Shaheen, R., et al. (2017). Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. Am. J. Hum. Genet. 1–12.

Zhan, Q., Ye, Z., Clarke, R., Rosenwaks, Z., and Zaninovic, N. (2016). Direct Unequal Cleavages: Embryo Developmental Competence, Genetic Constitution and Clinical Outcome. PLoS One 11, e0166398.

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., and Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. Nature 522, 179–184.

Zhao, J., Zhang, Q., Wang, Y., and Li, Y. (2014). Whether sperm deoxyribonucleic acid fragmentation has an effect on pregnancy and miscarriage after in vitro fertilization/intracytoplasmic sperm injection: a systematic review and meta-analysis. Fertil. Steril. 102, 998-1005.e8.

Zhu, H., Shang, D., Sun, M., Choi, S., Liu, Q., Hao, J., Figuera, L.E., Zhang, F., Choy, K.W., Ao, Y., et al. (2011). X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. Am. J. Hum. Genet. 88, 819–826.

Zini, A., and Sigman, M. (2009). Are Tests of Sperm DNA Damage Clinically Useful? Pros and Cons. J. Androl. 30, 219–229.

Zook, J., McDaniel, J., Parikh, H., Heaton, H., Irvine, S.A., Trigg, L., Truty, R., McLean, C.Y., Vega, F.M.D. La, Xiao, C., et al. (2018). Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. BioRxiv 281006.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. 32, 246–251.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. 51, 12–18.

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 19, 217.

Samenvatting

Tijdens de bevruchting waarbij een eicel en een spermacel samensmelten wordt een deel van het DNA van de moeder en de vader overgeërfd. Deze combinatie van het geërfde DNA van de vader en de moeder vormt het genoom, de complete DNA-code bestaande uit twee keer 3.2 miljard basenparen verdeeld over 46 chromosomen, van een individu. Het DNA bevat alle informatie voor de ontwikkeling van een enkele cel in een complex organisme bestaande uit biljoenen cellen. Humane genomen bestaan voor het grootste deel uit dezelfde code, maar (vrijwel) elk genoom is uniek. De verschillen tussen genomen, ook wel varianten genoemd, vormen de basis van veel van onze persoonlijke kenmerken. Genetische varianten worden overgeërfd van de ouders, maar er kunnen ook nieuwe ("*de novo*") varianten ontstaan in de ouderlijke geslachtcellen of in cellen van vroege embryo's die niet aanwezig zijn in de genomen van de ouders. Sommige van deze *de novo* varianten kunnen ertoe leiden dat de embryonale ontwikkeling niet goed verloopt, waardoor aangeboren aandoeningen kunnen ontstaan. Er zijn verschillende soorten variaties in de DNA-code: van variaties die een enkel basepaar veranderen tot grote structurele variaties (SVs) die miljoenen basenparen kunnen aandoen. Deleties, duplicaties, inserties, inversies en translocaties zijn verschillende typen structurele variaties, welke soms ook in complexe combinaties voor kunnen komen. In dit proefschrift zijn de oorzaken en gevolgen van complexe structurele variaties onderzocht die *de novo* zijn ontstaan in geslachtscellen van een ouder of in de vroege embryonale ontwikkeling.

De volgordes van DNA-codes en de varianten daarin kunnen worden ontrafeld met behulp van DNA sequencing technologie. In **hoofdstuk 1** worden de geschiedenis, de ontwikkeling en de impact van DNA sequencing besproken. De DNA sequencing technologie heeft zich razendsnel ontwikkeld in de laatste decennia. In 2001 werd voor het eerst een nagenoeg compleet humaan genoom gesequencet, wat destijds nog een enorme technische en financiële uitdaging was. Tegenwoordig kan een humaan genoom in een aantal dagen gesequencet en geanalyseerd worden voor een prijs van ongeveer 1000 euro. Deze indrukwekkende ontwikkelingen hebben ertoe geleid dat DNA sequencing steeds meer wordt toegepast in de wetenschap en de gezondheidszorg. Hierdoor is er veel kennis opgedaan over de rol van specifieke DNA-varianten in het veroorzaken van ziekten. Tegenwoordig liggen de grootste uitdagingen niet meer in het sequencen zelf, maar vooral in de interpretatie van de gigantische hoeveelheden sequencing data die worden geproduceerd.

