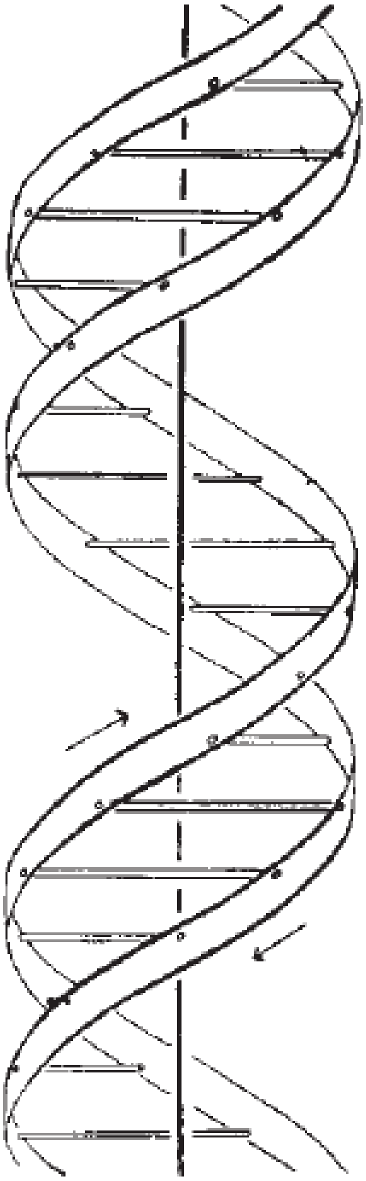
Structural variations in the human genome

The current focus in research on DNA



Myrthe Jager

July 2011

W.P Kloosterman

Department of medical genetics

Research on DNA has evolved from the discovery of the double-helix structure in 1953 to structural variations today. Structural variations are all genomic rearrangements bigger than one base pair. This definition includes deletions, insertions, translocations, inversions, and duplications. Genomic rearrangements can have an influence on phenotype, and are thus associated with diseases. A Structural variation in a somatic cell might change susceptibility to cancer while a *de novo* rearrangement in a germ cell might result in congenital defects. Sequencing the break point can aid in relating the variant to a phenotypic effect and may help identifying a mutational mechanism. Three major mechanisms have currently been suggested. NAHR and NHEJ are double strand DNA break repair mechanisms. FoSTeS (or MMBIR) is a replication-based mechanism. Chromothripsis, retrotransposition, alternative FoSTeS and alternative end-joining (MMEJ) are also suggested mechanisms, resulting in structural variations. Finding and defining both pathogenic and non-pathogenic structural variations is important, since we will then be able to establish the cause for some diseases.

In the project described in this article, the occurrence of four recurrent non-pathogenic deletions in the population was determined. This experiment shows that non-pathogenic rearrangements are quite common in the population. The deletions in chromosomes 1, 5, 22, and the X-chromosome are present in 35% to 93% of the population.

Furthermore, a second experiment was performed in which structural variations of two children with congenital defects were sequenced by capillary sequencing. The goal of this experiment was to identify a possible cause for their abnormalities and to establish which mutational mechanism could have led to the structural variation. No *de novo* mutations were found in one of the patients. Two mutations that he inherited from his mother were caused by MMEJ and retrotransposition. In the other patient, two *de novo* rearrangements were found. Sequencing of one of them failed. The other was a 1.4 Mb tandem duplication, containing five genes and two non-processed pseudogenes, of which the coding sequence was still intact. I conclude that this duplication is caused by FoSTeS. Each of the *de novo* mutations could in theory be the cause for the congenital defects found in the first patient.

Index

Introduction

Deoxyribose nucleic acid or DNA is without a doubt the most fascinating molecule in the entire world, perhaps even in the entire universe. Its massive amount of base pairs consisting of a varying number of genes (per organism) contains hereditary information that is used in the development and functioning of an entire organism. In fact, it is hard to imagine life or living without DNA being involved. The double helix structure that Watson and Crick (figure 1) (1) discovered in the nineteen fifties holds many more mysteries than any other molecule could ever do; mysteries that are in need of elucidation. This is probably what inspires us every day, in our quest of understanding DNA.

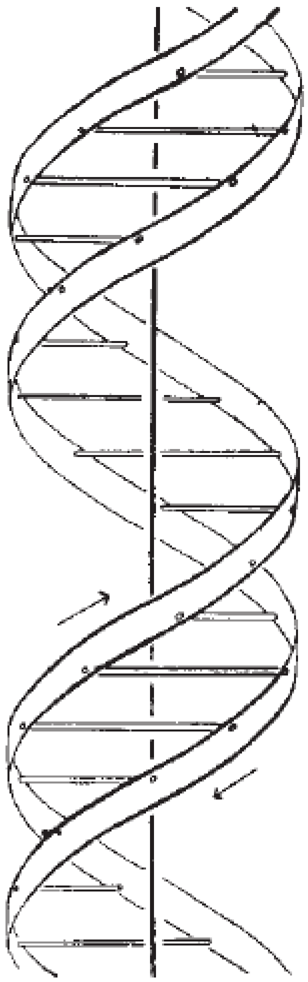
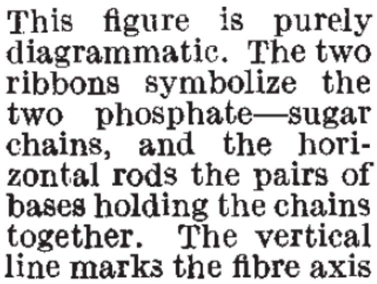


Figure 1: The double-helix structure of DNA that Watson and Crick proposed in 1953. Their words further speak for themselves (1):



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

With every single discovery that has been made, it seems as though ten new questions arise; the most important questions without exception being 'what can we do with this new information' and 'what are the clinical implications of this knowledge'. Answering these (and other) questions is not always easy. For this reason many questions remain unanswered. Even with current newly developed techniques, a seemingly simple fact such as the exact number of genes has (for instance) yet to be determined. In fact, research on DNA can probably keep researchers busy for decades. Perhaps you are wondering why researchers would still spend time on something that can seem so fruitless. There is a reason for everything however. Despite the fact that the amount of mystery that DNA holds is enormous, the fact that DNA contains important information in the development and existence of (almost) every organism (still) lures researchers into doing research on it. Its importance in living and life, its complexity and its mysteries are what make DNA such a fascinating molecule.

Research on DNA has evolved since the discovery of the famous Watson and Crick in 1953. For obvious reasons, it was merely focusing on discovering the function of genes at first: they contain the actual hereditary information. This focus shifted (due to some pressure of the media) later on to the revelation of the 3.2 million base pair sequence that the human genome consists of. In 2003, this sequence was completed (37). Nowadays, the entire genome sequence of almost sixty vertebrate species, thirty metazoan species, sixteen fungi, thirteen protist species, eleven plants, and many (many, many) bacteria has been unraveled (38). Still, understanding the human genome remains the ultimate goal.

The sequence of the human genome differs tremendously among individuals (2,3). These differences range from single nucleotides to gross alterations. All of these alterations can have an impact on human phenotype, like eye color. This impact on phenotype is a result of their ability to interfere with gene function, protein function and even gene expression. In some cases, it can eventually lead to certain (new or heritable) diseases (2,3). So even though the extent to which our genomes differ is not entirely clear yet, the fact that these differences can exist in humans that coexist is very spectacular on itself.

Differences that are quite common in the population (which occur in more than one percent of all individuals) are called polymorphisms. Perhaps the best researched example of a polymorphism is a single nucleotide polymorphism, or SNP (pronounced as 'Snip'). It is currently estimated that there are 10 to 15 million SNPs in the human genome (2). Our knowledge on common patterns of SNPs has increased rapidly over the past few years. Our understanding of variations bigger than one base pair however, is much less pronounced (2). One thing that is clear is that the human genome differs more due to these bigger variations than due to single nucleotide differences (4). Since these variations are not only bigger but also more common in the human genome than single nucleotide differences, their combined impact on the human phenotype (and therefore their association with diseases) might be of significant greater importance (3). Therefore research on DNA is currently paying attention to these variations in DNA, called structural variations (SV) (2).

In this essay I will first explain what structural variations are and what their effects are on disease, and disease susceptibility. Also, some population specific differences in structural variations will be discussed. Next, detection methods of structural variations will briefly be explained. I will then discuss the three main hypotheses on the molecular mechanism of structural variations. Finally, I have conducted two experiments on structural variations myself. It is important to do research on structural variations. Once common SVs are known, it is easier to distinguish pathogenic rearrangements from non-pathogenic rearrangements. Eventually, we will thus be able to develop specific medicines faster. Foremost, we will be one step further in understanding DNA.

Structural variations

Structural variations used to be defined as all genomic rearrangements that are bigger than one thousand base pairs (>1 kb) (4,5). Since our detection techniques have further developed, the current definition can be adjusted to include all variations bigger than 50 base pairs (4). Structural variations in its broadest sense can even simply be defined as all genomic variations in an organisms genome that are bigger than one base pair (2). Several different types of mutations fit these two last definitions: deletions, insertions (novel sequence insertions and mobile-element insertions), inversions, duplications (tandem duplications and interspersed duplications), and translocations (figure 2) (2,6).

The type of rearrangement can be identified by comparing the sequence of someone's DNA sample to the sequence of another DNA sample. Usually, a reference genome is used in this comparison. However, when trying to identify *de novo* rearrangements, the DNA sequence of the parents is used. *De novo* (or new) rearrangements are structural variations that a child has, but the parents of that child do not have. They are often a result of a rearrangement in the paternal chromosome of the germ cell during meiosis (7).

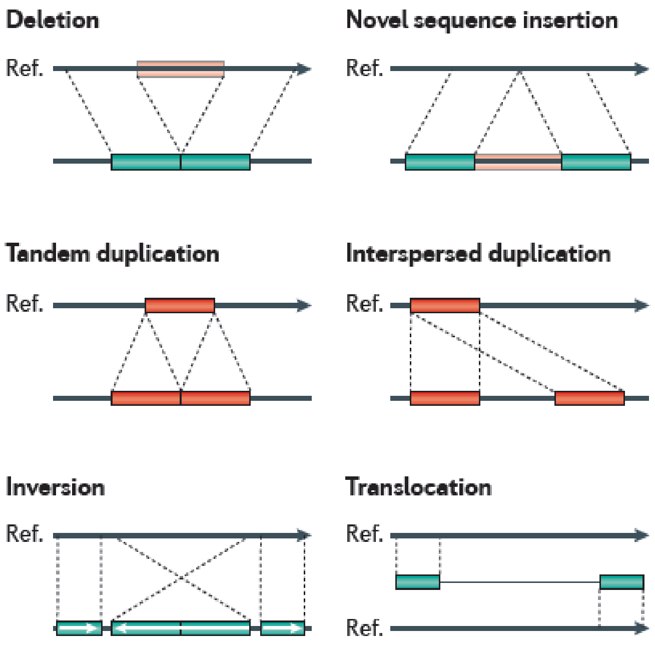


Figure 2: Different types of structural variations. The upper DNA-strand is a strand of the reference genome. When this sequence is compared to the sequence of the (lower) DNA-strand (which is of a different individual), differences between these two can be detected. Then, a structural variation could be found. Deletions, insertions, duplications, inversions and trans-locations are different types of structural variations (4).

Structural variations can be divided into several categories. Firstly, they are either recurrent or non-recurrent. Sometimes, rearrangements occur more often in a certain DNA fragment, due to favorable circumstances. They are therefore present in many individuals. These are recurrent structural variations, meaning that they happen more often. Non-recurrent structural variations on the other hand occur on rare spots in the DNA. Sometimes an individual can even seem to be the only one with a certain structural variation at a certain spot. Secondly, structural variations are either intrachromosomal or interchromosomal. Rearrangements in one chromosome are named intrachromosomal, while rearrangements between two chromosomes are called interchromosomal. Finally, structural variations can either occur in somatic cells or in germ cells. A rearrangement in a somatic cell only affects the organism in which the rearrangement has happened in. A mutation in a germ cell on the other hand will only have effect on the offspring.

Paired-end mapping (PEM) of two individuals, one African and one European, revealed that structural variations vary tremendously in size (figure 3). The majority of the variations are in between 0 and 25 thousand base pairs in size, so relatively small. Approximately 65 percent of all rearrangements are smaller than ten kilo base pairs and only fifteen percent of all structural variations are bigger than a 100 kb (3).

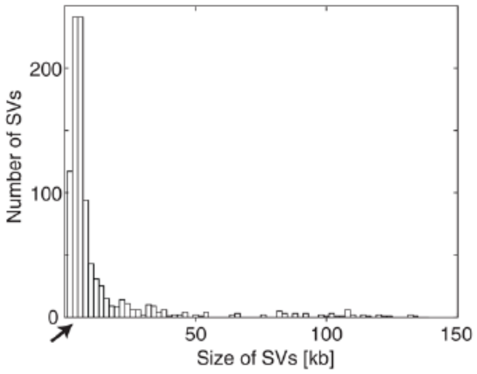


Figure 3: Distribution of different sizes of structural variations. Most SVs are quite small. Approximately thirty percent is five thousand base pairs or smaller in size. 65 percent are smaller than ten thousand base pairs. In contrast: only fifteen percent is larger than 100 kilo base pairs (3).

