Genomic structural variations and their role in carcinogenesis

Natalia Sánchez Collantes. Supervisor: Wouter de Laat
Cancer, Stem Cells and Developmental Biology MSc, Utrecht University.

Layman's Summary

Cancer is formed when one cell starts dividing in an uncontrolled manner. The acquisition of this cancerous state is the result of the accumulation of mutations that activate or disrupt genes involved in cell proliferation (oncogenes) or cell death (tumor-suppressor genes) signaling, respectively. With the rapid development of biomolecular technologies, one type of mutations has recently gained significance in cancer, named structural variations (SVs). SVs consist of genomic rearrangements that disrupt the structural organization of the DNA and can induce changes in the expression of the affected genes. SVs are recurrent in cancer genomes and can be used in cancer sub-type diagnosis, prognosis and as targets for novel personalized medicine. This review discusses the current knowledge on SVs, regarding their formation, functional consequences on oncogenesis, detection methods and their relevance in the clinic.

Throughout different cellular processes, breaks of both strands of the DNA may occur (DSBs). In order to overcome this, cells harbour DNA repair pathways that bring together the break ends and repair this damage. Homologous repair (HR) is the only pathway that repairs the break without any errors, whereas non-homologous end joining (NHEJ) and some variants of these two repair pathways can introduce deletions or duplications at the site of the break or even re-join the wrong DSB ends, generating inversions or translocations. These rearrangements are called SVs and have been found in many cancer types, where they can misregulate the expression of cancer-related genes by disrupting their sequence, altering their copy number or even juxtaposing two coding regions, resulting in the synthesis of a novel oncogenic protein. Most of the DNA consists of intergenic regions that harbour sequences important for the regulation of gene expression, such as enhancers. They are needed for the activation of target genes, which is accomplished through their interaction with the gene promoter (i.e. the sequence upstream of the gene, at which transcription starts). Duplications of enhancers can lead to the over-expression of oncogenes. SVs can also trigger the interaction between an enhancer and a different promoter, this is called enhancer hijacking and is a common cause of oncogene activation in cancer. Another form of regulation of gene expression is through epigenetic modifications: the chemical modification of DNA and histones, the proteins around which the DNA wraps to be packaged in the nucleus. SVs have been shown to alter the epigenetic marks of the genome by disrupting or activating genes coding for epigenetic modifiers or by re-positioning regions of the genome with different epigenetic marks.

The specificity of different SV patterns across cancer subtypes and their relation to cancer prognosis in certain cases makes the detection of SV highly valuable for cancer research, diagnosis and personalized medicine. Nowadays, next-generation sequencing (NGS) technologies, focused on deciphering the DNA sequence and comparing it to a reference genome in order to find differences, are capable of identifying all types of SVs and have been implemented in the clinic. However, NGS has some limitations such as the detection of SV breakpoints with high resolution in repetitive regions of the genome. In order to overcome some of the limitations of NGS, different approaches are being developed, although some remain too expensive for their implementation in the clinic, such as the most promising long-read sequencing.

Genomic structural variations and their role in carcinogenesis

Natalia Sánchez Collantes. Supervisor: Wouter de Laat
Cancer, Stem Cells and Developmental Biology MSc, Utrecht University.

Abstract

Rearrangements of the genome known as structural variations (SVs) that amplify, delete or reorder chromosomal material are an important source of genomic diversity across humans. Nowadays, with the recent improvements of sequencing technologies, SVs are conceived as a hallmark of cancer, and the molecular mechanisms underlying their oncogenic potential are being uncovered. SVs can alter cancer-related gene expression by disrupting the gene sequence, inducing copy number variations or forming fusion genes. Recently, 3D-genome organization has been shown to play an important role in the regulation of gene expression and the formation of genomic rearrangements. Disrupting genome structural organization, mainly topologically associating domains (TADs), can also trigger the activation of oncogenes and the onset of cancer. This review provides general insights into the current knowledge on SVs, including their formation, their significance in carcinogenesis, state-of-the-art detection methods and their relevance in the clinic, highlighting their potential implementation in early diagnosis and personalized medicine.

Different forms of variations in the human genome explain the genetic diversity among individuals. Approximately until 2004, with the use of traditional molecular biology technology, these genomic alterations were thought to predominantly comprise small variations such as single nucleotide polymorphisms (SNPs), small insertions or deletions polymorphisms (indels) and repetitive elements, such as mini- and microsatellites, whereas larger structural variants constituted rare events. With the exponential development of sequencing technology, now it is possible to perform genomic structural variant (SV) analysis with higher resolution. This has revealed the significant contribution of SVs to inter-individual genetic variability and certain disorders, making up the majority of varying nucleotides among human genomes[1].

SVs consist of genomic rearrangements that can be intra-chromosomal or translocations between chromosomes, and they often originate from double strand breaks (DSBs) improperly repaired by error-prone DNA repair pathways. SVs can also result from retrotransposition or insertions of external DNA elements[2]. Nowadays, it is known that SVs, including translocations, inversions, insertions, copy number variations (CNVs) caused by deletions and duplications, and complex SVs, such as chromothripsis, are the greatest cause of alterations in gene expression. Additionally, SVs are recognized as hallmarks in tumorigenesis[3]. Genomic rearrangements can occur at the coding sequence of cancer-related genes, resulting in the activation of oncogenes, inactivation of tumor suppressor genes or the creation of fusion genes[4], or they can alter the transcription regulatory landscape of genes by placing new enhancers or other regulatory DNA sequences within the reach of genes[5].

The detection of cancer-driver SVs at premalignant stages provides an opportunity to recognize high-risk populations and prevent the onset of cancer by patient's tracking and anticancer therapies. Furthermore, uncovering the aetiology and clinical outcome of cancer-related SVs helps in discovering novel biomarkers for cancer subtypes and prognosis, as well as in developing new therapeutic targets for personalized medicine.

Next-generation sequencing techniques have enabled the identification of distinct patterns of genomic rearrangements that were missed by conventional molecular biology technologies. Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium collaborated with The Cancer Genome Atlas (TCGA) to compare 2,658 whole cancer genomes with their corresponding healthy tissues across 38 tumour types. With the combination of whole-genome sequencing and transcriptomic data the authors reported new patterns of SVs in both coding and non-coding sequences, genomic breakpoint cluster regions and the effects of SVs on gene expression, some of which will be discussed here[6].

Nonetheless, the formation of SVs and their oncogenic potential remain poorly understood, with SV detection being the bottleneck. The current standard whole-genome high throughput sequencing (HTS) uses short reads. While this enables the direct identification of SNPs, SV detection requires computational inference, as

complex rearrangements with multiple breakpoints span a myriad of short reads. Short reads also do not allow the detection of breakpoints in repetitive regions, where DSBs occur more frequently. Long read sequencing (LRS) has been recently proposed as an alternative to more accurately identify SVs[7]. However, LRS technology currently still is too expensive for clinical application. Hence, further research is needed for the development of novel platforms to efficiently identify SVs at a whole-genome level, which is needed to gain further insight into their biological relevance and the molecular mechanism underlying their formation.

In this review, I provide an update on the current knowledge on SVs in the human genome. I will discuss the different types of SV, the proposed mechanisms for their formation and how these rearrangements can lead to the onset of cancer. Finally, I will also evaluate the detection methods currently in use and the relevance of SV detection in the clinic.

1 Types of structural variations

Molecular karyotyping, together with recent next-generation sequencing, has disclosed the large complexity of human genome structure. Genomic nucleotides are subject to continuous alterations. The most impactful changes to the primary nucleotide sequence come from rearrangements of the genome, also known as structural variations (SVs), that amplify, delete or reorder chromosomal material, ranging from single genes to entire chromosomes[8].

Chromosomal rearrangements can be unbalanced, termed copy number variations (CNVs), when they alter the copy number of chromosomes or chromosomal regions and thereby change the diploid status of the genome. When they have no effect on the amount of genetic material, they are termed balanced rearrangements. CNVs can arise from *insertions* of sequences into the genome, leading to copy number gain, *deletions*, which result from removal of genetic material from the genome, or *duplications* that result from the copy of a specific region of the genome and its insertion either contiguously (*tandem duplications*) or separated from the site of the original sequence. On the other hand, balanced rearrangements may originate from *inversions* of intra-chromosomal segments without the loss or gain of sequences, and *translocations* whereby segments are exchanged between or within chromosomes without the loss or gain of sequences (Figure 1). In some cases, inversions and translocations may also be accompanied by the gain or loss of genetic material, leading to unbalanced rearrangements[1].