Vaak is het onbekend hoe DNA-varianten zijn ontstaan en wat hun gevolgen zijn.

In ongeveer de helft van de patiënten met aangeboren aandoeningen (zoals een verstandelijke beperkingen of autisme) kan tegenwoordig een genetische oorzaak worden gevonden in het DNA. De meeste subchromosomale *de novo* varianten ontstaan op de chromosomen die worden overgeërfd van de vader. Beschadigingen van het DNA van spermacellen, die in tegenstelling tot eicellen constant worden aangemaakt tijdens de vruchtbare levensfase, komen vaak voor tijdens de ontwikkeling en het verplaatsen van de cellen. In **hoofdstuk 2** hebben we de effecten van deze DNA-schade in spermacellen op het genoom van twee- en achtcellige embryo's bepaald. Met behulp van single-cell genoom sequencing hebben we aangetoond dat sperma DNA-schade kan leiden tot genomische instabiliteit en grootschalige genetische afwijkingen in vroege embryo's. Onze resultaten suggereren dat de hoge sterfte van vroege humane embryo's (meer dan de helft van de bevruchtingen leidt niet tot een succesvolle zwangerschap) deels verklaard kan worden door de gevolgen van DNA-schade die al in de spermacellen ontstaat.

In **hoofdstuk 3** hebben we de gevolgen van complexe *de novo* structurele varianten onderzocht bij patiënten met aangeboren aandoeningen. Varianten in het DNA kunnen ervoor zorgen dat de functie en/of activiteit van specifieke genen (welke voor eiwitten, essentiële bouwstenen en machines van de cellen, coderen) veranderen. Structurele varianten kunnen verschillende effecten op nabij gelegen genen hebben. Ze kunnen direct de DNA-sequentie van genen veranderen en zelfs de hele code van een gen verwijderen of dupliceren. Ook kunnen structurele varianten indirect nabijgelegen genen aandoen, waarbij ze niet de code van de genen zelf veranderen, maar wel de DNA-elementen die belangrijk zijn voor de regulatie van gen activiteit. Mensen hebben rond de 20,000 verschillende genen, maar deze moeten niet in elke cel op elk moment actief zijn (je wilt bijvoorbeeld (simpel gezegd) niet dat een gen dat voor botaanmaak zorgt actief is in je hersencellen). De expressie van genen wordt daarom strikt gereguleerd door complexe regulatiemechanismen. Mede dankzij de ontwikkeling van nieuwe sequencing technieken is er de laatste jaren veel geleerd over deze regulatiemechanismen en over de indirecte effecten van genetische varianten op genen (ook wel positionele effecten genoemd). Meestal wordt bepaald of een specifieke variant een aandoening kan hebben veroorzaakt door te onderzoeken of er ook andere patiënten met eenzelfde variant én een vergelijkbaar ziektebeeld zijn. Vaak is er echter niet bekend welke moleculaire mechanismen tot de ziekte hebben geleid. In hoofdstuk 3 hebben we de volledige genomen van 39 patiënten met eerder gevonden *de novo* SVs in kaart gebracht om te kijken of we nieuwe varianten en effecten konden vinden die eerder niet waren gedetecteerd met andere technieken die in de reguliere genetische diagnostiek worden toegepast. In zeven gevallen ontdekten we met behulp van whole genome sequencing dat de SVs complexer waren dan eerder bepaald kon worden. Vervolgens hebben we een bio-informatische aanpak

ontwikkeld om effecten van SVs op genen die mogelijk betrokken zijn bij de ziekte te kunnen voorspellen door gebruik te maken van eerder gegenereerde data. Met onze methode hebben we meer inzichten gekregen in de moleculaire mechanismen die ten grondslag liggen aan het ziektebeeld van 16 van de 39 bestudeerde patiënten.