Effects of structural variations on phenotype

Like all human genomic alterations, structural variations can have an impact on human phenotype by disrupting the 'normal' DNA (if one can even speak of normal DNA). Diseases can be a result of this ability to interfere with gene function, protein function, and gene expression. Rearrangements can either occur in a germ cell or in a somatic cell; the consequences are entirely different. A mutation during meiosis of a germ cell can cause a congenital (and eventually hereditary) disease, while a somatic mutation can contribute to a tumor. Structural variations are thus associated with many different diseases. These range from aniridia to susceptibility to HIV infection to genomic disorders such as the Williams-Beuren syndrome (8,9,10).

Location of mutation

The (severity of the) effects of structural variations on phenotype depend on a combination of the location and the type of structural variation. The location is presumably even the most important factor in defining the consequence, since a mutation in so-called 'junk DNA' might not even have any consequences.

Non-coding DNA

Structural variations can in theory be present in the entire genome, but they are most often present in sequences that do not code for a protein as a result of selective constraint in germ cells (3). In fact, there are people who hypothesize that intercepting mutations that could also happen in coding DNA is the most important function of non-coding DNA. Obviously, the more 'not important' DNA there is, the smaller the chance of a mutation occurring in 'important' DNA. Even in the non-coding sequence however, structural variations can have their influence on the human phenotype in another way than coding for a protein. Two examples illustrate the diversity of the effects of structural variations in non-coding sequences.

Firstly, structural variations can occur in the regulatory sequence of a gene. If the promoter sequence of a certain gene for example changes, gene expression (could) changes as well. A deletion or inversion of (a part of) the regulatory sequence can cause a decrease in gene expression. Insertions can also decrease gene expression when they occur in the promoter. However, when a promoter of an active gene is coincidentally inserted right in front of a relatively inactive gene, an insertion can cause an increase in gene expression. A deletion in the downstream regulatory sequence of *TNFAIP3* is associated with systemic lupus erythematosus (11).

Another example of a change in phenotype due to a rearrangement in the non-coding DNA-sequence is in the non-coding functional RNA, among others: micro-RNA (miRNA). Micro-RNAs are thought to control the activity of approximately 30 percent of all proteins (12). When a structural variation changes a miRNA, the activity of a protein could change as well. Therefore it is no surprise that micro RNAs have been shown to play important roles in different diseases, such as cancer and immune diseases (12). A deletion of the miRNA *Dgcr8* in mice results in defects in the synaptic transmission of the prefrontal cortex, which could give insights in the pathology of human schizophrenia (13).

Coding DNA

Structural variations can also occur in genes, even though there is selective constraint against this in germ cells (figure 4). The effects of these mutations in coding DNA are more obvious (and often worse) than of non-coding DNA. Seventeen percent of all rearrangements for example directly alter gene function (3). The amount of genes affected by a variation obviously increases with an increase in size of the variation. This is especially true for mutations smaller than ten thousand base pairs. Approximately 125 genes are affected by a ten thousand base pair rearrangement (3).

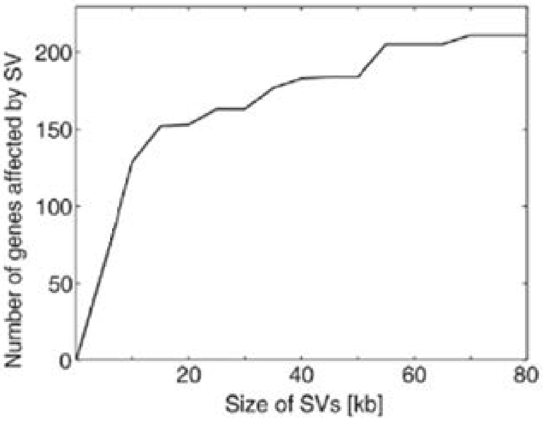


Figure 4: Amount of genes affected by structural variations. The number of genes that are affected by a structural variation grows as the size of the SV grows. One hundred genes are affected by a variation of approximately 8 thousand base pairs. Two hundred genes are affected by a rearrangement of almost 55 kilo base pairs (adjusted from (3)).

Genes can be affected by structural variations in different ways. Firstly, the gene dosage can be altered. When a person has a third 21st chromosome, he or she will suffer from Down syndrome. Secondly, a gene could be disrupted, by for instance an insertion. This would result in a disrupted non-functional protein. Thirdly, genes that are fused together by a rearrangement can form a new functional protein (2). An example of this is the BCR-ABL fusion gene that is caused by a translocation and that is found in leukemia patients (14,15). A fourth mechanism is the alteration of gene expression due to structural variations. Gene expression can for instance be increased when a gene with low transcription activity will translocate to another promoter of a gene with high transcription activity. A final mechanism is the unmasking of recessive mutations (2).

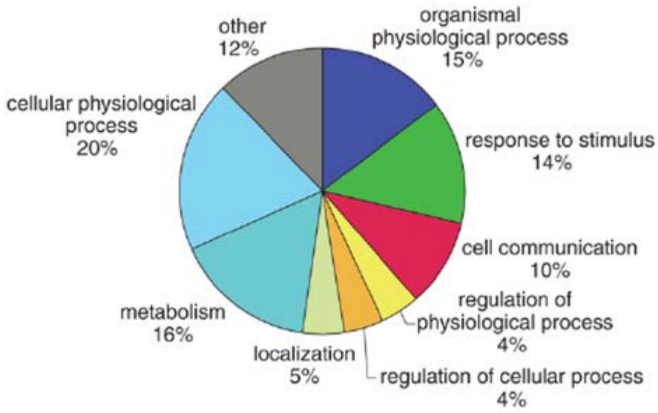


Figure 5: Function of protein coding genes that are affected by rearrangements. Structural variations especially occur in 'cellular physiological process' genes (20%) and 'organismal physiological process' genes (15%). Twelve percent ('other') of gene functions only have less than ten genes that are affected by a structural variation (adjusted from (3)).

Even within coding DNA, there is selective constraint on the type of genes most susceptible to mutations (16). Genes that code for proteins involved in cell adhesion, signal transduction, immunity and defense, and sensory perception are especially prone to mutations (16). In more general terms, most of the mutations are present in genes that code for proteins with a function in cellular physiological processes, organismal physiological processes and metabolism (figure 5). It is important to note that genes that code for proteins that regulate these same processes show only a quarter to a third of the amount of structural variations. So even within genes, it is still true that the most important fragments of the DNA encompass the least structural variations (3).

Structural variations are associated with diseases

The alteration of phenotype by structural variations can contribute to (or even cause) a disease. Rearrangements in somatic cells can lead to cancer, while rearrangements in germ cells can contribute to hereditary diseases. Some of these diseases or disease susceptibilities run through families while others are the result of a *de novo* rearrangement in a germ cell. The only way to elucidate what happened is by finding the (sometimes complex) rearrangements in the patients DNA (17).

De novo

Rearrangements that occur during meiosis in germ cells are associated with many diseases, like for instance: susceptibility to HIV infection, systematic autoimmunity, Williams-Beuren syndrome, Prader-Willi syndrome, velocardiofacial syndrome, color blindness, rhesus blood group sensitivity, classical hemophilia, several forms of beta- and alpha thalassemia, DiGeorge syndrome, and glomerulonephritis (2,3,8,10,16). The risk of a congenital disease is twice as high in children with a *de novo* structural variation as in children without it (17). In many cases, a rearrangement in the paternal DNA is the cause for these diseases .The result is often a child with a complex clinical phenotype, including many congenital abnormalities (7).

Two established models have been composed, to explain the association of structural variations with diseases (4). Firstly, gains or losses (of many base pairs) which are rare in (occur in less than one percent of) the population play a significant role in the cause of the disease. This is the case in many neurocognitive diseases (4). Secondly, changes in gene families will only contribute to a change in the susceptibility for a disease. Examples of the latter are immunity genes and cell-cell signaling genes (4). In conclusion: structural variations can lead to a disease either by altering the disease susceptibility or by causing the disease. I will give an example of both.

A change in susceptibility for colon Crohn disease can be caused by low beta-defensin 2 gene copy number (figure 6). Crohn disease is a chronic inflammatory disease of the bowel, most prominently present in the colon and the ileum. It seems to be a consequence of both genetic and environmental factors, but the exact cause has not been found yet. Several susceptibility genes had already been found, but none of them seem to exclusively lead to Crohn disease. It had previously been suggested that a change in defensin gene expression might contribute to increased disease susceptibility. They protect the bowel from bacteria. On average, a human has four copies of the beta-defensin 2 gene. Patients with colonic Crohn disease however have a slightly lower copy number. This suggests that these patients were already susceptible for colonic Crohn disease (18).

A deletion downstream of PAX6 (paired-box gene 6) at 11p13 is the cause for eye abnormalities such as aniridia. Aniridia is an autosomal dominant hereditary disorder in which the iris of the eye is (complete or partially) absent with in some cases additional hyperplasia of the residual iris. Deletions in the PAX6 gene, a transcription factor of 422 amino acids that is involved in eye development, are a known cause for aniridia. In a large Chinese family however, no mutations were found in the 14 exons of the PAX6 gene itself. A large deletion of 556 kilo base pairs, 123 kilo base pairs downstream of PAX6, was detected in the affected family members, but not in healthy family members. This deletion contains four genes: DCDC1, DNAJC24, IMMP1L, and ELP4. Since little is known about (the function of) these four genes, it is not possible to name them as possible biological candidates for the cause of aniridia. Several previous studies have found a different downstream deletion in aniridia patients. Therefore, it could also be hypothesized that the deletion of remote downstream regulatory elements of the PAX6 gene (of which no further knowledge currently exists) can be a cause for aniridia (9).

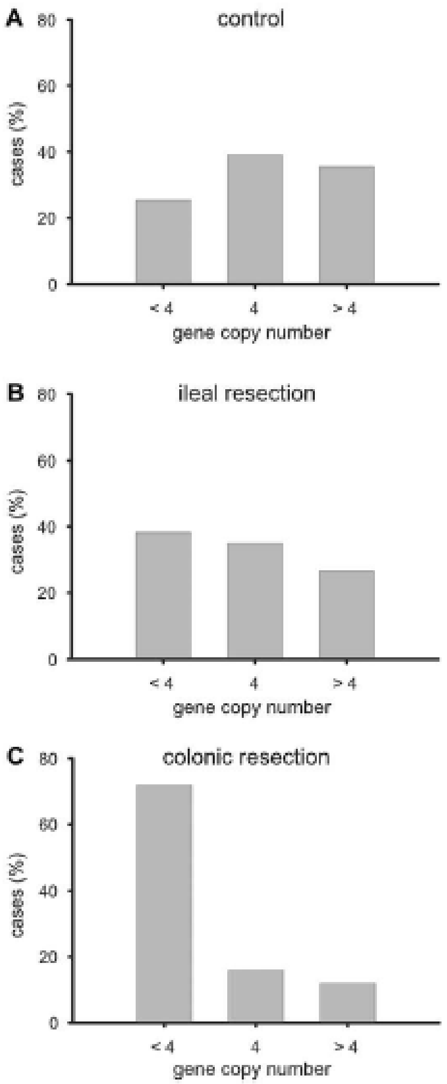


Figure 6: Gene copy number of beta-defensin 2. On average the gene copy number is four. **(A)** This is true for both controls **(B)** and subjects with ileal Crohn disease. **(C)** Subjects with colonic Crohn disease however have a significantly lower average gene copy, suggesting that they are predisposed for this disease (18).

Cancer

Interestingly, genomic rearrangements are not always the cause for a disease but they can also be the cause as well as the consequence of a disease, like in the case of cancer. As we know, cancer is a direct consequence of mutations in the human genome. Many studies have also proven that somatic rearrangements occur in human cancer genomes, from the earliest stages throughout tumor development (14,15). The prevalence of structural variations varies between different cancer types and different patients. In general, epithelial cancers show many structural variations (14). Identification of the type of structural variations in thirteen patients with metastatic pancreatic cancer has proven that even inter-individual variety in type of structural variation is enormous. Not one type of mutation was present in all of the patients (19). In breast cancers, the most frequent rearrangements are intrachromosomal (14). These differences highlight the diversity in structural variations in cancer and their possible contribution to this disease.

Chromosomal rearrangements can be used as a personal biomarker for tumor detection, since they are only present in tumor cells and not in healthy cells. This approach has for instance been shown to be effective in leukemia. In solid tumors however, recurrent rearrangements are not generally present. Response to therapies in solid tumors can be measured by a technique called PARE, or personalized analysis of rearranged ends. Patient specific rearrangements are identified by next-generation sequencing of the resected tumor. These are the new biomarkers for this specific patient. Then, the response to therapies can be measured by measuring the biomarker in bodily fluids. The chance of misdiagnosis can thus be decreased by the PARE-technique (15).