Individual rearrangements can be combined in more complex ones that result from the improper re-joining of multiple simultaneous DSBs occurred in one or more chromosomes. A particular type of

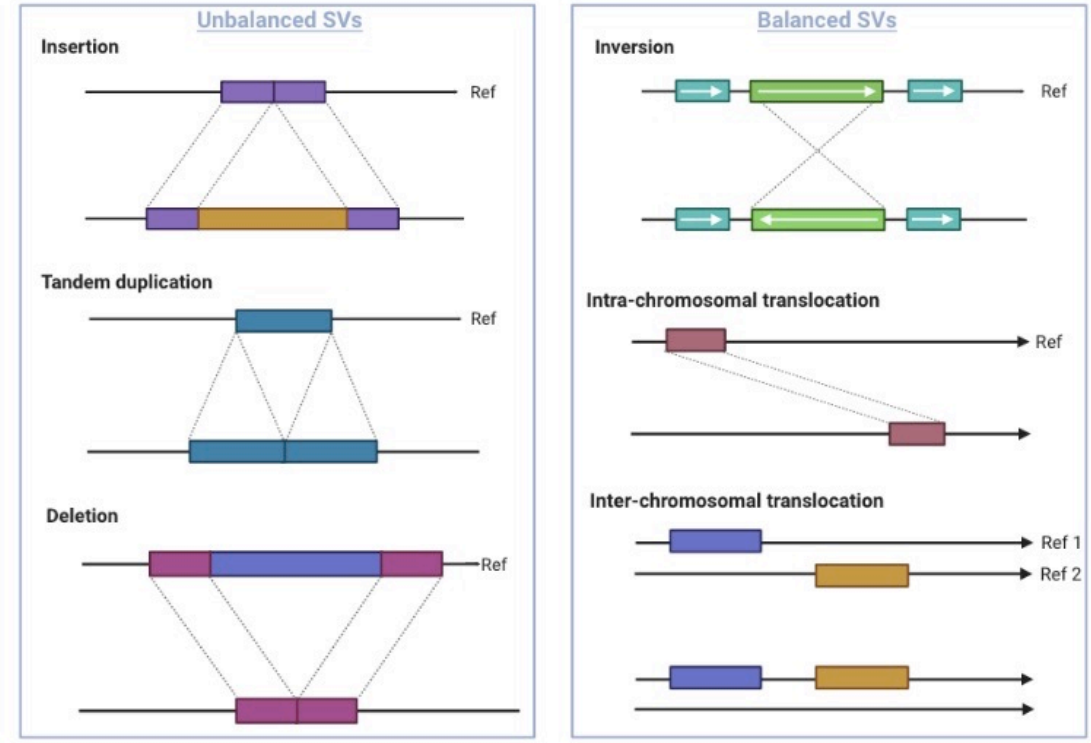

Figure 1. Different types of simple structural variations. On the left, unbalanced rearrangements including insertions, duplications and deletions. On the right, balanced rearrangements, including inversions and translocations, which can be intra- or inter-chromosomal. In some cases, inversions and translocations may also result in unbalanced SVs. Ref: reference. SVs: Structural variations. Image made in BioRender.com.

complex rearrangement is c*hromothripsis*, during which a series of chromosome breaks and the rearrangements of the resulting fragments occur at the same time, involving one or a few chromosomes. Chromothripsis was discovered in 2011 and was seen to be present in at least 2-3% of all human cancers with a particular prevalence in osteosarcomas[9]. Another important complex structural variation is *chromoplexy*, which also involves the improper re-joining of co-occurring DSBs but in different chromosomes[10]. This acquisition of massive genomic structural variants in a single catastrophic event at early oncogenic stages has challenged the traditional understanding of cancer development as a successive accumulation of somatic mutations throughout life[11].

Structural variations can occur in either somatic or germ cells and are linked to the development of cancer and disease. Germline SVs are considered a common cause of congenital diseases, such as the well-known Down's syndrome (trisomy of chromosome 21)[8], or other congenital abnormalities associated to rare CNVs, which include diabetes (17q12 deletion) and heart failure (1q21.1 deletion)[12]. On the other hand, as mentioned above, the accumulation of somatic SVs at high densities is a hallmark of many human cancers[4].

2 Spatial organization of the 3D genome

The onset of cancer and disease caused by SVs is explained by the disruption of gene function and gene-expression regulatory elements, but also by the alteration of the 3D structure of the human genome, which controls the regulatory landscapes sensed by genes. In order to be aware of the pathogenic mechanism of SVs it is essential to comprehend the basics of the human genome organization in the 3D space of the nucleus.

2.1 Chromatin architecture

Traditional fluorescent *in situ* DNA hybridization (FISH) and microscopy experiments have unveiled sub-nuclear compartments occupied by different forms of chromatin assembly, known as heterochromatin and euchromatin. Heterochromatin can be further divided into constitutive (cHet) or facultative (fHet) heterochromatin. cHet refers to compact, inaccessible chromatin present pan-cellularly at repetitive sequences such as the centromeres and telomeres that play an indispensable role in preserving genome stability[13]. fHet is found at developmentally expressed genes and their surrounding sequences, in cells that require them to be silenced[14]. Euchromatin corresponds to the open form of the chromatin, is gene-rich and harbours transcriptionally active genes[15]. The chromatin status, thus, dictates whether a given sequence is found in an active or inactive nuclear compartment.

Chromosomes occupy specific nuclear territories and harbour different regions of open/closed chromatin, known as compartments A and B. Compartment A correlates with open, transcriptionally active euchromatin and is located at the inside of the nucleus. On the other hand, compartment B harbours compact, transcriptionally silenced heterochromatin and is located towards the periphery of the nucleus forming lamina-associated domains (LADs) or at nucleoli as nucleolar-associated domains (NADs)[8].

Chromatin is built through DNA wrapping around nucleosomes. Nucleosomes are formed by an octamer of histone proteins surrounded by ~147 bp of DNA[16]. The conserved residues at histone tails are hotspots for modifications by acetylation, methylation and ubiquitination (mainly at lysine (Lys) residues) and phosphorylation (mainly at serine (Ser) residues). Interestingly, histone modifications are known to control chromatin architecture and gene expression. Methylation of histone H3 lysine 9 (H3K9me3) and methylation of histone H3 at lysine 27 (H3K27me3) are common epigenetic marks of heterochromatin and correlate with gene silencing[17,18]. Acetylation of histones H3 and H4 and H3K4 methylation (H3K4me1 and H3K4me3) are enriched in euchromatin, particularly in enhancers and promoters of active genes, and are largely associated with active transcription[18]. Modifications in histone epigenetic marks can lead to alterations in gene expression and the onset of cancer and disease. For instance, it has been shown that *GABARAPL1*, an autophagy-related gene, is epigenetically silenced through decreased acetylation of its promoter by histone deacetylases (HDACs) in breast cancer cells[19].

The nucleosome ribbon can be further folded by looping, which enables the interaction of enhancers and promoters distant in the linear genome and controls gene expression[20].

2.2 Topologically associating domains (TADs)

The development of three-dimensional chromosome conformation capture technology (3C)[21], which detects the spatial proximity between DNA sequences, and its higher-throughput derives, such as 4C[22] or Hi-C[23], have broadened our understanding of genome organization in the nuclear space.

High-throughput 3C methods have disclosed the presence of self-interacting chromosomal segments separated by low-interaction boundaries, known as topologically associating domains (TADs)[24]. TADs consist of looped domains formed by loop extrusion, during which cohesins form progressively larger loops until the encounter of boundary elements, such as CTCF present at TAD boundaries. Within TADs, genes and their regulatory elements interact by loop formation giving rise to sub-TADs. TADs regulate gene expression by concentrating regulatory interactions within and preventing them across domains. Epigenetic studies of TADs have shown that loci within one contact domain harbour the same histone marks, which correlate to the corresponding chromatin conformation of the nuclear compartment A or B it occupies[25] (Figure 2).

All this proves the indispensable role of TADs in the regulation of gene expression by promoting spatial proximity between promoters and enhancers and preventing their interaction across different domains. This is in line with previous evidence showing aberrant gene expression upon disruption of TAD boundaries, which has been linked to the onset of cancer[5].

3 Formation of genomic structural variations

SVs result from genomic double strand breaks (DSBs) that are unfaithfully repaired by rearrangements between different genomic regions[2]. Breaks occurring either inside genes or in intergenic regions within TADs may lead to disruption of genomic structural organization, affecting gene function and regulation of gene expression, respectively.

3.1 Sources of DNA breakage

DSBs can result from exogenous agents, such as ionizing radiation of X-rays and γ-rays, or occur during endogenous processes, including replication stress and telomere dysfunction[26], and, in lymphoid cells, recombination and hypermutation, as discussed below.

During DNA replication different replication forks open along the genome to copy the DNA. *Replication stress* refers to the stalling of these forks due to the lack of replication components, such as dNTPs, chromatin packaging elements and/or functional replication proteins, or due to their encounter with unrepaired DNA damage, with secondary structures commonly found in repetitive sequences or with highly-compacted DNA regions present in heterochromatic telomeres and centromeres. When the resolution of the hindered forks fails, the replisome complex is inactivated and DSBs arise at the stalling site[27].