De indirecte, positionele effecten van SVs op omliggende genen zijn vaak specifiek voor bepaalde celtypen. Het kan bijvoorbeeld zijn dat een SV geen verstorend effect heeft in bloedcellen, maar wel in specifieke hersencellen. Daarom kunnen positionele effecten van SV het beste worden bepaald in cellen die relevant zijn voor het ziektebeeld. Zulke cellen zijn meestal echter niet beschikbaar, omdat ze moeilijk toegankelijk zijn (zoals hersencellen) of omdat de effecten alleen optreden tijdens de embryonale ontwikkeling. Om toch de effecten van SVs in relevante celtypen te bestuderen hebben we daarom in **hoofdstuk 4** bloedcellen van een patiënt getransformeerd naar zogenaamde "induced pluripotent stem cells" (iPS cellen). Deze iPS cellen, welke het genoom met de complexe genetische afwijkingen van de patiënt bevatten, zijn vergelijkbaar met stamcellen van jonge embryo's en ze kunnen gebruikt worden als een model voor de vroege ontwikkeling. De iPS cellen kunnen, net als vroege embryonale stamcellen, veranderen ("differentiëren") in andere celtypen. De patiënt heeft complexe *de novo* structurele varianten waarbij meerdere chromosomen zijn betrokken (veroorzaakt door een proces wat "chromothripsis" heet). Uit een eerdere studie welke gebruik maakte van bloedcellen was niet precies duidelijk geworden welke genen precies zijn aangedaan en het ziektebeeld kunnen verklaren. We hebben de iPS cellen van de patiënt en de ouders gedifferentieerd naar neurale celtypen. Vervolgens hebben we de gen-activiteit (RNA-expressie) en de 3D-organisatie van het genoom van de cellen van de patiënt vergeleken met de cellen van de gezonde ouders. Met deze aanpak ontdekten we dat de expressie van het *TWIST1* gen, welke in de buurt van de structurele varianten lag en niet direct was aangedaan, specifiek was verstoord in de neurale cellen. Deze verstoring was niet detecteerbaar in bloedcellen van de patiënt, wat de toegevoegde waarde van het gebruik van iPS cellen aantoont voor de detectie van de moleculaire gevolgen van structurele varianten. De verstoring van de regulatie van het *TWIST1* gen heeft waarschijnlijk een groot deel van het ziektebeeld van de patiënt veroorzaakt.

Van meer dan duizend genen is al bekend dat specifieke genetische varianten in deze genen aangeboren aandoeningen kunnen veroorzaken. In **hoofdstuk 5** beschrijven we onze ontdekking van zeldzame varianten in het *POLR3GL* gen in drie patiënten met endostale hyperostosis (een botafwijking), oligodontie (ontbreken van een aantal tanden), een klein gestalte en milde gelaatskenmerken. Het eiwit dat vanuit de code van dit gen geproduceerd wordt, is onderdeel van het RNA-polymerase III complex, wat essentiële functies uitvoert in cellen. Varianten in andere onderdelen

van dit complex waren al eerder geassocieerd met een spectrum aan aangeboren aandoeningen, maar varianten specifiek in het *POLR3GL* gen waren nog niet eerder aan ontwikkelingsstoornissen gelinkt. De gevonden mutaties zijn zogenaamd homozygoot of compound recessief, wat betekent dat beide kopieën van het gen (de kopie overgeërfd van de vader en de kopie overgeërfd van de moeder) zijn aangedaan en de patiënten dus geen normaal *POLR3GL* RNA meer tot expressie brengen. Dit konden we aantonen door onder andere het RNA van bloedcellen van de patiënten te sequencen. Nu het duidelijk is dat zeldzame varianten in *POLR3GL* een specifiek ziektebeeld kunnen veroorzaken kan dit gen ook in de andere patiënten worden bestudeerd, wat een bijdrage kan leveren aan het verbeteren van de genetische diagnostiek.