Population-based differences

Structural variations not only have negative effects, but they also seem to have a function. Many deletions for instance (in some cases even encompassing the deletion of entire genes) have been found to be widespread in the genome. Structural variations can thus possibly also play a significant part in genome evolution (16).This might be the cause for the existence of population based differences in structural variations. The UGT2B17 gene for example is associated with ethnic differences in risk of prostate cancer (2,5). Moreover, let us not forget that different populations have different skin colors.

Copy number polymorphisms of five different populations (European Americans, Han Chinese from Beijing , Japanese, Yoruba, and Maasai) have been compared to each other (5). Thirty significant differences in copy numbers involving genes were found, most of them coding for proteins with a function in environmental response. Sixteen of these copy number polymorphisms had not previously been genotyped. These differences could explain for some difference in phenotype and disease susceptibility between populations. East and south East Asian people for example have a significantly lower copy number of a certain DNA fragment that includes a duplication of the last five exons of the OCLN gene than African individuals. The copies of this gene are separated by 1.4 mega base pairs of DNA sequence. The OCLN gene codes for occludin and is associated with a decreased susceptibility to hepatitis C viral infection. African individuals are thus more susceptible to Hepatitis C viral infection than Asian people (5).

Detection and identification

As mentioned, it is very important to detect and identify (non-)pathogenic structural variations. This will enable us to find pathogenic structural variations more quickly, so that we can establish what rearrangement is the cause for a disease. Eventually, we will thus be able to develop specific medicines faster. It is important to know the location of a mutation, in order to identify which sequences have been disrupted by it. A first indication of the location of the structural variation can be given by either FISH or karyotyping. Primers can then be designed for specific regions in the DNA. These primers can be used to sequence the DNA. Whole-genome sequencing can also be performed, but this is a more expensive strategy. CNV (copy number variant) arrays are a cheaper alternative to determine the copy number.

FISH and Karyotyping

FISH (fluorescent in situ hybridization) is a technique in which fluorescent probes are hybridized to a certain DNA-sequence. These can then be seen and analyzed under a fluorescent microscope. The probes are designed to specifically hybridize to the DNA fragment that is being analyzed (37). FISH is for instance used to determine the copy number of a certain gene. A fluorescent probe is designed to hybridize to that specific DNA fragment. If the DNA of a person only shows one fluorescent probe, he or she is missing one copy of the gene. When different genes need to be analyzed at one time, different colors of fluorescent probes can be used. The application in identification of structural variations differs somewhat from this. Probes have been designed to color each of the 23 chromosome pairs in a different color (figure 7). Now, especially interchromosomal rearrangements can be detected.

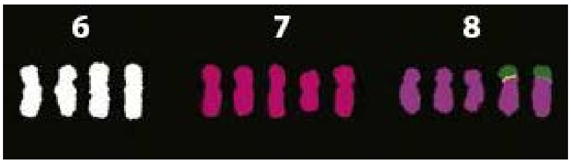


Figure 7: FISH of chromosome 6,7, and 8 of a cell line of renal cancer. Four copies are present of the sixth chromosome. Chromosome 7 shows five copies, one of them slightly shorter (due to a translocation or deletion). The cell line contains five copies of chromosome 8, two of them with translocations of another chromosome (17) (adjusted from (36)).

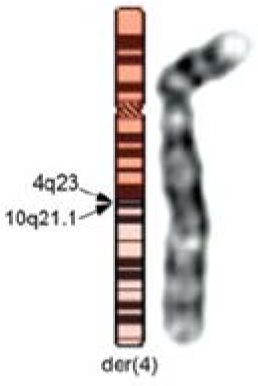


Figure 8: Karyotype of chromosome 4 of a child with congenital defects. The black and white bands were compared to their normal orientation. This led to the conclusion that a part of chromosome 10 has been attached to a part of chromosome 4 (adjusted from (33)).

Karyotyping can give a more exact location of chromosomal rearrangements. The karyotype of a patient needs to be compared to the karyotype of a healthy individual, in order to find structural variations. With Giemsa stain, so called G bands on the DNA are black, while G negative bands are pale (figure 8) (37). The pattern of black and white bands can elucidate what happened. Both interchromosomal rearrangements (like in figure 8) and intrachromosomal rearrangements can be identified by karyotyping. Furthermore, since the location of each black and white band is known, a more precise location of chromosomal rearrangement can be established. It is therefore perhaps a better strategy in quickly determining the location of a structural variation than FISH.

Next-generation sequencing

Sequencing of the structural variation allows us to identify the exact location to the base pair, type, and break point of structural variation. The genome is first broken into random pieces by for example nebulization (17). All DNA-fragments of a certain size are then selected for analysis and amplificated. Next, this DNA fragment library is sequenced (20). The orientation and span of for instance paired-end reads are mapped and the computer assembles the sequence of the analyzed genome (4). Finally, this is compared to the sequence of a reference genome. Differences between the reference genome and the analyzed genome can be explained by SVs.

The high costs and low-throughput of traditional sequencing led to the development of new sequencing techniques, called next-generation sequencing: Illumina/Solexa, SOLiD, and 454 sequencing (table 1). The technique which is chosen depends on read length, accuracy, cost and amount of base pairs per run. The highest accuracy can be obtained by SOLiD. This technique can also perform the most Gb per run. The read length however is very short. In terms of the best read length, 454 sequencing is the best technique to use. And when you want the best of both worlds, Illumina is the most practical, with a medium read length a medium amount of Gb per run (21).

Table 1: Sequencing techniques (21).

Technology	**Read length (bp)**	**Accuracy (%)**	**Gb/run**
Traditional Sanger	~1000	99.99	0.0003
454 sequencing	~450	99.00	0.6
Illumina/Solexa	36-100	98.00-99.00	3-20
SOLiD	35-50	99.94	50-100

454 sequencing was the first technique to become available (21). It is based on the principle of pyro sequencing. DNA fragments are exposed to one nucleotide at the same time. Whenever a nucleotide is incorporated, pyrophosphate is released. This causes a luciferase-driven reaction with as a result the emission of light (22). So when multiple DNA-fragments are exposed to adenine, the wells in which adenine is incorporated will light up. Then another nucleotide is analyzed, etcetera.

The second next-generation technique, Solexa, is based on the principle of reversible terminator chemistry (21). After preparation, the DNA fragments are exposed to reversible terminator nucleotides with a different fluorescent color per nucleotide. The terminator is then removed and the entire process is repeated for 36 to 100 times. Analysis of the order of colors per DNA-fragment can elucidate the DNA-sequence (22).

Finally, SOLiD is based on the principle of sequencing by ligation (21). A mixture of octamers, with a fluorescent color corresponding to the fourth and fifth base, is added to fragments of DNA. The color is analyzed and the octamer is then cleaved between the fifth and sixth base pair, which removes the label. Then, the mixture of octamers is added again. By using different primer lengths, it is possible to identify a fluorescent color corresponding up to the sequencing primer and the first base pair. Since the sequence of the primer is known, the other base pairs can then be identified as well (22).

Capillary sequencing

Refinement of the break points can done by capillary sequencing. The principle is the same as in traditional Sanger sequencing. Chains are elongated from a primer, until a ddNTP (dideoxy nucleotide phosphate) is incorporated, which terminates elongation (37). In capillary sequencing, each ddNTP is labeled with a different fluorescent color (figure 9). Then the DNA-fragments are loaded in a gel sorted according to size. The wavelength of each fragment size is analyzed. This corresponds to the DNA-sequence of the analyzed fragment (37). Approximately 800 base pairs can be sequenced by capillary sequencing.

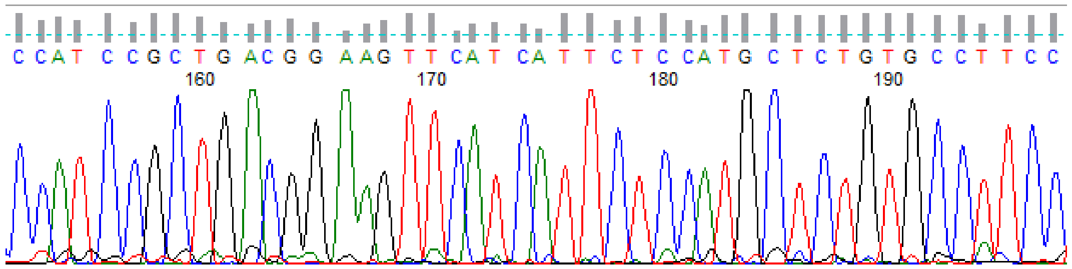


Figure 9: Output of traditional capillary sequencing. Each ddNTP (which terminate DNA elongation) is labeled with a different fluorescent color. Fragments are loaded in a gel and sorted according to size and the color of each size is analyzed. Then the DNA-sequence is determined.

CNV arrays

Copy number variant (CNV) arrays are microarray-based techniques that can detect the amount of copies of a certain DNA-sequence compared to a reference sample (37). In CGH-arrays sample DNA and reference DNA are co hybridized to oligonucleotides or BAC-clones (bacterial artificial chromosome) on the array (4). One sample is labeled with a red dye (Cy3), while the other is labeled with a green dye (Cy5) (37). The log of the ratio will then allow verification of the copy number. The expectation would be a yellow signal, indicating that there is no difference in copy number. Any other color is an indication of a copy number variation. SNP-arrays, another CNV-array, work in a slightly different way. They compare sample DNA and reference DNA as well, but the sample and the reference DNA are not hybridized on the same microarray (4).

CNV arrays can detect CNVs of 500 base pairs and longer, but are the best at detecting CNVs of approximately 1,500 base pairs (4). They are relatively cheap, in comparison to sequencing and can therefore be a good alternative, giving cheaper insights in human disease. They are however not fit to identify high copy numbers or copy numbers in heterochromatin (6). Also, the influence of the reference genome on the outcome is enormous and should therefore not be overlooked (4).

Mutational mechanisms

Even though many structural variations are known and we can identify them by sequencing them, the precise mechanism in which they arise remains a mystery (except perhaps the insertion of a transposable element). Many suggestions have been made to try to elucidate this phenomenon. The knowledge, on which these solutions are based, has mainly been derived from DNA studies in prokaryotes, cell lines and yeast (6). Three major mechanisms that result in structural variations have been suggested so far (6,23).

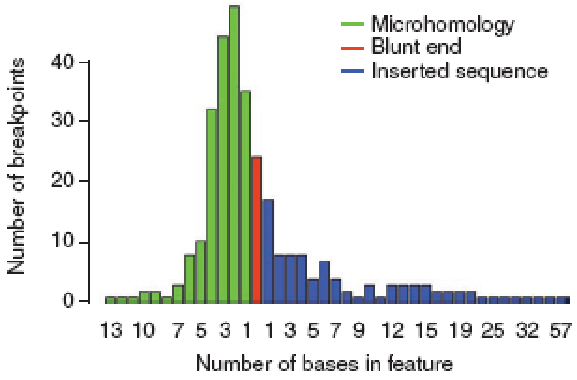


Figure 10: Sequence content at breakpoint of structural variations. Break points can show microhomology (70%), inserted sequence (33%), and blunt ends (8%). Approximately ten percent of all break points show both microhomology and insertion. Most of the microhomologies and insertions are smaller than seven base pairs. Blunt ends are zero base pairs in size (6).

The mutational mechanism can be defined by examining the sequence at the break points (6). Seventy percent of all structural variations show microhomology at the break points (figure 10). Thirty three percent of all structural variations contain an inserted sequence at the breakpoint. Ten percent of all break points show insertion, flanked by microhomology. The other eight percent are blunt ends. Microhomology is significantly more frequently present in break points without inserted sequence, showing that these two features do not arise independently of each other. There is no correlation between the size of structural variation and the frequency of sequence content at break point. Also, the length of microhomology and of inserted sequences does not correlated with the size of the structural variation (6).

Two of the mechanisms that have been proposed to induce structural variations, involve the repair of double-strand DNA breaks (DSBs) (6,24,25): NHEJ (nonhomologous DNA end joining) and HDR (homology directed repair). DSBs can both be pathological, for instance caused by ionizing radiation, and physiological, such as in the case with VDJ recombination (23,25). Pathological double-strand DNA breaks occur in all living cells (24,25). Approximately five to ten percent of dividing mammalian cells seem to have at least one chromosomal break at all times. Therefore, organisms need a mechanism to be able to repair these DSBs (25). It is possible to repair the DSBs both by means of NHEJ and HDR. Either one of them can result in a structural variation, when the two loose DNA-ends are adhered incorrectly.

It is unclear what defines whether NHEJ or HDR is used to repair a double-strand DNA break (24). DSBs induced by ionizing are most frequently repaired by NHEJ (24), which indicates that there is at least some selection between the two mechanisms. HDR is restricted to S and G_2 phase, while NHEJ can occur during the entire cell cycle. Therefore it seems as though NHEJ is the major mechanism in repairing DSBs (25). The contribution of each of them to pathogenic and non-pathogenic mutations however is not precisely known (6).