Other sources of DSBs formation during replication and also transcription are *topoisomerases*. Topoisomerases act to relax DNA topology at supercoiled regions formed during replication, transcription and DNA compaction, and they do so through the transient formation of DSBs. In the worst scenario, high topoisomerase activity, e.g. during chromosome condensation and active transcription, may lead to an accumulation of unrepaired DSBs that form genomic rearrangements, as shown for *MLL* translocations in leukemia cells[28].

Lymphoid cells are special in that they naturally undergo *recombination and hypermutation* events at the immunoglobulin and T-cell receptor genes, in order to create a diverse repertoire of antigen

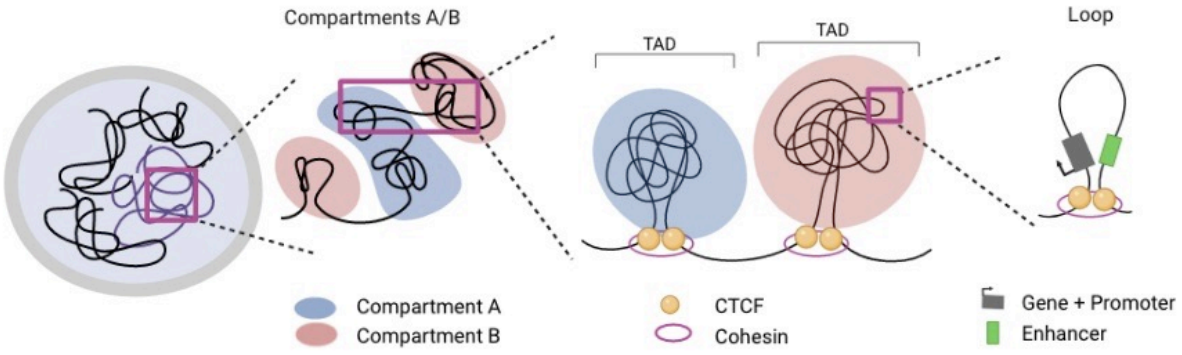

Figure 2. 3D-genome organization in TADs. Along chromosomes different compartments A and B can be distinguished based on the chromatin compaction levels. Within each compartment there are regions with high interaction frequencies known as topologically associating domains (TADs). CTCF and cohesin proteins are present at TAD boundaries. Promoters and enhancers in the same TAD can interact by loop formation (sub-TAD), through cohesin-mediated loop extrusion. Image made with BioRender.com. Modified from Wang et al., *Epigenetics and Chromatin*, 2020.

receptors and antibodies. While recombination-activating gene (RAG) endonuclease is involved in the recombination of V(D)J segments of antigen receptors in lymphocytes, activation-induced cytidine deaminase (AID) accounts for the hypermutation and class-switch recombination processes that occur at Ig genes during the activation of B lymphocytes. Both enzymes, either directly or indirectly, induce DSBs at their target sites. However, they can also have off-target effects and trigger undesired genomic rearrangements, such as genomic translocations present in some lymphomas, which relate to poor prognosis and tumor relapse[29,30].

Aberrant DSBs can also result from *telomere dysfunction*. Telomeres correspond to tandemly repeated sequences of single-stranded DNA (ssDNA) located at the ends of our chromosomes. Certain telomeric proteins recognize these sequences and protect them from DNA damage responses[31]. Dysfunctional telomere capping leads to the recognition of telomeres as DSBs and the subsequent activation of non-homologous end-joining (NHEJ) repair pathway and its derivatives, which join non-homologous or sister chromatids and generate dicentric chromosomes. Dicentric chromosomes can lead to DSBs during anaphase, when they undergo stretching forces caused by the attachment of both centromeres to opposite spindle poles. The asymmetric inheritance of the resulting fragments effectively results in CNVs in the daughter cells. Furthermore, if the inherited chromosome remains uncapped it can enter successive breakage-fusion-bridge (BFB) events that result in the accumulation of aberrant chromosomal rearrangements, including non-reciprocal translocations, aneuploidies or even gene amplifications if fusions between sister chromatids occur[32]. The Protection of Telomeres 1 (POT1) protein is an essential component of the shelterin complex that protects telomere ends from DNA repair responses. About 2.9% of pan-cancer genomes have been shown to harbor oncogenic *POT1* mutations, being more prevalent in angiosarcomas and cutaneous squamous cell carcinoma[33]. This highlights the importance of telomere function in preserving genome integrating and preventing the onset of cancer.

In addition to DNA breaks leading to single rearrangements, there are also more complex structural variations, such as chromothripsis and chromoplexy. Their origin is believed to be the consequence of multiple DSBs that occurred at the same time. Little is known about the causative mechanism underlying these events. Some research puts forward degradation of chromatin bridges derived from dicentric chromosomes caused by telomere dysfunction[32] and micronuclei[34] formed by miss-segregation errors as the origin of chromothripsis. For chromoplexy, the underlying biological process remains unknown. Chromoplexy basically involves simultaneous translocation events among multiple chromosomes. Interestingly, previous evidence shows that chromoplexy breakpoints are clustered at actively transcribed and open chromatin regions along the human genome[35], in line with the known DSBs vulnerability of actively transcribed regions, discussed later in this review.

3.2 Improper double-strand break repair leads to structural variations

SV formation following DSBs occurs through undesired recombination and re-joining events during DSB repair, which is carried out by different DNA repair pathways:

Homologous recombination (HR) is the most faithful DSBs repair pathway and plays its role during the S/G2 phase of the cell cycle. It consists of the recombination between sister chromatids to repair DSBs in an error-free manner due to the almost-complete similarity between sister chromatids. However, *non-allelic homologous recombination (NAHR)* pathway is an alternative HR activated in the absence of sister chromatids throughout meiosis or when DSBs occur at repeated regions. Hence, the recombination step during NAHR occurs between paralog sequences, leading to SVs formation[36] (Figure 3A).

Non-homologous end-joining (NHEJ), the most commonly used DNA repair pathway to deal with DSBs, is activated almost at any phase of the cell cycle and consists on the rapid re-joining of the DSB ends that share little or no homology. This process is error-prone and introduces insertions or deletions at the site of the junction. Hence, SV patterns induced by this mechanism can be detected as deletions or insertions at junction points with no or little homology. Indeed, WGS analysis of the PCAWG data shows that the vast majority of the identified SVs do not present homology at the breakpoint junctions and, thus, derive from NHEJ. An alternative error-prone NHEJ is *microhomology-mediated end-joining (MMEJ)* which repairs DSBs by rearranging sequences with small homologies upon deficiency for HR and NHEJ. This mechanism, unlike NHEJ, involves the resection of the 5' strand to then anneal the resulting 3' overhangs of

proximal small homologies. Similar to MMEJ is the mechanism of *single-strand annealing (SSA)*, which anneals homologous 3' single-stranded ends exposed after more extensive end resection events. In both cases, resulting not-annealed single-stranded DNA is digested, which often also causes big deletions[35]. Most inter-chromosomal translocations with microhomologies at the junction sites are explained by MMEJ activity, specially in cells deficient for the classical NHEJ[37] (Figure 3B).

Aberrant rearrangements between distant chromosomes or chromosomal regions can result from errors occurring during the repair of DSBs derived from replication fork stalling. The repair pathways involved in this process are*: Fork stalling and template switching (FoSTeS)*, activated when the hindrance is caused by structural obstacles, and *microhomology-mediated break-induced repair (MMBIR)*, when the fork stalling is due to breaks in the template strand[38]. During these pathways, the lagging strand switches template by microhomology annealing with a different fork, extends some base pairs and re-anneals with the original template strand. The two forks, although close in the replication space, may be distant in the linear genome. Furthermore, the invasion of new templates can occur several times between different forks, leading to the juxtaposition of different chromosomal segments and additional duplications, triplications or even inversions if the leading strand anneals with the lagging strand[39] (Figure 3C).