Dankwoord

En dan het gedeelte waar iedereen met smacht naar heeft uitgekeken. Allereerst wil ik graag mijn promotor **Edwin** bedanken. Bedankt dat je me zonder al te veel toezicht in het lab hebt durven rond laten lopen. De vrijheid heeft zeker bijgedragen aan mijn ontwikkeling naar een zelfstandige wetenschapper. Het was erg fijn dat je altijd de knopen doorhakte als ik weer eens twijfelde over iets. Ik denk dat jij en je groep het perfecte voorbeeld zijn van hoe je hard werken met plezier kunt combineren!

Ook wil ik uiteraard de leescommissie bestaande uit **Alexander van Oudenaarden**, **Elzo de Wit**, **Geert Kops**, **Hans Kristian Ploos van Amstel** en **Nine Knoers** bedanken voor het doornemen en (gelukkig positief) beoordelen van mijn proefschrift. Daarnaast wil ik Alexander en Geert bedanken voor deelnemen aan mijn PhD begeleidingscommissie en het kritisch en waardevol monitoren van mijn PhD traject.

Ik wil mijn paranimfen bedanken voor hun onmisbare bijdrage aan de totstandkoming van dit proefschrift. **Ewart** bedankt voor de prettige samenwerking en alle begeleiding, gefundeerd op gortdroge humor en fervente pogingen om toch nog goede muziek op het lab te houden. Heb veel van je opgestoken en je enthousiasme voor de biologie heeft altijd erg aanstekelijk gewerkt! **Judith** bedankt voor alle samenwerkingen gecombineerd met oer-Hollandse nuchterheid. Het was niet altijd een makkelijk project, maar we hebben er zeker wat moois van gemaakt! Niet te veel naar 90's Now feesten gaan en nog veel succes met de afronding van je PhD traject en het onderwijs traineeship!

Ik wil alle leden van de Cuppen groep bedanken voor al het moois in deze 4,5 jaren. **Myrthe** en **Francis** bedankt voor alle gezelligheid in ons kantoor. Jullie hebben me zeker het goede voorbeeld gegeven! **Roel** zonder jou en je Guix had het proefschrift er waarschijnlijk een stuk dunner uitgezien. Bedankt voor al je hulp bij al onze computer vragen! **Monique**, bedankt voor het uitzoeken en regelen van alles! **Sander**, fijn dat het je toch best aardig is gelukt om met een broek aan op het werk te komen de laatste 4,5 jaren. Naast dit essentiële punt wil ik je ook wel bedanken voor het runnen van de pipelines en het managen van onze data! **Lisanne** en **Nicolle**, de massa-pipetteerders van de Cuppen groep, bedankt voor al jullie hulp met onze experimenten en de gezelligheid op en buiten het lab. Ook nog veel succes met de paarden en de vogels! **Robin**, bedankt dat ik de loodzware en verantwoordelijke taak van biermanager met een gerust hart aan je heb kunnen overdragen. Succes met het hooghouden van het promillage in de Cuppen groep de komende jaren! **Arne**, geniet

van je drie kleinmannen, een goed excuus om extra veel Samson en Gert te kijken! **Sharon**, alvast veel plezier gewenst met het kopen van baby-cadeaus voor Arne's aankomende aanwinst en veel succes met je onderzoeksprojecten (niet te veel 5-FU binnenkrijgen)! **Don Luan**, ik weet dat je dit stiekem wel kunt lezen, succes met al je digitale en analoge Chords. **Barisha** (B-a-s-t-i-a-a-n), veel succes met Lewis leren zijn naam juist te schrijven en ook nog veel succes met al je modellen! **Ies**, bedankt voor al je bio-informatica adviezen! Rustig aan met het gaspedaal en kijk uit voor de tram (als ie eens gaat rijden).