The other mechanism that has been suggested as an inducer for structural variations is the replication-based mechanism FoSTeS (fork stalling and template switching). This is the human form of MMBIR (microhomology mediated break induced replication). This mechanism is entirely different from the previous two, and might be able to explain more complex non-recurrent structural variations (26,27).

NHEJ

Nonhomologous DNA end joining, or NHEJ, is the major mechanism in which DNA breaks are repaired during the (entire) cell cycle of both prokaryotes and eukaryotes (25). Normally the two DNA-ends are ligated back to each other after a DSB has occurred. When the DNA strands are not ligated in the same way as before the break occurred, NHEJ can create non-recurrent structural variations such as deletions and translocations. NHEJ is therefore associated with several (congenital) diseases, including cancer (23).

NEHJ, just like any DNA repair pathway, requires the involvement of three different kinds of enzymes: endonuclease, polymerase, and ligase (23,25). Polymerase λ and µ are responsible for the polymerase activity during NHEJ. The enzymatic ligase-activity is due to a complex, which consists of ligase IV, XLF and XRCC4. In this complex, XRCC4 stabilizes ligase IV and XLF increases the ability of ligase IV to ligate. Finally, DNA-PKcs (Artemis-DNA-dependent protein kinase catalytic subunit) execute the endonuclease activity that is crucial in NHEJ (25).

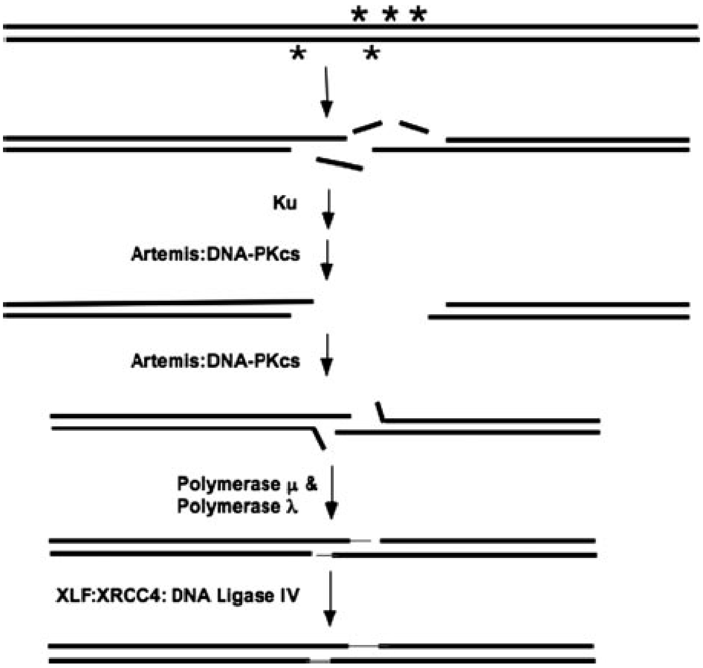


Figure 11: The NEHJ pathway in its simplest form. The exact order of NEHJ is not fixed. This is the simplest scenario of NHEJ. First, a DSB (*) occurs. Then Ku binds to the loose DNA-ends and attracts the enzymes, while strengthening their affinity for DNA. After activation, Artemis DNA-PKcs execute endonuclease activity. Then µ or λ polymerase extends the DNA. Finally, a ligase complex ligates the two loose ends. The grey sequences in the DNA represent information scars, formed by DNA polymerase (25).

NEHJ presumably starts with Ku (a heterodimeric protein of Ku70 and Ku80) binding to the loose DNA ends of a DSB. This protein strengthens all interactions between the enzymes and the DNA. Next, Ku recruits the enzymes that are needed to repair the DSB. The exact order in which these enzymes are recruited to the break can differ each time. There are multiple ways in which one DNA break might be repaired. In the simplest form, the order would be: endonuclease, polymerase and then ligase (figure 11) (25). First DNA-PKcs interacts with the DNA. Then the kinase activity of this protein activates itself, by either *cis* or *trans* autophosphorylation. This causes a conformation change. As a result, Artemis is able to function as an endonuclease. Next, polymerase µ or λ is attracted to elongate the loose DNA ends. Finally, the ligase complex ligates the loose DNA ends to each other (25).

NEHJ produces so called 'information scars'. These information scars are often the addition of 1-4 (microhomologous) base pairs, loss of 1-10 base pairs or inverted repeats at the break point (6,25). This is due to 'template slippage' of (either µ or λ) DNA polymerase. A possible benefit of this slippage, is that the ends that show microhomology are easier to ligate by ligase IV. Since µ polymerase executes template-independent DNA synthesis and λ polymerase executes template-dependent DNA synthesis, only µ polymerase can result in inverted repeats at the break point. The other two break point signatures can be a result of both polymerases (25).

Alternative EJ pathways

Not all of the above named proteins are essential in all end joining processes. In cell lines in which HDR is not possible, mutation of for instance DNA-PKcs only results in a much slower NHEJ, called B-NHEJ (back-up NHEJ), instead of an entirely deficient end-joining. The half time of normal NHEJ is approximately 20 minutes, while the half time of B-NHEJ is between two and ten hours. This means that more time is allowed for exchanges and thus mutations. B-NHEJ seems to be an evolutionary older variant of the current NHEJ as we know it in eukaryotic cells (24).

In fact, this is not the only study that claims that there are alternative EJ pathways. Some other recent studies report that a mutation in any protein that is essential in NHEJ results in an alternative end-joining process. One that is quite frequently discussed is MMEJ, or microhomology-mediated end joining. Unlike NHEJ, MMEJ requires microhomology during end-joining, due to a different ligase-enzyme activity (6,25). Ku deficient cell lines also show an alternative end joining pathway (28). MMEJ and all of these other 'alternative end-joining' pathways have however only been found in cells with a mutation in a NHEJ protein. Therefore, it is very possible that these alternative pathways are merely a possibility that is never applied, unless NHEJ itself is not working properly anymore (25).

HDR

HDR, or homology-directed repair, includes a major pathway which is suggested to induce recurrent structural variations: NAHR (nonallelic homologous repair) (6). During NAHR, double-strand DNA breaks are repaired by annealing homologous DNA fragments to each other. Due to the presence of abundant LCRs (low copy repeats) however, this process is prone to errors. LCRs, or segmental duplications (SD), are repeats of at least 200 base pairs to as much as several (hundred) thousand base pairs with so much similarity (often over 95%), that it is hard to distinguish them from one another (6,23). When a double-strand break occurs, the strands are therefore not always annealed in the right way. This cross-over, or 'misalignment', can result in multiple types of chromosomal rearrangements. NAHR mostly causes structural variations in between two LCRs. The LCRs are thus mediators for structural variations caused by NAHR, like deletions, duplications and inversions (figure 12) (23). On more rare occasions, NAHR can also use repetitive elements like *Alu* and *LINEs* as a substrate (23,26).

NAHR between two LCRs on the same chromosome and in the same orientation can result in a deletion or a tandem duplication of one LCR and the sequence in between the two LCRs. NAHR between two segmental duplications on the same chromosome but in opposite orientation can result in an inversion of the sequence of DNA in between the two LCRs. Furthermore, rearrangement of two low copy repeats on different chromosomes can result in a translocation.

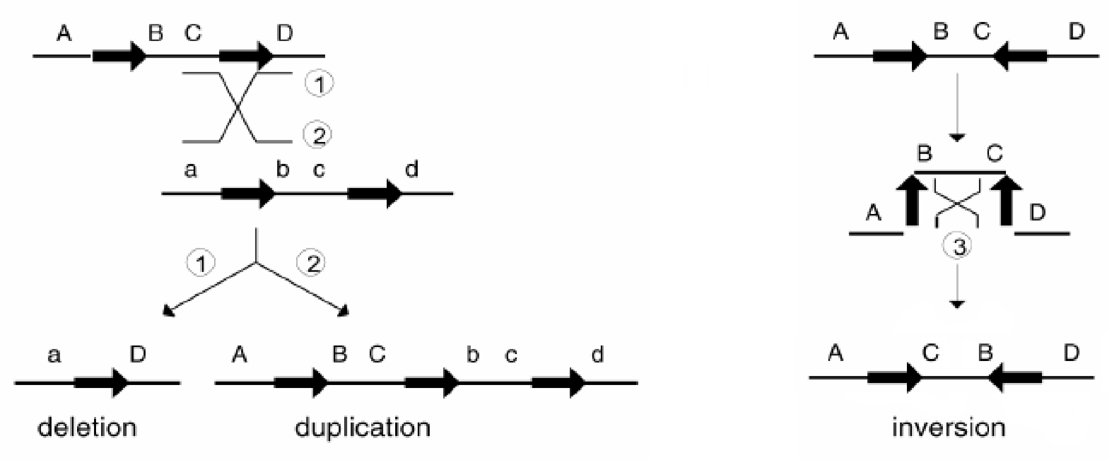


Figure 12: Different rearrangements caused by NAHR. The sequence in between two LCRs is affected by NAHR. **(Left)** Two tandem LCRs can cause a deletion or a tandem duplication. **(Right)** Two inverted LCRs can cause an inversion (adjusted from (23)).

The human genome is susceptible to the formation of structural variations as a result of NAHR. At least 3.6 percent of the human genome consists of segmental duplications (29). Since many possible SDs were excluded, the actual number might even be five percent. The Y-chromosome, with 10.9% intrachromosomal and 13.1% interchromosomal duplications, is at the highest risk for a structural variation due to NAHR (29). Recurrent nonallelic homologous recombination especially occurs in so-called 'hot spots'. These are short intervals which are prone to mutations. Hot spots have been indentified in several different diseases, including Smith-Magenis syndrome (SMS) and Potocki-Lupski syndrome (PTLS) (30).

PLTS is a congenital disease that is associated with mental retardation, autism, infantile hypotonia and cardiovascular abnormalities. The common type is associated with a 3.7 Mb duplication at 17p11.2. During primate evolution, several segmental duplications have arisen in the short arm of the seventeenth chromosome, resulting in a sequence that is susceptible to structural variations due to NAHR. SMS in fact, is associated with a 4 Mb deletion on that same chromosome 17p11.2 (30,31). This deletion is flanked by large low copy repeats, called Smith-Magenis syndrome repeats, or SMS-REPs. The sequence that is often deleted in SMS patients also contains an inverted SMS-REP in the middle (31). Species that are evolutionary older than 40 (to 65 million) years however (like the lemur), do not have these three repeats. Other studies have even found LCRs that only exist in humans and chimpanzees, but not in gorillas (31). This suggests that LCRs have originated relatively recently, making humans and other primates especially prone to chromosomal rearrangements during the cell cycle as a result of NAHR (31).

Replication-based mechanisms

Not all break points can be explained by NAHR or NHEJ. Some non-recurrent structural variations are too complex for these mechanisms (23). Therefore two replication-based models have been proposed: FoSTeS (fork stalling and template switching) and MMBIR (microhomology mediated break induced replication). MMBIR is a model that can be applied to all organisms, which encompasses FoSTeS for human recurrent rearrangements.

MMBIR

MMBIR is based on BIR, or break induced replication. BIR has been developed in yeast, but seems to be applicable to human cells as well. When the replication fork reaches a single-strand DNA break in the template strand, one arm breaks off, resulting in a collapsed fork. An exonuclease then crops the 5’ end of this arm, leaving an overhang at the 3’ end. This single-strand overhang will then interact with a long length homologous DNA sequence nearby (often the sister chromatid). A replication fork is formed, but DNA synthesis quickly ends again due to inefficiency problems. The 3’ end is then detached from its homologous DNA sequence. After a few repeats of this process (repeats of this process are however not required), DNA replication becomes more efficient and continues to the end of the chromosome. BIR is a relatively accurate process, which is associated with duplications and deletions of the genome. It requires microhomology of over fifty base pairs and mediation of the RecA/Rad51 protein during interaction with the homologous sequence (32).

MMBIR requires less homology and is therefore a more likable candidate as a structural variation inducer than BIR in some cases. The process is very similar to BIR. MMBIR is independent of RecA/Rad51, which makes it possible for the 3’ end to interact with sequences with shorter stretches of homology than fifty base pairs. The absence of Rad51 might be due to stress. Another protein, Rad52, is proposed to take over in absence of Rad51. MMBIR can cause non-recurrent structural variations with microhomology, such as duplications, deletions, translocations and inversions (32).

FoSTeS

The FoSTeS model (figure 13) is related to the MMBIR model (33). In this recently proposed model, structural variations occur as a result of stalling of the replication fork. The replication fork pauses during replication at or near for instance a LCR due to their genomic instability. This can induce collapsing of the replication fork. Then the lagging strand disengages and switches to another template with an active replication fork (sometimes even mega base pairs from the original replication fork) at a sequence with microhomology. Then, DNA replication might proceed normally, or FoSTeS might happen again. In the end, DNA replication will finish, but (multiple) errors have been made in the meantime (26).