More recently, *non-coding RNAs (ncRNAs)* have been ascribed a role in DSBs repair. Damage-induced long non-coding RNAs (dilncRNAs) are

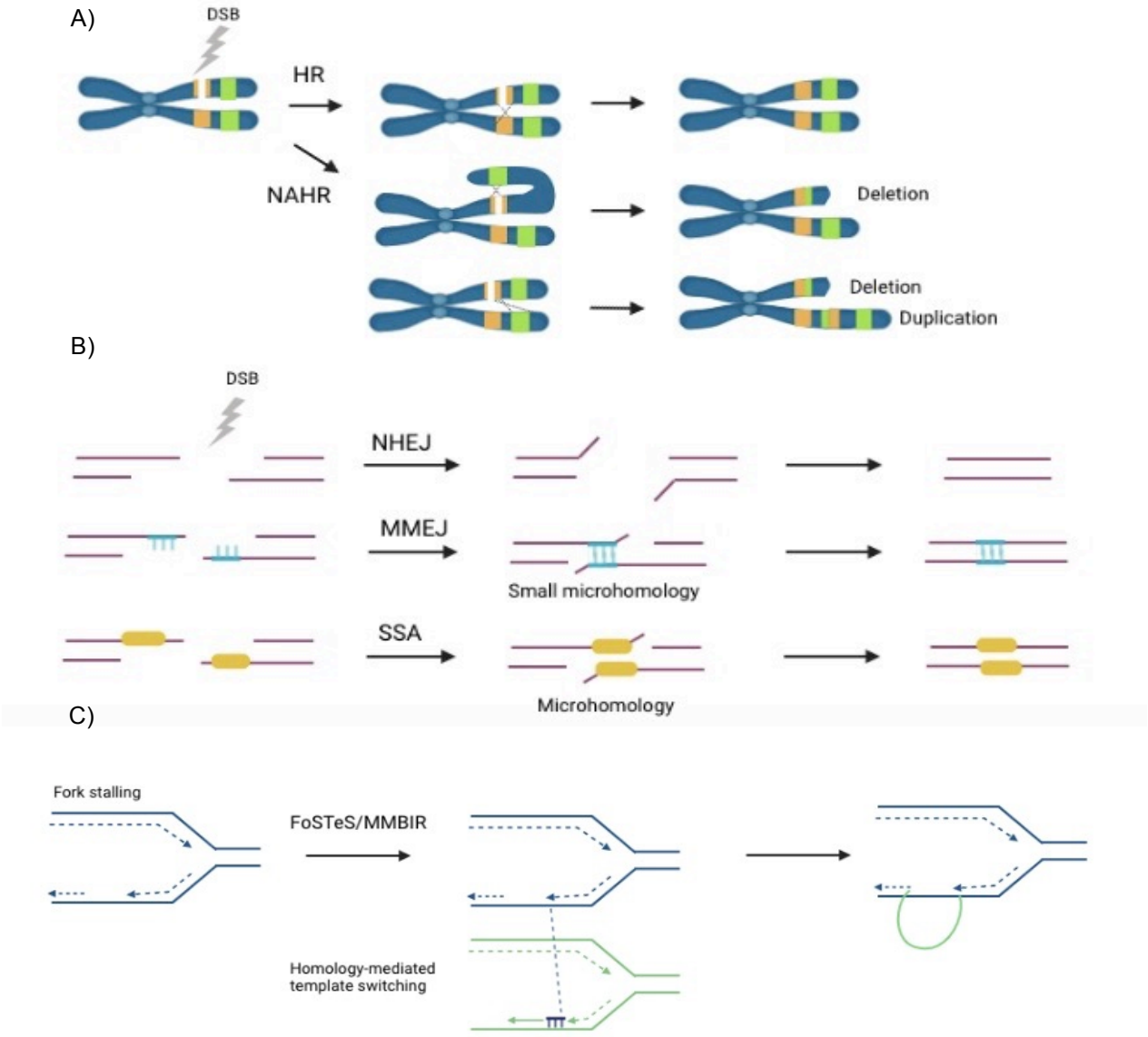

Figure 3. Mechanisms for SV formation by error-prone DSB repair. DSBs repair by HR, NHEJ and their alternative repair pathways. During replication, fork stalling can be a source of DSBs that, upon repair by FoSTeS or MMBIR, lead to the formation of genomic rearrangements. Image made with BioRender.com.

transcribed at DSB breakpoints by RNA polymerase II that binds to the Mre11-Rad50-Nbs1 (MRN) complex present at DSB ends. These dilncRNAs can be subsequently processed into small ncRNAs, known as DNA-damage RNAs (DDRNAs), which localize at the break site by pairing with new emerging dilncRNAs at the break junctions and activate the recruitment of DSB-repair proteins[40]. This localization of ncRNAs at DSB sites to promote DSBs repair may be involved in the generation of chromosomal rearrangements by ncRNA-mediated DNA-DNA interactions that may provide spatial proximity between distant chromosomal regions. Indeed, previous evidence shows that RUNXOR, a lncRNA upregulated in acute myeloid leukemia (AML), may be involved in the formation of t(8;21), a translocation found in 30-40% of AML, through its interaction with the *RUNX1* gene located in chromosome 21 and the *ETO* gene in chromosome 8[41]. Additionally, some of the breakpoint clusters in these two genes associate with topoisomerase II cleavage sites[42], which highlights the above-mentioned role of this enzyme in DSBs formation and subsequent chromosomal rearrangements.

3.3 Factors affecting DNA breakage: 3D genome spatial organization and epigenetics

Unrevealing what features of the DNA increase the susceptibility of certain genomic regions to undergo structural rearrangements, or be positively selected during clonal selection in cancer development, can provide more insight into the formation of SVs. Putative genomic regions with higher chromosomal-rearrangement risk can be studied by analysing breakpoints of different SVs. Recent research points towards epigenetics and genomic spatial organization as two main influencers of SV vulnerability.

For instance, previous evidence shows that common translocations breakpoints in malignant B-cells correlates with CTCF and cohesin binding regions. Interestingly, DSBs at these regions were mainly caused by topoisomerase 2B[43]. Disruption of loop-anchor binding sites and TAD boundaries may lead to changes in the enhancer-promoter contact landscape, altering gene expression and potentially causing disease.

Additionally, it is known that AID targets ssDNA to induce DSBs during somatic hypermutation and class switch recombination in B lymphocytes. ssDNA is present at actively transcribed regions, which demand the relaxation of chromatin to the most open conformation in order to provide access for transcription proteins. Hence, it is not surprising that AID-induced off-target DSBs are enriched at transcriptionally active genomic regions of B lymphocytes[44].

Changes in the epigenome can also influence the frequency of genomic rearrangements. DNMT3B is a *de novo* methyltransferase involved in DNA methylation during early development. Mutations in DNMT3B that trigger hypomethylation in constitutive heterochromatin can lead to immunodeficiency-centromeric instability-facial anomalies syndrome 1 (ICF1). ICF1 is characterized by increased chromosome rearrangements in heterochromatic regions of chromosomes 1, 9 and 16 in lymphocytes[45]. In line with this, a correlation between DNA methylation levels and the frequency of structural rearrangements was suggested when it was found that a considerable enrichment of SVs in hypomethylated regions of the human genome accumulated throughout evolution since the branching of chimpanzee[46].

3.4 *Other sources of structural variations*

Besides DSBs and their repair by error-prone DNA repair pathways there are other mechanisms of induction of genomic SVs, such as transposition events and the integration of external DNA (e.g. by viral infections).

Transposable elements

45% of the human genome consists on transposable elements (TEs), repetitive sequences that can jump to a different region of the genome by ‘copy-and-pasting’ (retrotransposition) or ‘cut-and-pasting’ (transposition). Although most of these sequences are currently inactive, there are some important TEs to bear in mind when talking about SVs. The most important self-sufficient mobile element is Long INterspersed Element-1 (L1). L1 retrotransposition entails its reverse transcription, the synthesis of the corresponding cDNA and its insertion in a different region of the genome[47]

The deleterious consequences of L1 retrotransposition comprise the amplification of L1 sequences by the copy-pasting of its coding sequence during transposition and, in some cases, of their adjacent genes or regulatory elements, the disruption of cancer-related genes by the insertion of these elements into their ORFs or regulatory elements and the generation of other genomic rearrangements by the L1-cDNA invasion of pre-existing DSBs. Interestingly, L1 retrotransposition events are frequently present in cancer genomes.

Indeed, L1-related rearrangements are found in about 50% of the Pan-cancer project genomes[35,26].

Insertion of external DNA elements

Apart from rearrangements among internal sequences in the human genome, SVs can derive from external DNA elements that integrate into the genome, e. g. during viral infections. Different viruses have been linked to cancer through the integration of their genetic material into genomic regions that harbour cancer-driver genes, or through the insertion of viral regulatory elements near proto-oncogenes.

The most widely studied is Human Papillomavirus (HPV), which is known to cause invasive cervical cancer (ICC) progression through the integration of its DNA sequences nearby cancer-related genes of the host genome. A recent multi-omics approach focused on the study of ICC integration targets that can affect the expression of ICC-related genes, found one integration site adjacent to *BNC1* enhancer RNA coding sequence that leads to increased *BNC1* expression levels in ICC[48].

Another example is the Epstein-Barr virus (EBV)-associated malignancies. Recently novel insight on EBV genetic material integration into the host genome has been provided. Most integration sites in EBV-related nasopharyngeal carcinomas (NPC) and gastric cancers are adjacent to tumor suppressor genes or their regulatory sequences, probably triggering the downregulation of these tumor suppressor genes. Interestingly, these breakpoints are skewed towards DNA-damage vulnerable sites and present microhomology, suggesting microhomology-mediated DNA repair as a putative mechanism for viral genome integration[49].