Daarnaast wil ik ook de alle oud-Cuppen members en aanhang bedanken voor al het werkplezier de laatste jaren: **Anna** (leuk dat ik je nu weer tegenkom!), **Annelies** (heb veel over trouwjurken geleerd), **Esther** (ik weet je te vinden als ik ooit nog een yoga-guru zoek), **Ewart de B.** (staat de B voor bier?), **Jerome** (bedankt voor de nagenoeg foutloze METC-aanvraag), **Joep** (the Force is strong in this one), **Pim** (met jou, subtiele ratten- en Joepknuffelaar, een biertje drinken is altijd een feest!), **Mark V.** (je bent nog niet van me af!), **Nico** (the man with a dog who is called the dog officially known as prince), **Pjotr** (bedankt voor je bijdrage aan mijn digitalisering), **Ruben** (fijn dat je me van de straat af hebt gehouden) en **Terry** (ik weet nog steeds niet wat je precies doet, maar bedankt voor de gezelligheid en je unieke eigenzinnige en waardevolle kijk op de genetica). **Martin** en **Robert** bedankt voor onze heldhaftige en epische avonturen in het hooggebergte van Italië. Dankzij jullie ben ik gitaarmuziek nóg meer gaan waarderen! Ook erg bedankt voor het ontwikkelen en onderhouden van de onmisbare IAP. Ook wil ik graag mijn stagestudenten **Joris**, **Laura** en **Imke** bedanken voor jullie inzet en nuttige bijdrages aan mijn projecten. Ik heb zelf veel geleerd van het begeleiden van jullie en ik hoop dat jullie er ook iets van hebben opgestoken!

Ik wil graag alle mensen van "beneden" uit de groepen van **Wigard Kloosterboer**, **Jeroen de Ridder**, **Bo(o)bby Koeleman**, **Gijs van Haaften** en **Nine Knoers** bedanken voor alle Chinese Highways, spaghettitorens, bierproeverijen, culinaire hoogstandjes, barbecues, popcorn uit de magnetron, geheime drankvoorraden, illegale bevoorradingen en alle gezelligheid tijdens de retreats en borrels. Jullie zijn inmiddels met te veel om op te noemen, maar ik doe toch een poging (klachten kunnen bij paranimfen terecht, een extra biertje voor eenieder die ik heb gemist): **Alessio, Alexandra, Amin, Anukrati, Chris, Cristina, Daiane, Edith, Ellen C, Ellen S, Flip, Glen, Helen, Iris, Joanna (x2), Joline, Jose, Joske, Karen, Kirsten, Luca, Marijn, Marleen, Mircea, Nayia, Roy, Rozemarijn, Ruben, Sanne, Tilman** en **Wout**. **Wigard** bedankt voor het ontdekken van chromothripsis in de germline en je input op onze projecten. Veel succes met je nieuwe uitdaging! **Mark(us)** bedankt voor alle filtering scripts en je hulp met de analyses, zal er de komende tijd nog veel gebruik van gaan maken. **Ivo**, bedankt voor de essentiële levenslessen (aka biertips). Ook nog bedankt

voor je hulp met de experimenten in het lab, al zullen de zebravissen daar vast anders over hebben gedacht.

Ook wil iedereen van de **2e en 3e verdieping van het Straatje** bedanken voor alle leuke borrels. Mede dankzij jullie is onze overgang naar het UMC toch erg meegevallen! Alle mensen van de **van Boxtel groep** bedankt voor het warme welkom en het wegwijs maken in het wondere en vaak ook heftige wereld van het Prinses Máxima Centrum. We gaan hopelijk nog veel mooie avonturen beleven!

Uiteraard wil ik ook alle mensen van buiten het Stratenum met wie ik heb samengewerkt bedanken. Allereerst wil ik **Jacques** bedanken voor de samenwerking en voor het wegwijs maken van Judith en mij in de klinische genetica. **Paulien** en **Koen** bedankt voor de samenwerking aan het *POLR3GL* project, het is een mooi paper geworden! **Bernard** en **Leni** bedankt voor alle hulp met de embryo kweken en voor jullie inspirerende passie voor embryo's en interesse in onze projecten. Ook wil ik graag **Diana**, **Victor** en **Peter** van het ERIBA uit het hoge noorden bedanken voor de samenwerking, het single-cell sequencen en jullie enthousiasme voor ons project. **Chloé**, merci for working together and operating the (stupide) robot.