FoSTeS is already associated with some diseases. Rett syndrome can for instance be caused by non-recurrent duplications of the MECP2 gene at Xq28 due to FoSTeS. Rett syndrome is a neurodevelopmental disorder that one out of ten thousand girls suffer from. A similar mutation in boys can even result in neurodevelopmental delay. Mutations in the MECP2 gene showed three to four nucleotide microhomology, which excludes NAHR as a possible mechanism, since NAHR needs longer stretches of homology. Additionally, 27 percent of the mutations that have been found in MECP2 were too complex to be explained by NHEJ. Therefore it is suggested that FoSTeS is sufficient to induce most structural variations that have been found in human with an altered MECP2 copy number. The LCRs in proximity of the MECP2 gene can induce the collapsed forks and thus cause FoSTeS (27).

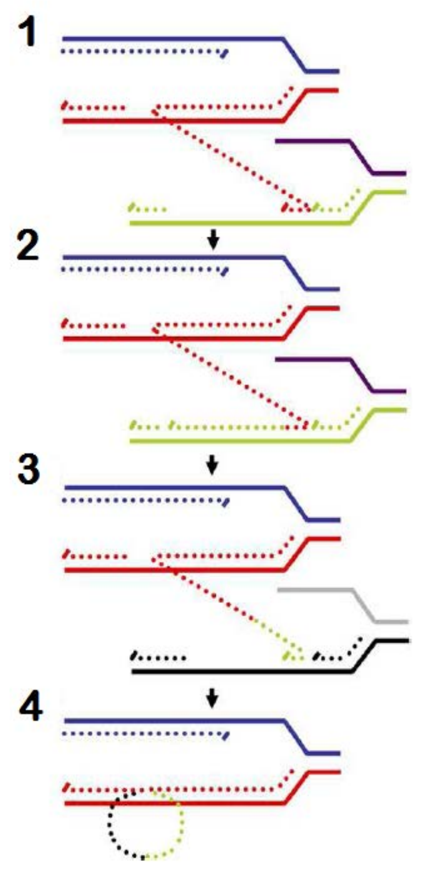


Figure 13: Fork stalling and template switching. (1) After the replication fork pauses, the lagging strand disconnects and switches to a replication fork at another template. **(2)** DNA is elongated. **(3)** The lagging strand disconnects again and often switches to another replication fork. DNA is then elongated again. **(4)** Finally, DNA replication will continue as usual. The end-product of FoSTeS is a DNA strand with some additional DNA sequence(s) (adjusted from (26)).

Other mechanisms

Even though NAHR, NHEJ, and FoSTeS are the known cause for many structural variations, not every chromosomal rearrangement can be explained by one of the mechanisms described above. Some rearrangements are too complex for instance to be a possible result of one of them. Others, which are probably created by a replication-based mechanism, might not have the necessary microhomologous sequence that is needed in FoSTeS. Or maybe in other cases the inserted sequence at the breakpoint is too long to agree with either HDR, NHEJ, or FoSTeS. Therefore, researchers are still trying to think of new mutational mechanisms that could eventually lead to (complex) chromosomal rearrangements. Most of the novel mechanisms are still based on previous research in bacteria, yeast or cell lines.

Retrotransposition

Retrotransposition of processed pseudogenes is actually an old and proven concept of a mutational mechanism that can lead to a structural variation. A processed mRNA is converted back to DNA by reverse transcriptase. With help of an endonuclease, it then integrates into the genome. This retrotransposition is caused by the enzymes that LINE-1 (long interspersed nuclear elements) codes for. Seventeen percent of the human genome consists of LINE-1s. These transpose LINEs, SINEs (short interspersed nuclear elements), and processed pseudogenes. Retrotransposition can be phenotype altering when the DNA-sequence is integrated in another gene or into functional non-coding DNA (34).

Alternative fork stalling

A more recent proposed mechanism is an alternative FoSTeS. It was suggested to be the cause of a deletion at 6p25(35). 6p25 is a region in the DNA that is very prone to chromosomal rearrangements. Seven duplications and three deletions in this region were sequenced. Most of them were caused by NAHR or NHEJ. One 1.2 Mb deletion with an 367bp inserted sequence at the breakpoint was too complex to have been caused solely by either of them. The inserted sequence consists of two homologous motifs (M1 and M2) with three blocks of $(GTG)_n$-repeats. The motifs are separated by a 13bp DNA-sequence (35).

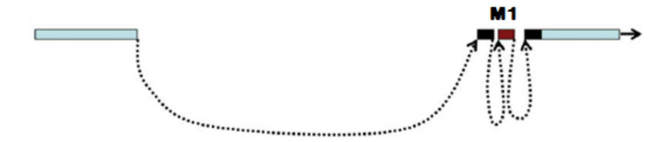


Figure 14: Alternative FoSTeS. The replication fork stalls due to a GTG-repeat. DNA-synthesis then continues at another GTG-repeat, ahead of the first GTG sequence. The result of (two repeats of) this process is a large deletion and thus the formation of the M1-motif (adjusted (35)).

The formation of such a chromosomal rearrangement starts with stalling of the replication fork due to the $(GTG)_n$-sequence. Then the fork continues DNA-synthesis at another $(GTG)_n$ sequence ahead of the sequence that caused the stalling. Repeating this process twice is the cause for the deletion and the formation of the first motif (figure 14) (35). Single strand dependent repair of a DSB using M1 as a template then results in the second motif. The difference between this mechanism and FoSTeS is that there is no evidence of template switching (35).

Chromothripsis

A novel proposed mechanism, initially developed in tumors, is Chromothripsis (or chromosome scattering). Mutations that contribute to cancer development usually happen over time. However, some complex rearrangements seem to have arisen at the same time, due to scattering of several chromosomes (Chromothripsis) and subsequent reassembly (figure 15A). Chromothripsis can be recognized by many rearrangements in (between) a few chromosomes (figure 15B). It has occurred in 2-3% of all tumors (36). More recently, chromothripsis has also been found to be able to induce complex de novo structural rearrangements resulting in complex congenital defects (33).

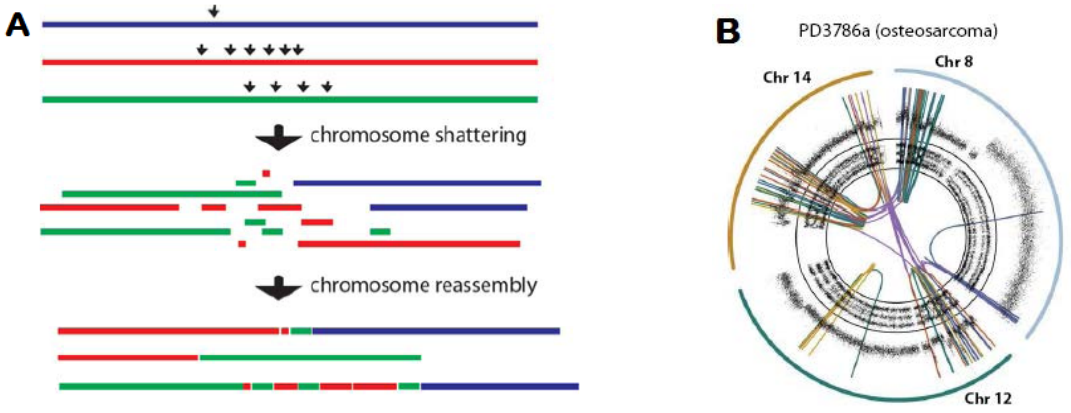


Figure 15: Chromothripsis. (A) First chromosome scattering or chromothripsis occurs. Then the chromosomes are reassembled (33). **(B)** Circos plot of the result of chromothripsis in a patient with osteosarcoma. The patient has 88 rearrangements in (between) chromosome 8, 12, and 14 (36).

Introduction to experiments

I performed two experiments. In the first experiment I tested how often four known deletions are present in 96 Dutch individuals. In the second experiment I have tested two family trios with a child with congenital defects on certain structural variations. These structural variations were sequenced to define their identity and possible mechanism in which they arose.

Materials and methods

PCR and gel electrophoresis analysis

PCR was performed in 27 cycles with an elongation time of 1:30 minutes for Taq Polymerase. The samples were then loaded in a 1% agarose gel. Gel electrophoresis was carried out for one hour at 120V. Visualization has been made possible by use of ethidiumbromide. A 50 base pair GeneRuler^{TM} marker was used in order to be able to determine the size of each fragment.

Significance testing

H_0: $\pi_1 = \pi_2$

H_1: $\pi_1 \neq \pi_2$

$\alpha = 0.05$ (two-sided)

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p_0(1 - p_0)(1/n_1 + 1/n_2)}}$$

$p > \alpha$ → H_0 rejected

Primer pairs and DNA-samples

Experiment 1

Four primer pairs (table 2) that have previously been identified as primers that can detect certain non-pathogenic deletions were tested on 95 random individuals (48 men and 47 women) that have donated their blood for use in genomic research.

Table 2: Primer pairs of the first experiment

Primer pair		Location	Sequence	Deletion
4	Forward	1:89248102-89251653-164F	ATTGGGTTTCTGTCTCTTGG	2717bp
	Reverse	1:89248102-89251653-416R	CTCTTTCAGGAGGCATCAAG	
51	Forward	22:19280163-19284607-108F	ATAAGTGGCTTCCAAGAAGG	1983bp
	Reverse	22:19280163-19284607-424R	CCCTAAATGGCCAATAACTC	
71	Forward	5:57358891-57369976-96F	CAGGCGATTCTAGCCTATTC	10293bp
	Reverse	5:57358891-57369976-486R	TGCATTCCATCTTAGGTTCC	
76	Forward	X:126425103-126430447-157F	CATTGCTATATGCCAACAGTG	4638bp
	Reverse	X:126425103-126430447-436R	ATTAGAGCTCCTCTGCCAAG	

Experiment 2

Six primer pairs (table 3) and then their nested primers (table 4) were tested on the DNA-samples of two children with different congenital defects (the first female and the second male) and on DNA-samples of their parents.

Table 3: Primer pairs of the second experiment

Primer		Name	Sequence
1.1	F	8_53499878_53503376_-_8_55161236_55165294-224F	AGGGAAACAGGTCCCTTG
	R	8_53499878_53503376_-_8_55161236_55165294-601R	GTGTGTGCTTGTAGTTTCAGC
2.1	F	8_80535197_80538185_-_8_81944205_81947323-167F	GGAAGGCTAAATTGATCCAG
	R	8_80535197_80538185_-_8_81944205_81947323-485R	CAAGGAACAAGGCAACATC
3.1	F	1_154445695_154448216_-_3_156530557_156533390-216F	TGTAGAGCTGGGCTCAGTG
	R	1_154445695_154448216_-_3_156530557_156533390-563R	GGTGACAGAGCAAGACTCC
4.1	F	1_153746565_153748683_-_3_158389428_158392484-127F	AGTTTCTGTGGCTCTGGTTC
	R	1_153746565_153748683_-_3_158389428_158392484-536R	TAAATGATGTGCACCCTCTG
5.1	F	14_93652833_93653396_-_3_169726126_169727784-42F	AACATGTGATTAGGGAGCTATC
	R	14_93652833_93653396_-_3_169726126_169727784-528R	CGTCTGGGCAACAGAGC
6.1	F	14_93652776_93653405_-_3_169730758_169733339-124F	ATTGCAAATAACTGCCAAGC
	R	14_93652776_93653405_-_3_169730758_169733339-480R	TGAATAATGATGCCACAAGG

Table 4: Nested primer pairs of the second experiment

Primer		Name	Sequence
1.2	F	8_53499878_53503376_-_8_55161236_55165294-257F	GCAGTTGATAGATGGGCATAG
	R	8_53499878_53503376_-_8_55161236_55165294-544R	GAGGTTGAGGCTGCTGTG
2.2	F	8_80535197_80538185_-_8_81944205_81947323-244F	TAAAGTGGAAGCAGGAGAGC
	R	8_80535197_80538185_-_8_81944205_81947323-449R	TAACCCTTATTTGGGTGTCG
3.2	F	1_154445695_154448216_-_3_156530557_156533390-267F	CTGAGACAGGCGGATCAC
	R	1_154445695_154448216_-_3_156530557_156533390-455R	GCCCACACAGCTAATACTTG
4.2	F	1_153746565_153748683_-_3_158389428_158392484-265F	CTGGAGCTCCGAACTGAC
	R	1_153746565_153748683_-_3_158389428_158392484-492R	AGGCCTCAGCAATCACTAAC
5.2	F	14_93652833_93653396_-_3_169726126_169727784-216F	CTGTGCAACATAGTGATGATTC
	R	14_93652833_93653396_-_3_169726126_169727784-469R	CAACCTAAACCTCCGATTTG
6.2	F	14_93652776_93653405_-_3_169730758_169733339-257F	GGCAAGGAGAGTAATTGAGC
	R	14_93652776_93653405_-_3_169730758_169733339-438R	GCCAGATGCAATTTAAGAGG

Capillary sequencing of breakpoints and analysis of sequence reads

PCR samples were purified and then prepared for a sequence reaction with Bigdye®. The sequence products were purified by using a sephadex loader. Next, samples were stored at -20^{o}C in the freezer for two weeks, until capillary sequencing was performed. Sequence reads were analyzed with BLAT software. Break points were analyzed manually to define the exact break point and break point signature. Affected genes were found in Ensembl and NCBI. A potential mechanism that could have led to the structural variations has also been composed, based on the DNA-sequence of the breakpoint, of the structural variation, and of the flanking DNA-sequences.