Besides that, insertional mutagenesis and chromosomal instability caused by Hepatitis B Virus (HBV) genomic integration are known to be a common cause of hepatocellular carcinomas (HCC). WGS of HBV integration sites in HCC unravelled that HBV integration sites present in HCC but not in control samples showed a tendency towards telomere regions, which suggests a causative role for HVB integration in telomere dysfunction and chromosome instability[50].

4 Consequences of structural variations on carcinogenesis

SVs can occur either inside genes, giving rise to gene disruption, CNVs or fusion genes, or in intergenic regions within or across TADs, changing the regulatory landscape of proximal genes. In this context, SVs may have functional consequences on carcinogenesis by, directly or indirectly, affecting cancer-related genes. SVs have been described as the underlying cause of the activation or inhibition of oncogenes and tumor suppressor genes, respectively, in certain cancer genomes.

4.1 Tumor suppressor genes disruption

Translocations and insertions occurring in tumor suppressor genes can interrupt these sequences, inactivate the gene and subsequently cause cancer. A recent whole-genome analysis of 101 metastatic castration-resistant prostate cancer (mCRPC), the lethal progression of prostate cancer, identified prevalent SVs affecting tumorigenesis-and-progression related genes. For instance, the sequence and promoter of *PTEN*, a tumor suppressor gene that codifies for a negative regulator of the AKT/PKB signaling pathway that leads to cell survival, proliferation, angiogenesis, and migration, was frequently disrupted by translocations or inversions in mCRPC. The number of affected alleles correlated with the levels of the corresponding mRNA, which suggests that a mono-allelic mutation is sufficient to induce an effect on gene expression[3].

Another commonly inactivated tumor suppressor gene across cancer subtypes is *RB1,* which encodes a negative regulator of the cell cycle. In PCAWG cancer genomes, *RB1* is often hit by templated insertions (insertions of duplicated regions from within or outside the gene) that disrupt the *RB1* sequence and render a non-functional transcript[4].

In pediatric osteosarcoma, characterized by high levels of chromothripsis, non-coding translocations have been shown to interrupt the first intron of *TP53*[51], a tumor suppressor gene involved in signaling pathways that ultimately induces cell-cycle arrest, apoptosis or senescence.

4.2 CNVs of oncogenes and tumor suppressor genes

CNVs of oncogenes and tumor suppressor genes are mainly the result of duplications, deletions or templated insertions. Many PCAWG liver cancers show overexpression of *TERT*, an oncogene involved in telomeres length maintenance, caused by templated insertions, which duplicate the entire oncogene and juxtapose it with other genomic regions[4]. Chromothripsis is another source of oncogene amplification in cancer. For instance, chromothripsis involving different chromosomes

induces *TERT* and *MDM2* co-amplification in 20% of PCAWG liposarcomas[6]. On the other hand, tumor-suppressor-gene copy losses are mainly caused by deletions. In lung adenocarcinomas, *CDKN2A* and *PTEN* genes are the most significant focal regions of deletions causative of the corresponding gene loss[51].

lncRNAs may also be involved in the development of CNVs of cancer-related genes. For instance, tandem duplications affecting lncRNAs upstream *MYC* lead to *MYC* copy number gain and higher *MYC* mRNA levels in mCRPC[3].

One cancer-related gene can undergo different rearrangements across cancer types. For example, *MYC* is amplified as a result of tandem duplications in breast cancer and translocations or chromoplexy with *IGH* gene in lymphomas, among others[4]. However, some cancer genes acquire oncogenic potential only with specific structural events, such as fusion genes.

4.3 Fusion genes

Specific genomic rearrangements may lead to the juxtaposition of coding regions to form what is known as fusions genes. These rearranged sequences encode chimeric proteins that play a driver role in tumorigenesis. The specific association between the fusion gene and the tumor phenotype increases the interest of fusion genes for their use as diagnostic biomarkers and in targeted therapy. Many targeted therapies against oncogenic chimeric proteins have shown a dramatic improvement of patient outcomes[52].

Before the development of deep-sequencing technologies. fusion genes were thought to be mainly caused by translocations of large chromosome segments. Among these early-detected fusion genes is *BCR-ABL1*, a product of translocation t(9;22) in chronic myelogenous leukaemia, which encodes a chimeric protein with abnormal tyrosine kinase activity[53,52].

The introduction of next generation sequencing allowed the detection of fusion genes arising from other types of SVs, such as intrachromosomal rearrangements rather than translocations between chromosomes. For example, an inversion in chromosome 12 has been shown to juxtapose *NAB2* and *STAT6* genes in solitary fibrous tumour. Contrary to *NAB2* known function, *NAB2-STAT6* fusion-gene product acts as an activator of *EGR1*-targeted transcription genes, whereas it has no effect on *STAT6* target genes[54].

Another important source of fusion genes are focal deletions, such as the 2.8Mb genomic loss between *TMPRSS2* and *ERG* genes that results in *TMPRSS2-ERG* fusion. *TMPRSS2-ERG* is present in about half of prostate cancers and relates to poor clinical outcome. In this particular case, the rearrangement leads to the overexpression of *ERG* by the placement of the oncogene downstream *TMPRSS2* androgen-responsive promoter[55], which ultimately triggers dedifferentiation, cell invasion and neoplastic transformation of prostate epithelium[33]. This abnormal gene expression as a consequence of fusion-gene-associated 'promoter swapping' is commonly found in cancer[52].

Besides the juxtaposition of two coding genes, it can occur that one non-protein-coding gene fuses with other coding regions resulting in the synthesis of a novel protein involved in the onset and progression of cancer. This is the case of the non-coding gene *PVT1* which recurrently fuses with *MYC* and *NDRG1*, among others, as a result of chromothripsis in medulloblastoma genomes[56].

Specific fusion genes are characteristic of many pediatric malignancies and seem to play a driver role in carcinogenesis, such as *EWS*-associated fusion genes (*EWS-FLI1, EWS-ERG*, among others) in Ewing sarcoma. However, fusion genes are less frequent, albeit considered major drivers, in adult cancer genomes, such as the already mentioned *TMPRSS2-ERG* in prostate cancer[51]. Interestingly, this suggests a differential molecular mechanisms for tumor initiation and progression in pediatric malignancies compared to those occurring in adult cancer.

4.4 Epigenetic alterations

In addition to aberrant expression levels of cancer-related genes, as well as synthesis of novel oncogenic proteins, SVs may also drive changes in the epigenome in cancer, which alters the 3D genome structure and triggers aberrant gene expression of cancer-related genes. SV-derived epigenetic alterations can occur through two main mechanisms: the aberrant expression of epigenetic factors or the rearrangements of genomic regions with different epigenetic marks.

One of the mechanisms by which SVs may have an effect on the epigenome is the disruption or overexpression of genes coding for epigenetic factors, such as DNA methyltransferases or histone modifiers. A common characteristic of cancer genomes is the global DNA hypomethylation, mainly in repetitive, coding and intronic regions, which contributes to the two main hallmarks of cancer: genomic instability and aneuploidy. Decreased methylation levels can be explained by

mutations disrupting genes that encode DNA methyltransferases (DNMTs), involved in the addition of methyl groups to cytosine bases in the DNA. On the other hand, tumor-suppressor gene silencing by the hypermethylation of CpG islands (genomic regions rich in CpG dinucleotides) in promoters is a major event in the onset of multiple cancers, and can occur by the overexpression of specific DNMTs[57].

Histone modifications are also involved in the regulation of gene expression. Unbalanced H3K36me2 and H3K27me3 antagonist histone marks are commonly present in a range of cancers. Accumulation of H3K36me2, a histone mark of active transcription, and its spread outside the active gene bodies caused by an overexpression of NSD2 histone methyl transferase, induces the contraction of H3K27me3 domains, which ultimately changes gene expression patterns. This overexpression of *NSD2* gene has been found as a result of a t(4;14) translocation between the *IGH* and *NSD2* loci in many multiple myeloma patients with poor clinical outcome[58]. How an imbalance between H3K36me2 and H3K27me3 histone marks may have an effect on alterations in gene expression in cancer is poorly understood. A recent study puts forward intra-TAD interaction modifications and A/B-compartment switch as putative mechanisms underlying this aberrant gene expression resulting from changes in histone epigenetic marks[59].

Changes in the epigenome upon SVs can also be explained by the rearrangements of genomic regions with different methylation levels. A recent study based on TCGA and PCAWG data, elucidates putative mechanisms for the altered methylation patterns present in cancer cells. In this study the authors show that many of the changes in DNA methylation are related to SV-derived rearrangements between differentially methylated regions[60]. This is in line with previous evidence putting forward a role for NHEJ, involved in SV formation, during DSBs repair in the alteration of DNA methylation seen in cancer[61]. *TERT* and *FASN* were among the most affected oncogenes by SV-associated decreased methylation and subsequent overexpression seen across TCGA cancer types[60].