Verder wil ik graag **papa**, **mama**, **Daan**, **Mikki**, **Myriam** en **Patrick** bedanken voor alle steun, interesse en vertrouwen, ook al was het soms lastig te begrijpen wat ik nou de hele dag uitvoer. Ook wil ik graag **King Beastman**, de **Worldpolice** en alle **B&Bers** bedanken voor alle afleiding en gezelligheid naast het werk. Als laatste wil ik mijn geliefde **Ila** bedanken voor alle liefde, steun, toeverlaat en boterhammen. Bedankt voor alle aanmoedigingen en interesse in mijn projecten, ook al moet je daarvoor meestal achter de TV eten.

A

List of publications

Middelkamp S*, Vlaar JM*, Giltay J, Korzelius J, Besselink N, Boymans S, Janssen R, de la Fonteijne L, van Binsbergen E, van Roosmalen MJ, Hochstenbach R, Giachino D, Talkowski ME, Kloosterman WP, Cuppen E**. Prioritization of genes driving congenital phenotypes of patients with *de novo* genomic structural variants. *bioRxiv* (2019)

Middelkamp S, van Tol HTA, Spierings DCJ, Boymans S, Guryev V, Roelen BAJ, Lansdorp PM, Cuppen E**, Kuijk EW. Sperm DNA damage causes genomic instability in early embryonic development. *bioRxiv* (2019)

Terhal PA*, Vlaar JM*, Middelkamp S*, Nievelstein RAJ, Nikkels PGJ, Ross J, Créton M, Bos JW, Voskuil-Kerkhof ESM, Cuppen E, Knoers N, van Gassen KLI**. Biallelic variants in *POLR3GL* cause endosteal hyperostosis and oligodontia. *Eur. J. Hum. Genet.* (2019)

Kalkan T, Bornelöv S, Mulas C, Diamanti E, Lohoff T, Ralser M, Middelkamp S, Lombard P, Nichols J, Smith A**. Complementary Activity of ETV5, RBPJ, and TCF3 Drives Formative Transition from Naive Pluripotency. *Cell Stem Cell* 24, 785-801.e7. (2019)

Middelkamp S, van Heesch S, Braat AK, de Ligt J, van Iterson M, Simonis M, van Roosmalen MJ, Kelder MJE, Kruisselbrink E, Hochstenbach R, Verbeek NE, Ippel EF, Adolfs Y, Pasterkamp RJ, Kloosterman WP, Kuijk EW**, Cuppen E**. Molecular dissection of germline chromothripsis in a developmental context using patient-derived iPS cells. *Genome Med.* 9, 9 (2017)

Cretu Stancu M*, van Roosmalen MJ*, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP**. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* 8, 1–13 (2017)

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet. 49*, 36–45.

Hupkes M**, van Someren EP, Middelkamp SH, Piek E, van Zoelen EJ, Dechering KJ. DNA methylation restricts spontaneous multi-lineage differentiation of mesenchymal progenitor cells, but is stable during growth factor-induced terminal differentiation. *Biochim. Biophys. Acta* 1813, 839–49 (2011)

** Equal contribution*

*** Corresponding author*

Curriculum vitae

Sjors Harrie Antoon Middelkamp finished his prenatal development in Nijverdal (Hellendoorn, The Netherlands) on the 5th of September 1989. After completing his pre-university education at Het Rhedens Lyceum in Rozendaal, he started his Bachelor studies in Biology at the Radboud University in Nijmegen. He continued his university education by following the Research and the Management & Technology tracks of the Medical Biology Master's programme of the Radboud University. During his studies he performed internships in the labs of Joop van Zoelen (Radboud University) and Gert-Jan Veenstra (Nijmegen Centre for Molecular Life Sciences), focussing on stem cell differentiation and early embryonic development. During his final Master's research internship in the lab of Austin Smith (Cambridge Stem Cell Institute, UK) he became fascinated by next-generation sequencing and bioinformatics. After obtaining his Master of Science degree in 2014, he started his PhD research in the group of Edwin Cuppen, first at the Hubrecht Institute and later at the University Medical Center Utrecht. During his PhD, he combined wet-lab with bioinformatics to study the causes and consequences of *de novo* structural variation. In June 2019 he continued his scientific endeavors as a postdoctoral researcher in the lab of Ruben van Boxtel at the Princess Máxima Center for Pediatric Oncology in Utrecht.