Results

Experiment 1

The goal of this experiment was to determine how often some recurrent deletions occur in the entire population. Four primer pairs were tested on 95 random individuals (48 men and 47 women) and a positive control was used to determine whether the PCR had succeeded (figure 16). The primer pairs had previously been identified as primers that can detect certain non-pathogenic deletions. PCR was performed as described above. Table 5 shows the results of this experiment. Differences between prevalence in men and women were tested on significance as described.

Table 5: Results of the first experiment

Primer pair	Band (bp)	Male	Female	Prevalence
4	500	28 / 48 = 58%	29 / 47 = 62%	60%
51	1000	16 / 48 = 33%	17 / 47 = 36%	35%
71	600	43 / 48 = 90%	45 / 47 = 96%	93%
76	400	12 / 48 = 25%	26 / 47 = 55%	40%

Only bright bands were counted as deletions, and less bright bands at the same height were considered to be a contamination. This reduces the chance of overestimating the occurrence of the deletions in these individuals.

Primer 4

Individuals with a deletion of 2717 base pair in chromosome 1 have a 500 base pairs band. This deletion in between 1:89475928 and 1:89478645 encompasses exon 7 of GPB3 (guanylate binding protein 3). This deletion was present in 58% of all males and 62% of all females. Three out of five individuals thus have this deletion in the first chromosome. There is no significant difference in prevalence between men and women (p= 0.74).

Primer 51

Primer 51 will show a band of 1000 base pairs in case of a intergenic deletion of 1983 base pairs in the 22th chromosome. This deletion (at 22:20950624-20952607) was present in 33% of all men and 36% of all women. Thirty-five percent of all individuals thus have a deletion in between the forward and reverse primer of primer pair 51. There is no significant difference in prevalence between men and women (p=0.77). Interestingly, a band at 500 base pairs was found in four individuals (one man and three women) (figure 16), indicating that this primer pair might encompass two different deletions, one 500 base pairs shorter (which is more common) than the other.

Primer 71

PCR with primer 71 will show a band of 600 base pairs in case of a 10293bp intergenic deletion in between 5:57323478 and 5:57333771. This deletion was present in a staggering 90% of all males and 96% of all females. 93 percent of all individuals thus have this deletion in the fifth chromosome. There is no significant difference in prevalence between men and women (p = 0.25).

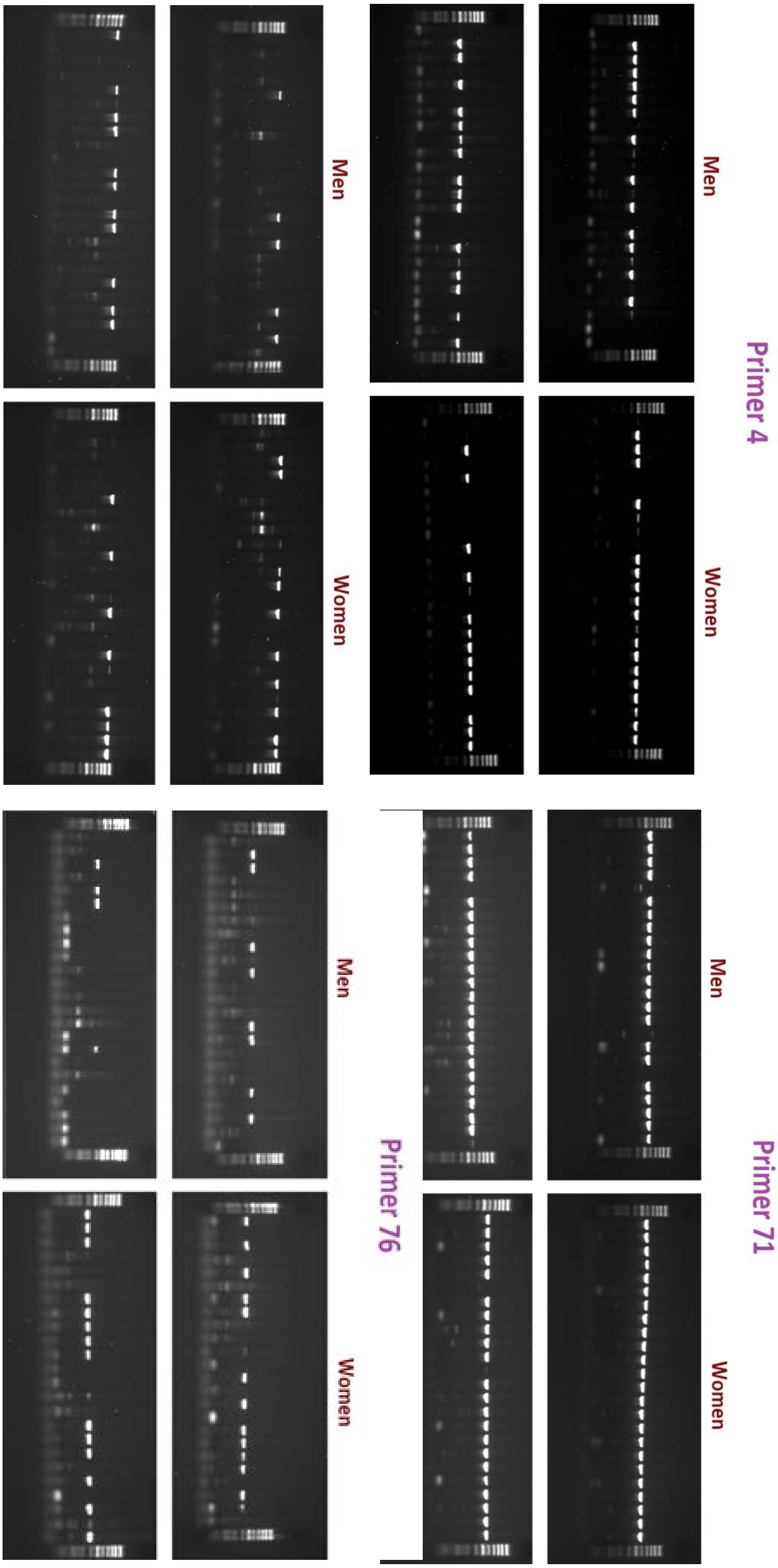


Figure 16: Gel of four primer pairs tested on 95 individuals. The 96th sample in the bottom right position of the bottom right gel of every primer is a positive control . **(Primer 4)** 28 men and 29 women show a band at 500bp. 60 percent of all individuals thus have a 2717bp deletion in the first chromosome in between 1:89475928 and 1:8947864. **(Primer 51)** 16 men and 17 women (prevalence = 35%) have a band at 1000bp. Interestingly, three women and one man have an unexpected band at 500bp. **(Primer 71)** 43 men and 45 women show a band at 600bp. A 10293bp deletion in the fifth chromosome is thus present in 93 percent of all individuals. (**Primer 76)** 12 men and 26 women have a band at 400bp. The prevalence of the4638bp deletion in the X-chromosome is thus 0,4.

Primer 76

PCR with primer 76 will show a band of 400 base pairs in case of a 4638bp intergenic deletion. This deletion (at: X:126597760-126602398) was present in 25% of all men and 55% of all women. Forty percent of all individuals thus have a deletion on the X-chromosome in between the forward and reverse primer of primer pair 76. Women encompass this deletion significantly more often than men (p=0.026). This increased chance in women is due to the fact that women have two X-chromosomes. When corrected for the amount of X-chromosomes, there is no significant difference in prevalence between men and women (p=0.72).

Experiment 2

The goal of the second experiment was to sequence break points of structural variations that have been found in two children with congenital defects (and their family) and identify which genes are affected by the structural variation. Six primer pairs and their nested primers were tested on two patients with severe congenital abnormalities and their parents. PCR was performed as described above. In the first family, I observed a band at 300bp with the first primer and a band at 350bp with the second primer in the first patient (figure 17). Neither of the parents had any bands, which means that it concerns *de novo* structural variations. In the second family, both the patient and his mother showed a 800bp band with primer four and a >1000bp band with primer six (figure 17). This patient also showed a second band at 400 to 500bp with primer 6. This could be a contamination and this sample was thus not further analyzed. The other five samples (table 6) were purified and sequenced as described above. Unfortunately, sequencing of the first sample has failed. The sequence of the other four samples was analyzed as described above. Also, I tried to identify the potential mutational mechanism that could have led to each of the structural variations.

Sample 2

The first patient has a *de novo* tandem duplication of 1.4 Mb in chromosome 8 (table 7). The duplication encompasses five genes. HEY1 (Hairy/enhancer-of-split related with YRPW motif 1) is involved in multiple processes like for instance angiogenesis, (independent) regulation of transcription, nervous system development, and organism development. MRPS28 codes for mitochondrial ribosomal protein S28. This is the small ribosomal subunit of the mitochondrion, which functions in the protein synthesis of the mitochondrion. TPD52 codes for tumor protein D52, which is associated with many different types of tumors. ZBTB10 (Zinc finger and BTB domain containing 10) has a function in regulation of transcription. Finally, another zinc finger was duplicated, of which the function is not entirely clear yet: ZNF704 (Zinc finger protein 704) (39).

Table 6: samples that were sequenced for further analysis

Sample ID			
1	Family 1	Patient (F)	Primer 1
2		Patient (F)	Primer 2
3	Family 2	Mother (F)	Primer 3
4		Patient (M)	Primer 3
5		Mother (F)	Primer 5

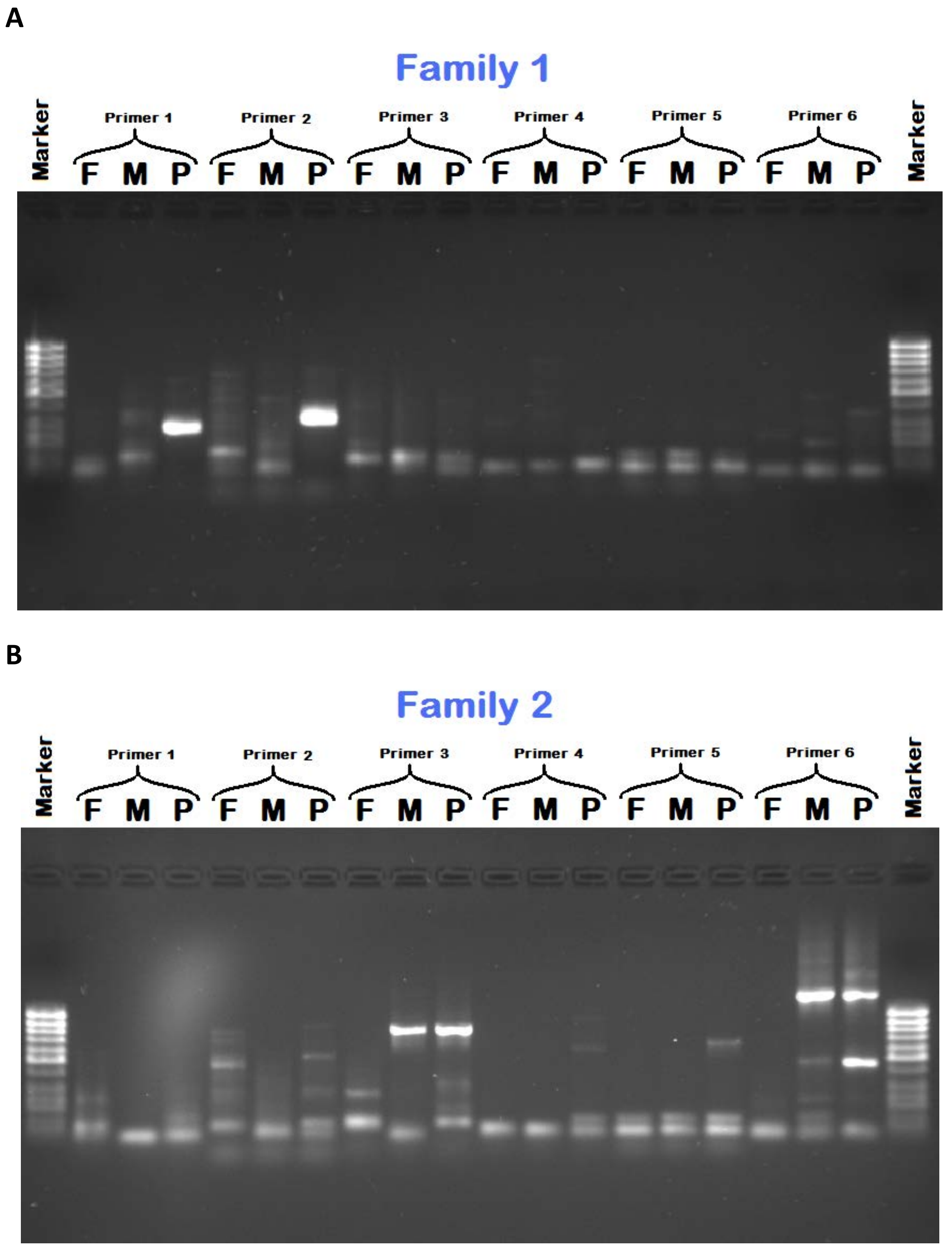


Figure 17: Gel of family 1 and 2. (A) The patient shows a band of ±300bp with the first primer and a band of ±350bp with the second primer. Neither of the parents show any bands with any primer. **(B)** Both the patient and his mother have a ±800bp band with primer pair four and a >1000bp band with primer pair six. The patient also has a band at approximately 400-500bp with primer six, which could indicate that this sample has been contaminated.