4.5 Topologically associated domains disruption

The regulatory landscape of the genome is largely controlled by the formation of TADs, which delimit genes and *cis*-regulatory elements that interact with each other and promote their spatial proximity. Structural rearrangements that disrupt these domains have major consequences on gene expression and may trigger the onset of cancer and disease. SVs can either occur within or across TADs, repositioning TAD boundaries and reorganizing enhancer-promoter interacting groups.

Intra-TAD structural variations

Considering the role of TADs in gene expression, regulatory elements within a domain are then supposed to be able to interact with genes in the same domain and, thus, control their expression patterns[24]. SVs affecting regulatory elements within TADs may increase (duplications) or decrease (deletions) the enhancer dosage, resulting in the overexpression or downregulation of their target genes[8]. However, the deletion of one enhancer usually has minor effects on its target gene. This is explained by enhancer redundancy, which consists of the partial or complete overlap of the activity of different enhancers that cooperate to rigorously enhance tissue-specific gene expression. Highly transcribed genes are usually controlled by clusters of enhancers that act redundantly, known as 'super-enhancers'[62].

The acquisition of super-enhancers to strongly activate oncogenes is a common feature of cancer development and plays an indispensable role in preserving cancer-cell identity[63]. In this context, copy number variations that amplify the enhancers of oncogenes may lead to the overexpression of the target gene and the development of tumors (**Figure** 4A). A study focused on chromosomal rearrangements present in a subset of breast-cancer genomes has revealed multiple tandem duplications (>100kb) enriched in breast-tissue specific super-enhancers that regulate the expression of oncogenes, such as *MYC* and *ESR1*[64].

Another example of oncogene activation by enhancer CNVs is the overexpression of the androgen receptor gene (*AR*), present in 81% of mCRPC. Specifically, this was explained by an amplification of the enhancer upstream of the oncogene. About 44% of these amplifications were caused by tandem duplications in the regulatory element. Amplification of both the enhancer and the *AR* gene itself was also found in mCRPC genomes, albeit less frequently. These cases had significantly higher expression levels of *AR* compared to the genomes harbouring the SV only in the enhancer region, which suggests an additive effect of the amplification event[3].

Inter-TAD structural variations

3D chromatin organization can be altered by structural rearrangements across TADs that drive the reordering of TAD boundaries and/or enhancers into other chromosome compartments or genomic regions, changing tissue-specific gene-expression patterns and subsequently causing disease. The repositioning of an enhancer to a gene different from its usual target is known as enhancer 'hijacking' and has been studied in cancer since the implementation of karyotyping[8]. TAD disruption by inter-TAD SVs can take place through different mechanisms: TAD fusion, neo-TAD formation or TAD shuffling.

Large deletions embracing CTCF-binding sites at the boundaries fuse different TADs together. The removal of one or more TAD boundaries allows for

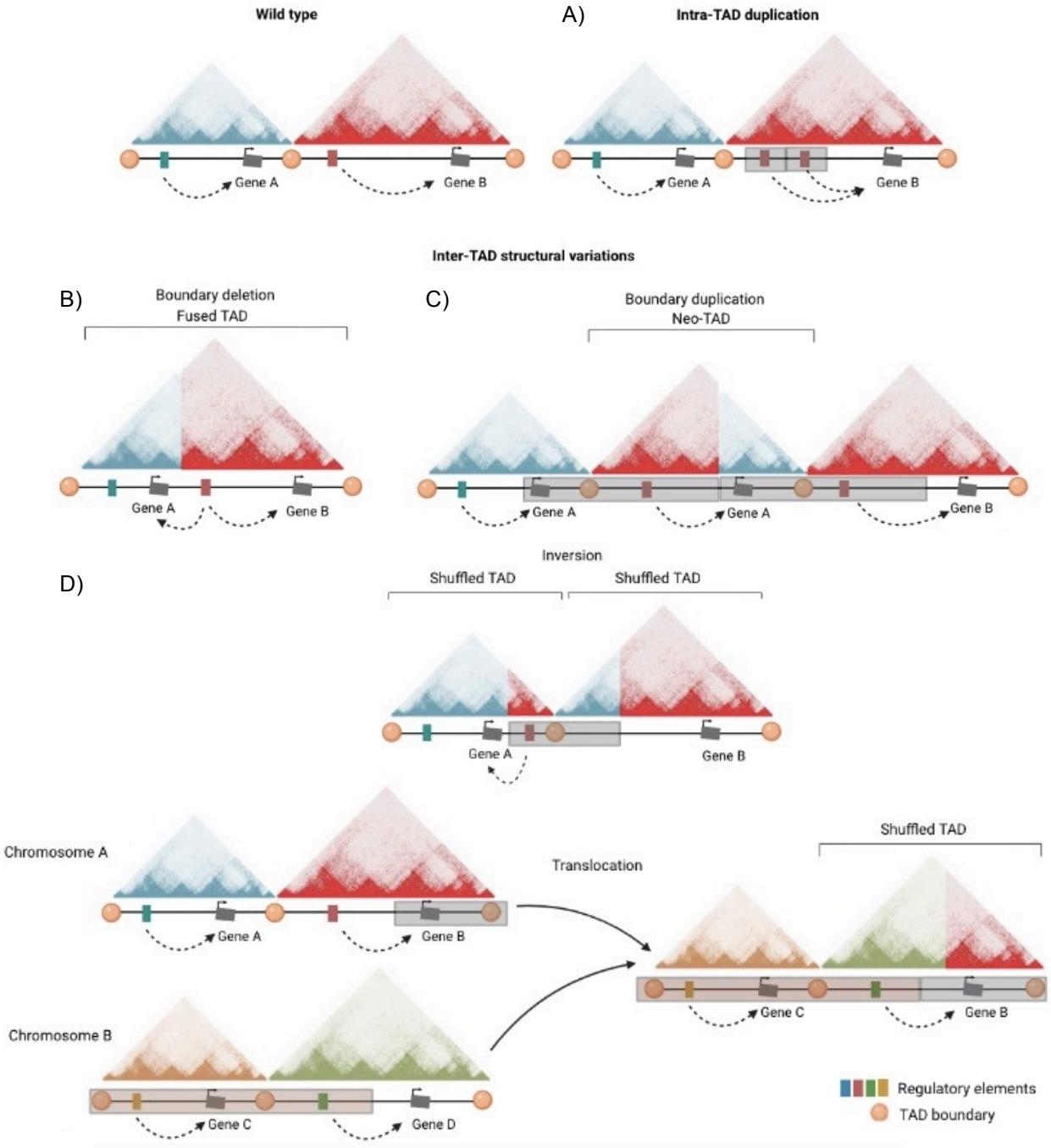

Figure 4. Structural rearrangements within and across TADs can result in oncogene activation. A) Intra-TAD duplications of enhancers that do not disrupt TAD boundaries lead to the formation of super-enhancers that activate oncogenes within the same TAD. B) TAD fusion by deletion of the boundary. C) Neo-TADs formation by duplication of a boundary. D) Inversions and translocations spanning TAD boundaries result in TAD shuffling. Inter-TAD SVs can lead to enhancer hijacking by oncogenes in cancer. Rearrangements are highlighted in shadow. Enhancer-promoter interactions are depicted as dotted arrows. Image made with BioRender.com. Modified from Spielmann et al., *Nature Reviews Genetics*, 2018.

the interaction between clusters of enhancers and genes that originally belonged to different domains (**Figure** 4B). Indeed, in multiple myeloma it has been shown that a large deletion between two distant TADs brings together the oncogene *MYC* and *NSMCE2* super-enhancer and drives the overexpression of *MYC* by enhancer hijacking[65].

TAD disruption can also occur without changing TAD boundaries by the formation of new interacting domains that are insulated from the rest of the genome (neo-TADs) (**Figure** 4C). For example, in colorectal cancer, the tandem duplication of a region comprising the *IGF2* locus, a super-enhancer of the adjacent TAD and the boundary separating both TADs, mediates the formation of a neo-TAD in which *IGF2* interacts with the ectopic super-enhancer, leading to high transcription levels of *IGF2*[66].

Inversions and translocations embracing TAD boundaries commonly lead to rearrangements across TADs, also known as TAD shuffling, resulting in the exchange of TAD regions between two different domains and the subsequent enhancer hijacking (Figure 5D). In some acute myeloid leukaemia (AML) genomes it has been shown that an inversion in chromosome 3 disrupts two TADs. As a result of the rearrangement, an enhancer originally located in the *GATA2* TAD moves to the TAD with the *EVI1* oncogene, leading to the improper activation of *EVI1,* resembling super-enhancer activity. At the same time, the expression levels of *GATA2* decreased as a result of the enhancer loss in *GATA2* TAD[67].