Table 7: Identification of structural variation and their mutational mechanism

ID	Mutation	Breakpoint analysis	Chromosomal pieces	Mechanism
Sample 2	Tandem duplication	8-CAATCAATCTTA TCAATTGA(CAAT) -8 insertion	8: 81947257 - 81947409 → 8: 80535131 - 80535255	FoSTeS
Sample 3 **Sample 4**	Insertion - translocation	CG microhomology	3: 156533319 – 156534038 → 1: 154445664 - 15444729	MMEJ
Sample 5	Insertion - inverted translocation & deletion	Blunt end & Blunt end	3: 169730770 – 169729734 → 14: 93651319 – 93651377 → 14: 93652605 – 93652769	Retrotransposition

The 1.4 Mb tandem duplication in the eight chromosome of the first patient also resulted in two non-processed pseudo genes: a partial duplication of the PAG1 gene and the STMN2 gene. PAG1 on the reverse strand codes for the phosphoprotein associated with glycosphingolipid microdomains 1 protein (39). The sequence that is duplicated in the patients' genome includes the sequence of PAG1 from the intron in between the first and second exon, indicating that the entire coding sequence, but no regulatory elements, have been duplicated. The STMN2 gene on the forward strand codes for Stathmin-like-2 protein. Reductions in the expression of this gene have been associated with Down's syndrome and Alzheimer's disease (39). The duplication of STMN2 lacks the first exon and thus part of the 5'UTR, but the protein coding sequence of one out of four transcripts (STMN2-007) is still intact. The upstream regulatory sequences of this non-processed pseudogene have not been duplicated.

The entire tandem duplication is 1.4Mb and the breakpoint has an insertion of 20 base pairs, with a 4bp microhomologous pattern (CAAT). Due to its size, it is not likely that the mutation has been caused by either NAHR or NHEJ. However, at first sight there is no microhomology which would suggest a replication-based mechanism such as FoSTeS.

Sample 3 and 4

Both the mother of the second patient and the second patient showed a band of approximately 750bp in between the forward and reverse primer of primer three. However, the sequence obtained by both of the primers can almost entirely be aligned to a normal fragment of the third chromosome (table 7). Sequencing with the forward primer did not yield any further sequence which could indicate a mutation. A ±65bp fragment of the ±750bp sequence obtained by the reverse primer on the other hand does have many alignments other than the third chromosome (almost 200). Also, the sequence is the same in both mother and child, which means that it is not likely to be the result of a failed sequence reaction. These observations indicate that this fragment might be the mutation that should be found. However, the sequence has too many alignments and is too short, to irrefutably prove on which chromosome the DNA-sequence is originally found. The best alignment (by far) is with a sequence of the first chromosome (1: 154445664 – 15444729). No genes seem to be interrupted.

A piece of the third chromosome is translocated to the first chromosome, or the other way around. The two alignments share two base pairs of microhomology. This is an indication that this rearrangement has been caused by a mechanism that needs microhomology, such as NAHR or alternative end-joining processes such as MMEJ. The overlap is too short for NAHR, meaning that MMEJ is the most plausible mutational mechanism.

Sample 5

The mother of the second patient (and perhaps the second patient as well) also has another structural variation (table 7). This is an inverted insertion of the third chromosome in the fourteenth chromosome or of the fourteenth chromosome in the third chromosome and an additional deletion of 1228bp in the fragment of chromosome 14 that has been sequenced. The fragment of the third chromosome that has been sequenced does not contain any genes. In chromosome 14 however, a gene is interrupted: part of the C14orf109 (chromosome 14 open reading frame 109) has been deleted. The deletion is exactly the one (only) intron of transcript C14orf109-201.This can be an indication that the fragment of chromosome 14 is a processed pseudogene, retrotransposed into the third chromosome with help of LINE-1.

Discussion

Research on DNA, the most fascinating molecule in the entire universe, has progressed from defining a double-helix structure in the nineteen fifties to determining variations in the human genome. These variations range from single nucleotides to gross alterations and can alter phenotype. Structural variations are all variations longer than one base pair: deletions, insertions, translocations, inversions and duplications.

The extent to which our genomes differ is not entirely clear yet. The goal of the first experiment was to determine how often four recurrent deletions occur in the entire population. Sixty percent of all men and women have a 2717bp deletion in the first chromosome. Thirty-five percent have a 1983bp deletion in the 22th chromosome. A staggering 93% have a deletion of 10293 base pairs in the fifth chromosome. Finally, forty percent of all individuals have a 4638bp deletion in the X-chromosome. Women encompass this last deletion twice as often, since they have two X-chromosomes. These results show that (at least some) recurrent non-pathogenic deletions are commonly present in the population.

The change in phenotype caused by rearrangements can even result in diseases. Rearrangements can change genes or gene function by altering gene dosage, disrupting the sequence of a gene, creating a fusion gene, altering gene expression or unmasking recessive mutations. In non-protein-coding sequence, they also can have an effect on phenotype, by for instance disrupting a miRNA or a promoter. Structural variations are in fact associated with many different diseases, ranging from color blindness to the Prader-Willi syndrome.

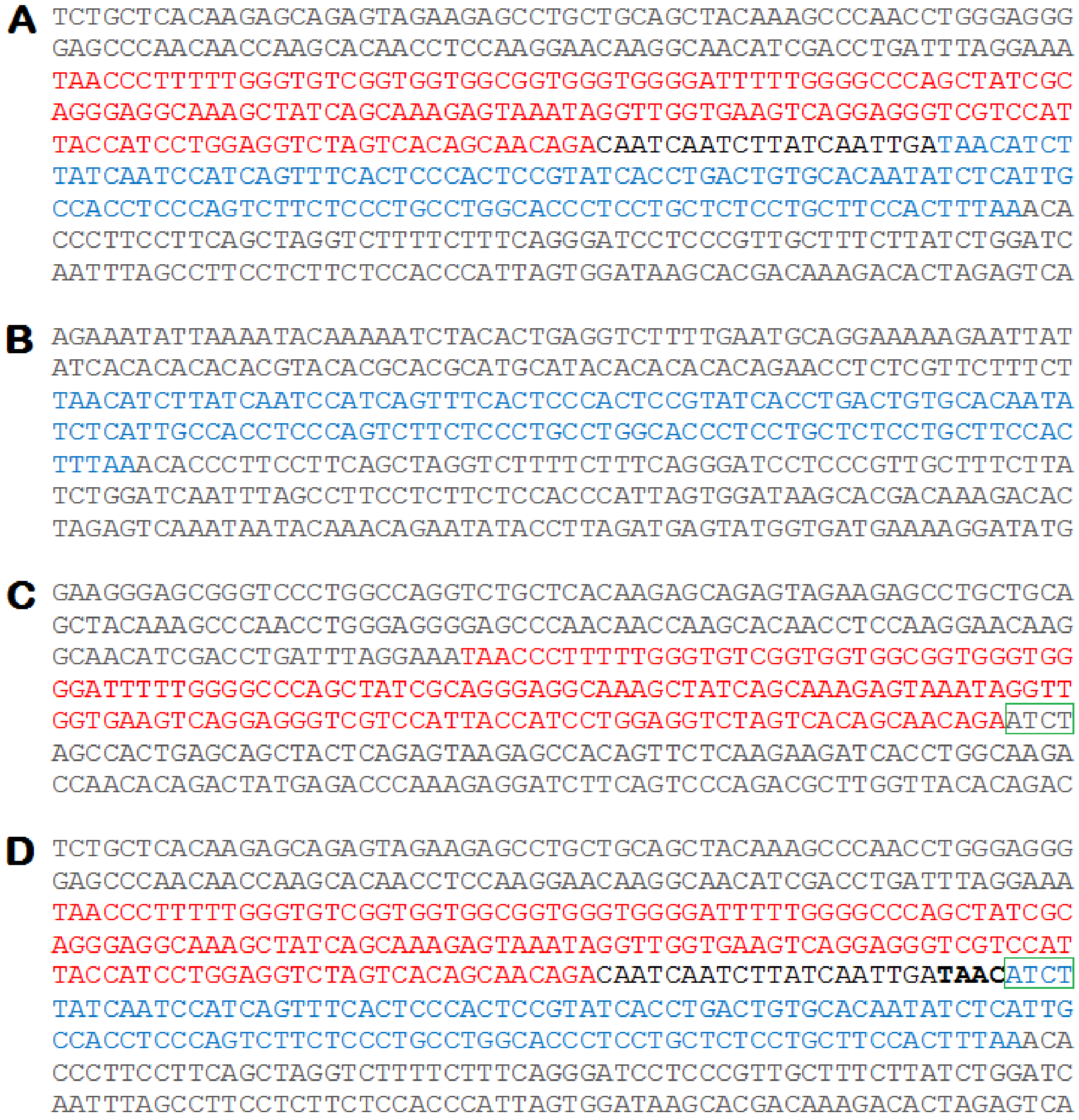
Structural variations can be found by FISH or karyotyping. Then primers can be designed so that specific regions of the DNA can be sequenced. CNV arrays are a cheaper alternative in defining copy number variations. Sequencing the break point can also give a hint of the mutational mechanism by which the structural variation has been caused. This is very important, since we might be able to predict structural variations better if we know mechanisms in which they arise. Unfortunately, mutational mechanisms have not been established thoroughly. NHEJ and NAHR are mechanisms that can repair DSBs in the DNA, sometimes resulting in a SV. FoSTeS is a replication-based mechanism, which can result in a SV. Some other models have been proposed, like: retroposition, alternative fork stalling, and chromothripsis.

Finding and defining structural variations is important, since we will then be able to quickly establish the cause for some diseases. This will enable us to develop specific medicines more quickly. The goal of the second experiment was to sequence certain structural variations in two families with children with congenital defects and define the mutation and the mutational mechanism. The end goal was obviously to see whether these structural variations found, could explain the congenital defects of the two patients.

The rearrangement (sample 3 and 4) found in both the second mother and the second patient was an inserted translocation of chromosome 3 in chromosome 1 or of chromosome 1 in chromosome 3. No genes were affected by this translocation. Interestingly, there was two base pair microhomology between the chromosomal pieces. This is an indication that this structural variation was caused by MMEJ. The structural variation (sample 5) found in the mother of the second patient was a processed pseudogene of chromosome 14 inserted in inverted orientation into the reverse strand of chromosome 3. A seeming deletion in the fragment of chromosome 14 encompassed exactly the only intron of the transcript C14orf109-201, which indicates a structural variation caused by retroposition. Neither of these structural variations disrupted any genes or altered gene numbers. Since they were both present in the mother, neither could have explained the phenotype of the second patient. Therefore, more extensive research should be conducted in the future to establish the rearrangements in the patients DNA that could have led to his congenital defects.

The structural variation (sample 2) in the first patient was the only *de novo* structural variation found (and sequenced), and was identified to be a tandem duplication of 1.4 Mb. Five genes have been duplicated as a result of this duplication: HEY1, MRPS28, TPD52, ZBTB10, and ZNF704. Also, two non-processed pseudogenes (PAG1 and STMN2) had been formed. The coding sequence of these genes is still intact. Either of these changes could be the cause for the congenital disease of the child. Due to the size of the variation, it has to have been caused by a replication-based mechanism, such as FoSTeS. The essential microhomology needed to make FoSTeS possible however seems to be missing (figure 18A-18C). When looking more closely at the sequence it can nevertheless be a result of FoSTeS (figure 18D). The four base pairs that have been aligned to 8: 80535131 – 80535134 are, coincidentally, TAAC. These are the four base pairs of the microhomologous pattern at the break point (only in a different order) and could thus be part of the break point signature. The four next bases are the exact same as would have been, if DNA replication continued as usual. I therefore suggest that this mutation is a result of fork stalling and template switching.

Another possible explanation would be that FoSTeS is affected by the presence of a sequence that is similar to the microhomologous sequence added to the break point. It could even be a combination of both: the microhomologous sequence of the break point and the matching next few base pairs. No previous data however support either of these statements, and I thus find it unlikely that this would be the case.



A

```
TCTGCTCACAAGAGCAGAGTAGAAGAGCCTGCTGCAGCTACAAAGCCCAACCTGGGAGGG
GAGCCCAACAACCAAGCACAACCTCCAAGGAACAAGGCAACATCGACCTGATTTAGGAAA
TAACCCTTTTTGGGTGTCGGTGGTGGCGGTGGGTGGGGATTTTTGGGGCCCAGCTATCGC
AGGGAGGCAAAGCTATCAGCAAAGAGTAAATAGGTTGGTGAAGTCAGGAGGGTCGTCCAT
TACCATCCTGGAGGTCTAGTCACAGCAACAGACAATCAATCTTATCAATTGATAACATCT
TATCAATCCATCAGTTTCACTCCCACTCCGTATCACCTGACTGTGCACAATATCTCATTG
CCACCTCCCAGTCTTCTCCCTGCCTGGCACCCTCCTGCTCTCCTGCTTCCACTTTAAACA
CCCTTCCTTCAGCTAGGTCTTTTCTTTCAGGGATCCTCCCGTTGCTTTCTTATCTGGATC
AATTTAGCCTTCCTCTTCTCCACCCATTAGTGGATAAGCACGACAAAGACACTAGAGTCA
```

B

```
AGAAATATTAAAATACAAAAATCTACACTGAGGTCTTTTGAATGCAGGAAAAAGAATTAT
ATCACACACACACGTACACGCACGCATGCATACACACACACAGAACCTCTCGTTCTTTCT
TAACATCTTATCAATCCATCAGTTTCACTCCCACTCCGTATCACCTGACTGTGCACAATA
TCTCATTGCCACCTCCCAGTCTTCTCCCTGCCTGGCACCCTCCTGCTCTCCTGCTTCCAC
TTTAAACACCCTTCCTTCAGCTAGGTCTTTTCTTTCAGGGATCCTCCCGTTGCTTTCTTA
TCTGGATCAATTTAGCCTTCCTCTTCTCCACCCATTAGTGGATAAGCACGACAAAGACAC
TAGAGTCAAATAATACAAACAGAATATACCTTAGATGAGTATGGTGATGAAAAGGATATG
```

C

```
GAAGGGAGCGGGTCCCTGGCCAGGTCTGCTCACAAGAGCAGAGTAGAAGAGCCTGCTGCA
GCTACAAAGCCCAACCTGGGAGGGGAGCCCAACAACCAAGCACAACCTCCAAGGAACAAG
GCAACATCGACCTGATTTAGGAAATAACCCTTTTTGGGTGTCGGTGGTGGCGGTGGGTGG
GGATTTTTGGGGCCCAGCTATCGCAGGGAGGCAAAGCTATCAGCAAAGAGTAAATAGGTT
GGTGAAGTCAGGAGGGTCGTCCATTACCATCCTGGAGGTCTAGTCACAGCAACAGAATCT
AGCCACTGAGCAGCTACTCAGAGTAAGAGCCACAGTTCTCAAGAAGATCACCTGGCAAGA
CCAACACAGACTATGAGACCCAAAGAGGATCTTCAGTCCCAGACGCTTGGTTACACAGAC
```

D

```
TCTGCTCACAAGAGCAGAGTAGAAGAGCCTGCTGCAGCTACAAAGCCCAACCTGGGAGGG
GAGCCCAACAACCAAGCACAACCTCCAAGGAACAAGGCAACATCGACCTGATTTAGGAAA
TAACCCTTTTTGGGTGTCGGTGGTGGCGGTGGGTGGGGATTTTTGGGGCCCAGCTATCGC
AGGGAGGCAAAGCTATCAGCAAAGAGTAAATAGGTTGGTGAAGTCAGGAGGGTCGTCCAT
TACCATCCTGGAGGTCTAGTCACAGCAACAGACAATCAATCTTATCAATTGATAACATCT
TATCAATCCATCAGTTTCACTCCCACTCCGTATCACCTGACTGTGCACAATATCTCATTG
CCACCTCCCAGTCTTCTCCCTGCCTGGCACCCTCCTGCTCTCCTGCTTCCACTTTAAACA
CCCTTCCTTCAGCTAGGTCTTTTCTTTCAGGGATCCTCCCGTTGCTTTCTTATCTGGATC
AATTTAGCCTTCCTCTTCTCCACCCATTAGTGGATAAGCACGACAAAGACACTAGAGTCA
```

Figure 18: Sequence of break point of tandem duplication in the first patient. Red = 8: 81947257 – 81947409; Blue = 8: 80535131 – 80535255; Black = insertion of sequence at breakpoint; Grey = sequence attached to the fragments, that has not been sequenced. **(A)** Break point sequence. **(B)** Normal DNA-sequence surrounding 8: 80535131 – 80535255. **(C)** Normal DNA-sequence surrounding 8: 81947257 – 81947409. **(D)** If the first four base pairs of the aligned sequence are attributed to the break point sequence, microhomology of four base pairs (framed in green) is found with the sequence that would have been sequenced if replication continued as normal.

This sequenced *de novo* tandem duplication in the eight chromosome could in theory be the cause for the congenital defects of the first patient, since several genes have been (partially) duplicated. The phenotype of the patient should therefore be compared to the (known) consequences of changes in CNV of these genes in future research. One other *de novo* mutation was also found. Sequencing of this rearrangement unfortunately failed. This *de novo* structural variation can also be the cause for the congenital defects. As mentioned in the introduction, answering questions in doing research on DNA gives rise to many new mysteries in need of elucidation.

Acknowledgements

I thank W.P. Kloosterman for helping me write this essay. He was always available for answering questions, he taught me how to process my results and he helped me finish something that seemed unattainable at some moments. I thank both W.P. Kloosterman and K. Duran for teaching me the ropes on genomic research. Finally, I thank a third (anonymous) person for spending time on checking this essay.

References

1. Watson J.D. and Crick F.H. (1993) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. 1953. *JAMA,* **269**, 1966-1967.

2. Human Genome Structural Variation Working Group, Eichler E.E., Nickerson D.A., Altshuler D., Bowcock A.M., Brooks L.D., Carter N.P., Church D.M., Felsenfeld A., Guyer M. *et al.* (2007) Completing the map of human genetic variation. *Nature,* **447**, 161-165.

3. 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.

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

5. Campbell C.D., Sampas N., Tsalenko A., Sudmant P.H., Kidd J.M., Malig M., Vu T.H., Vives L., Tsang P., Bruhn L. *et al.* (2011) Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.,* **88**, 317-332.

6. Conrad D.F., Bird C., Blackburne B., Lindsay S., Mamanova L., Lee C., Turner D.J. and Hurles M.E. (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.,* **42**, 385-391.

7. Grossmann V., Hockner M., Karmous-Benailly H., Liang D., Puttinger R., Quadrelli R., Rothlisberger B., Huber A., Wu L., Spreiz A. *et al.* (2010) Parental origin of apparently balanced de novo complex chromosomal rearrangements investigated by microdissection, whole genome amplification, and microsatellite-mediated haplotype analysis. *Clin. Genet.,* **78**, 548-553.

8. Gonzalez E., Kulkarni H., Bolivar H., Mangano A., Sanchez R., Catano G., Nibbs R.J., Freedman B.I., Quinones M.P., Bamshad M.J. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science,* **307**, 1434-1440.

9. Cheng F., Song W., Kang Y., Yu S. and Yuan H. (2011) A 556 kb deletion in the downstream region of the PAX6 gene causes familial aniridia and other eye anomalies in a chinese family. *Mol. Vis.,* **17**, 448-455.

10. Osborne L.R., Li M., Pober B., Chitayat D., Bodurtha J., Mandel A., Costa T., Grebe T., Cox S., Tsui L.C. *et al.* (2001) A 1.5 million-base pair inversion polymorphism in families with williams-beuren syndrome. *Nat. Genet.,* **29**, 321-325.

11. Adrianto I., Wen F., Templeton A., Wiley G., King J.B., Lessard C.J., Bates J.S., Hu Y., Kelly J.A., Kaufman K.M. *et al.* (2011) Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. *Nat. Genet.,* **43**, 253-258.

12. Li M., Marin-Muller C., Bharadwaj U., Chow K.H., Yao Q. and Chen C. (2009) MicroRNAs: Control and loss of control in human physiology and disease. *World J. Surg.,* **33**, 667-684.

13. Schofield C.M., Hsu R., Barker A.J., Gertz C.C., Blelloch R. and Ullian E.M. (2011) Monoallelic deletion of the microRNA biogenesis gene Dgcr8 produces deficits in the development of excitatory synaptic transmission in the prefrontal cortex. *Neural Dev.,* **6**, 11.

14. Stephens P.J., McBride D.J., Lin M.L., Varela I., Pleasance E.D., Simpson J.T., Stebbings L.A., Leroy C., Edkins S., Mudie L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature,* **462**, 1005-1010.

15. Leary R.J., Kinde I., Diehl F., Schmidt K., Clouser C., Duncan C., Antipova A., Lee C., McKernan K., De La Vega F.M. *et al.* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.,* **2**, 20ra14.

16. Conrad D.F., Andrews T.D., Carter N.P., Hurles M.E. and Pritchard J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.,* **38**, 75-81.

17. Chen W., Ullmann R., Langnick C., Menzel C., Wotschofsky Z., Hu H., Doring A., Hu Y., Kang H., Tzschach A. *et al.* (2010) Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. *Eur. J. Hum. Genet.,* **18**, 539-543.

18. Fellermann K., Stange D.E., Schaeffeler E., Schmalzl H., Wehkamp J., Bevins C.L., Reinisch W., Teml A., Schwab M., Lichter P. *et al.* (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to crohn disease of the colon. *Am. J. Hum. Genet.,* **79**, 439-448.

19. Campbell P.J., Yachida S., Mudie L.J., Stephens P.J., Pleasance E.D., Stebbings L.A., Morsberger L.A., Latimer C., McLaren S., Lin M.L. *et al.* (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature,* **467**, 1109-1113.

20. 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.

21. Nowrousian M. (2010) Next-generation sequencing techniques for eukaryotic microorganisms: Sequencing-based solutions to biological problems. *Eukaryot. Cell.,* **9**, 1300-1310.

22. Rusk N. and Kiermer V. (2008) Primer: Sequencing--the next generation. *Nat. Methods,* **5**, 15.

23. Gu W., Zhang F. and Lupski J.R. (2008) Mechanisms for human genomic rearrangements. *Pathogenetics,* **1**, 4.

24. Iliakis G., Wang H., Perrault A.R., Boecker W., Rosidi B., Windhofer F., Wu W., Guan J., Terzoudi G. and Pantelias G. (2004) Mechanisms of DNA double strand break repair and chromosome aberration formation. *Cytogenet. Genome Res.,* **104**, 14-20.

25. Lieber M.R. (2008) The mechanism of human nonhomologous DNA end joining. *J. Biol. Chem.,* **283**, 1-5.

26. Lee J.A., Carvalho C.M. and Lupski J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell,* **131**, 1235-1247.

27. Carvalho C.M., Zhang F., Liu P., Patel A., Sahoo T., Bacino C.A., Shaw C., Peacock S., Pursley A., Tavyev Y.J. *et al.* (2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum. Mol. Genet.,* **18**, 2188-2203.

28. Fattah F., Lee E.H., Weisensel N., Wang Y., Lichter N. and Hendrickson E.A. (2010) Ku regulates the non-homologous end joining pathway choice of DNA double-strand break repair in human somatic cells. *PLoS Genet.,* **6**, e1000855.

29. Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature,* **409**, 860-921.

30. Zhang F., Potocki L., Sampson J.B., Liu P., Sanchez-Valle A., Robbins-Furman P., Navarro A.D., Wheeler P.G., Spence J.E., Brasington C.K. *et al.* (2010) Identification of uncommon recurrent potocki-lupski syndrome-associated duplications and the distribution of rearrangement types and mechanisms in PTLS. *Am. J. Hum. Genet.,* **86**, 462-470.

31. Park S.S., Stankiewicz P., Bi W., Shaw C., Lehoczky J., Dewar K., Birren B. and Lupski J.R. (2002) Structure and evolution of the smith-magenis syndrome repeat gene clusters, SMS-REPs. *Genome Res.,* **12**, 729-738.

32. Hastings P.J., Ira G. and Lupski J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.,* **5**, e1000327.

33. Kloosterman W.P., Guryev V., van Roosmalen M., Duran K.J., de Bruijn E., Bakker S.C., Letteboer T., van Nesselrooij B., Hochstenbach R., Poot M. *et al.* (2011) Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.,* **20**, 1916-1924.

34. Pavlicek A., Gentles A.J., Paces J., Paces V. and Jurka J. (2006) Retroposition of processed pseudogenes: The impact of RNA stability and translational control. *Trends Genet.,* **22**, 69-73.

35. Chanda B., Asai-Coakwell M., Ye M., Mungall A.J., Barrow M., Dobyns W.B., Behesti H., Sowden J.C., Carter N.P., Walter M.A. *et al.* (2008) A novel mechanistic spectrum underlies glaucoma-associated chromosome 6p25 copy number variation. *Hum. Mol. Genet.,* **17**, 3446-3458.

36. 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.

37. Chapter 2, 7, 8, 14 and 19. In: Strachan T. and Read A.P. (2004) Human molecular genetics 3. Third edition. New-York: Garland Science.

38. www.ensembl.org (24/06/2011)

39. www.ncbi.nlm.nih.gov (24/06/2011)