5 Relevance of structural variations in the clinic

Driver structural variations widely vary in frequency and pattern across different tumor types. Hence, the detection of SVs provides a powerful tool for classification of cancer subtypes. At the same time, SVs specific to certain cancer subtypes can be used as biomarkers for cancer diagnosis and prognosis. Additionally, identifying driver SVs at pre-malignant stages enables the detection of high-risk patients, their monitoring and provision of anticancer therapies. This early detection of SVs may be possible thanks to the timeframe in between the occurrence of early driver rearrangements and the clinical presentation of cancer. Their tumor specificity and cancer-related potential make SVs good candidates for targeted therapy and personalized medicine.

5.1 Detection of structural variations

Past technologies for the detection of structural rearrangements preceding the NGS era underestimated the presence of SVs by exclusively rendering the identification of CNVs resulting from unbalanced SVs such as deletions, insertions, duplications or unbalanced translocations. On the other hand, balanced rearrangements (except from those sufficiently large to be detectable by karyotyping) require the implementation of NGS technologies[68]. Array comparative genome hybridization (Array-CGH) is one of the oldest SV analysis methods still in use for robust detection of CNVs in the clinic, ranging from whole chromosomes to CNVs of a few kilobases long[68]. Array-CGH uses a chip containing thousands of probes homologous to specific genomic regions. The reference and the query DNA are then labelled with different fluorescent dyes and added to the array slide. CNVs are measured as the ratio of fluorescent sample DNA to reference DNA for each probe[69]. The main limitations of this technique are the impossibility to detect absolute copy numbers, as it compares only two genomes, nor balanced SVs. Besides, array-CGH does not provide precise definition of breakpoints. Conventional targeted approaches such as fluorescent *in situ* hybridization (FISH) and PCR, are also commonly used in the clinic for CNVs analysis. However, these cannot be used for *de novo* detection of SVs, as they require prior definition of the regions of interest[70]. With the continuously drop of sequencing price parallel to the improvement of molecular biology techniques, NGS technologies are being gradually implemented for the detection of whole-genome SVs in the clinic. NGS allows for high-throughput calling of all classes of SVs at a single-base resolution. The state-of-the-art sequencing method is paired-end sequencing and consists on the fragmentation of DNA, amplification and sequencing starting from both ends of the DNA fragments. In comparison to single-end sequencing, paired-end provides better alignments to the reference genome, which increases the resolution of SV detection[1].

There are multiple algorithms for the analysis of SVs by paired-end sequencing that pay attention to different features of the NGS results to infer structural rearrangements. The read depth, which refers to the amount of reads mapped to a specific genomic region of the reference sequence, can be used as a measurement of CNVs with both paired-end and single-end reads[71] (Figure 5A). Although it is currently the preferred sequencing-based approach to accurately detect absolute CNVs, its

low breakpoint resolution and possible PCR-based biases in sequence coverage, which have an effect on read depth, are two main drawbacks that hamper this method[1].

How paired reads map to the reference genome with respect to each other provides useful information for the detection of all classes of SVs. Aberrant mapping of two paired reads can appear as being more separated or closer than expected, being both in the same orientation in the reference, pointing apart from each other in the reference or mapping in different chromosomes[72]. Different SV classes correlate to each of these possibilities (Figure 5B). In this approach, breakpoint resolution depends on the coverage and average fragment size. Mate-pair sequencing offers a variation of paired-end technology for the detection of large rearrangements. It consists on the circularization of large library fragments and their posterior cut in order to sequence the junction point[73].

Another different method for SV identification by paired-end sequencing is *de novo* sequence assembly, which allows for the reconstruction of the original sequence by joining groups of overlapping paired-reads, without mapping them to a reference genome. Overall, this method is time consuming, computationally expensive and more error-prone[74]. The major challenge for *de novo* assembly are repeated regions[74], which usually resolve incompletely and which have high susceptibility to structural rearrangements[27]. However, *de novo* assembly is a good strategy for the precise detection of SVs and insertion of external genetic material[75].

NGS-based techniques are the most commonly used in current routine molecular genetic diagnosis. However, short-read sequencing still has some limitations, including the challenging detection of variants in repetitive regions and complex rearrangements in a haplotype-resolved manner. Long-read sequencing (LRS) is currently emerging in medical genetics as a solution to these limitations. Its main advantages are the power to span challenging genomic regions and complex SVs by using long reads and providing single-molecule resolution, which makes this technique independent of a prior PCR step that introduces GC-content dependent coverage bias[7]. Although LRS remains too expensive for clinical purposes, its increasing throughput and great potential for SVs detection is expected to make them affordable for their future implementation in the clinic.

Different integrative approaches have been implemented to combine multiple SV-calling techniques for better-resolution diagnosis. For example, a multiplatform approach for the detection of SVs in cancer genomes has recently been suggested, which combines optical mapping, based on the analysis of the restriction patterns resulting from the digestion of fluorescently labelled ultra-long linearized DNA molecules; Hi-C, for the detection of aberrant genomic interactions; and

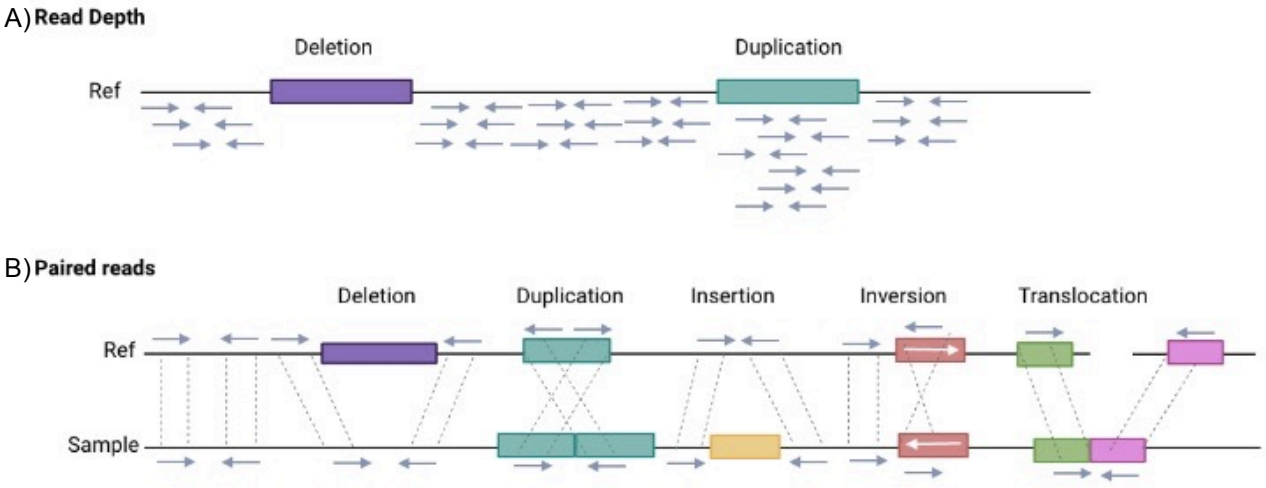

Figure 5. Paired-read sequencing methods for SV detection. A) Read depth is commonly used for the detection of CNVs caused by deletions and duplications. Upon deletion of a region no reads or fewer reads will map in that region of the reference, compared to the average depth. At duplicated regions, a higher depth than the average will be observed when mapping the reads to the reference. B) Paired-reads mapping properties can be used to detect all types of SVs. Reads more separated than expected in the reference indicate there is a deletion, upon duplication reads will point apart from each other in the reference, an insertion is detected as reads mapping closer than expected in the reference, reads pointing in the same direction are obtained with inversions and mapping in different chromosomes with translocations. Arrows: Sequencing reads. Ref: reference. Image made with BioRender.com. Modified from Escarami´s et al., *Brief. Funct. Genomics*, 2015.

WGS. This method allowed for the detection of novel gene fusions and TAD rearrangements in cancer samples and cell lines[70]. Another integrative approach combining NGS and the Bionano Genomics (BNG) platform (optical mapping technology) has unravelled somatic SVs in new putative oncogenes, such as *TRIO* and *SESTD1*, in non-small cell lung carcinoma genomes[76].

5.2 *Detection of SVs for early cancer diagnosis, subtype classification and prognosis prediction*

Driver structural rearrangements usually occur in the early stages of life, decades before the clinical presentation of cancer. Detection of these driver SVs offers the opportunity for early cancer detection and intervention. For example, it has been shown that the loss of chromosome 3p, which spans multiple tumor suppressor genes, appears in the early lifetime of patients before clear cell renal carcinoma (ccRCC) diagnosis. Furthermore, a few hundred of cells harbouring this mutation have been proven capable of initiating sporadic tumors[77]. In a similar way, the already discussed *TMPRSS2-ERG* fusion gene is present at early stages in prostate cancer and has been suggested to be involved in carcinogenesis[78].

Another important application of SVs in the clinic is cancer subtype classification, the same way driver mutations have been widely used for cancer subtype diagnosis. Previous evidence shows a relation between specific SV patterns and different tumor subtypes. For example, in gastric cancer, one signature of tandem duplications has been found enriched in tumors with *cadherin 1* mutations, which also associated with poor clinical outcome[79]. Specific fusion genes are also related to different histologic tumor subtypes. For instance, in childhood supratentorial ependymoma, C11orf95-*RELA* translocations, albeit co-occurring with low rates of coding mutations, have been linked to poor prognosis, which highlights the oncogenic potential of the fusion gene by itself, without a need for accumulating coding mutations. Nowadays the WHO considers *RELA*-positive supratentorial ependymoma as a distinct diagnostic category[51]. This also underlines the possible application of SVs as prognosis biomarkers. One further example are complex rearrangements in Ewing sarcoma. *EWSR1*-related fusion genes resulting form chromoplexy arise at early stages and associate with an aggressive form of the cancer, generating both primary and relapse Ewing sarcoma tumors[80]. All this proves the powerful role of SVs in the clinic as novel biomarkers for early diagnosis, cancer stratification and therapeutic decisions.

5.3 *Structural variants as therapeutic targets*

Besides providing novel strategies for tumor characterization and diagnosis, SVs also offer novel targets for cancer treatment and personalized medicine that have provided important improvements in patient survival and tumor relapse. However, when looking for novel therapeutic drugs for cancer treatment it is important to bear in mind, like with any other cancer monotherapies, the possibility for the development of tumor resistance, which highlights the need for combinational treatments.

The functional consequences of SVs in oncogene overexpression by super-enhancer activation and hijacking has put an eye on the development of novel inhibitors targeting these super-enhancers. BET inhibitors against specific bromodomain-and-extraterminal family proteins (BRDs) and cyclin-dependent kinase inhibitors (CDK7s) are capable of inhibiting oncogenic transcription driven by super-enhancers and selectively kill cancer cells. The mechanisms of action is basically based on inhibiting the recognition of H3K27ac marks in super-enhancers by BRD4 (BET family member) or the subsequent recruitment of CDK7 and CDK9, complexed with other proteins, involved in transcription initiation and elongation[81]. Despite showing promising results in tumor relapse in preclinical models, resistance to these inhibitors has also been reported. For instance, in *MLL*-fusion AML, the activation of Wnt signaling pathway as a response to BETi treatment brings back the activation of *MYC*[82]. With the aim to overcome this resistance, combination of BETi and *NOTCH1* activation inhibitors has shown synergistic effects in mouse models of T-cell acute lymphoblastic leukaemia (T-ALL)[82].

Novel oncogenic proteins derived from fusion genes are also powerful targets in cancer therapy. In Ewing sarcoma, characterized by the *EWS-FLI1* fusion gene, which codes for an aberrant transcription factor that promotes tumorigenesis, targeting *EWS-FLI1* has shown great promise in preclinical stages. *EWS-FLI1* oncogenic effects can be prevented by impairing the fusion-gene transcription with either antisense oligodeoxynucleotides or siRNAs; by disrupting the binding of EWS-FLI1 protein to its transcriptional modulator targets, such as its binding to Poly(ADP-ribose) polymerase 1 (PARP1) impaired by PARP inhibitors (e. g. olaparib), also commonly used in

breast cancer treatment; or by targeting genes downstream *EWS-FLI1*[82].

As already mentioned, SVs may also have an oncogenic effect by inducing changes in epigenetics. The overexpression of *NSD2* generated by a translocation in multiple myeloma leads to unbalances in histone marks that result in aberrant gene expression patterns and carcinogenesis. This overexpressed histone methyltransferase can be targeted by the specific inhibitor LEM-14, which has been proposed as a potent therapeutic drug for cancer personalized treatment[82].

6 Discussion and future perspective

Cancer is understood as the uncontrolled proliferation of cells. This neoplastic potential is known to be acquired by the accumulation of somatic mutations affecting genes involved in cell-cycle progression, apoptosis and cellular senescence. The scientific community has widely focused on studying mutations that implied SNPs and small indels to decipher the underlying mechanisms in tumor progression. With this aim, different mutagenic processes have been linked to specific mutational signatures in cancer genomes[83]. It was not until the development of next-generation sequencing that structural variations started to gain considerable significance in tumorigenesis, showing the important role genomic rearrangements may play in misregulating oncogene and tumor-suppressor gene expression[1]. Nowadays, we are still increasing our understanding of SVs and their functional consequences in tumor development, which allows to gradually incorporate the detection of SVs into the clinic in order to move towards improved diagnosis and novel personalized treatments.

SVs, including deletions, insertions, duplications, translocations and inversions, are known to result from DNA double-strand breaks (DSBs) unfaithfully repaired by forming rearrangements between different genomic regions[2]. These DSBs can be caused by replication stress, by topoisomerases at actively transcribed regions, during replication or chromosome condensation, or by the breakage of di-centric chromosome derived from telomere dysfunction during cell division, among others[26]. However, little is known about how multiple simultaneous DSBs my occur during the formation of complex SVs. While chromothripsis is thought to be the result of the degradation of micronuclei[34] and di-centric chromosomes[32], the origin of chromoplexy, which embraces rearrangements among multiple chromosomes, remains unknown. Previous evidence shows that chromoplexy breakpoints cluster mainly at actively transcribed and open chromatin genomic regions[35], but the origin of the DSBs occurred in these regions remains unexplored. In line with this, and taking into account the role of topoisomerase in relaxing supercoiled DNA during transcription, it would be interesting to study whether this enzyme is involved in the formation of DSBs during chromoplexy at actively transcribed regions.

Following the breakage of both strands of DNA, its repair by alternative error-prone DNA repair pathways leads to the formation of SVs. It is known that the cell-cycle stage may influence which DNA repair pathway is activated to solve DSBs, being HR most commonly activated during the S phase of mitosis and NHEJ when no sister chromatids are available[36]. However, little is known about how HR-alternative (NAHR) and NER-alternative (MMEJ and SSA) pathways are chosen over HR or NHEJ. Previous evidence suggests that alternative pathways may be activated upon a deficiency of some players of HR or NHEJ. For example, prostate cancer genomes with biallelic mutations in *BRCA2* have high levels of deletions with microhomology at the breakpoint, which suggests that MMEJ was used as repair mechanism in HR deficient cells[3].

Future research is needed in order to completely understand how multiple DSBs may occur during the formation of complex SVs, as well as which repair mechanisms are involved in certain rearrangements and the signalling pathways leading to their activation. The main challenge here is to identify the succession of rearrangements that explain specific SV signatures. A recent study has proposed a computational approach in which different SV signatures in human cancers were classified based on the number of events and size of deletions, tandem duplications, translocations, inversion and other rearrangements[4]. Besides the already-mentioned relation between *BRCA2* deficiency and small deletions, the authors found other associations between cancer-driver mutations and SV signatures across tumor types[4]. This explains the different prevalence of specific SVs found across cancer subtypes but does not elucidate the occurrence of cancer-driver SVs in the absence of previous driver mutations, such as oncogenic fusion genes driving many pediatric cancers[51]. Whether the activation of specific repair pathways for the repair of DSBs could lead to different SV signatures in different cancer types is

not well understood, and puts forward future research lines to provide insight into SV formation and their specificity across cancer subtypes.

SVs have functional consequences on cancer by activating or inactivating oncogenes and tumor suppressor genes, respectively. While the molecular mechanisms by which SVs in coding regions disrupt tumor suppressor genes or induce CNVs of cancer-driver genes are well studied, how rearrangements involving non-coding regions drive oncogenesis remains partially unexplored. Enhancer hijacking inside or across TADs and super-enhancer acquisition seems to be the main cause of oncogene activation by SVs in intragenic regions. However, several aspects remain unclear. For example, the inversion in chromosome 3 found in some acute myeloid leukaemias that results in *GATA2*-enhancer hijacking by *EVI1* oncogene has been shown to activate *EVI1* and trigger the adoption of super-enhancer activity[67]. Nonetheless, it is not well studied how the repositioning of an enhancer on a different genomic region enables the acquisition of super-enhancer activity. When talking about rearrangements of the genome it is important to bear in mind that epigenetics also play an important role in the regulation of gene expression. It is known that histone marks that trigger an open chromatin conformation also correlate with active gene expression[17,18]. Hence, it could be that the new position of the enhancer presents more open chromatin structure and, thus, results in super-enhancer activity.

Future research on how SVs are involved in tumorigenesis will pave the way towards improved diagnosis and personalized treatments by, not only detecting SVs at early cancer stages, but also predicting the functional consequences of the structural rearrangement and, subsequently, providing the patient with effective and personalized treatments.

Needless to say, in order to study SV formation and functional consequences on carcinogenesis, sequencing technologies with high resolution and sensitivity for the detection of SVs are needed. The recently-developed long-read sequencing solves the limitations of short-read NGS, including the detection of breakpoints at repeated regions and in a haplotype-resolved manner. However, there are some cases in which LRS cannot be applied, such as when sequencing formalin-fixation and paraffin embedding (FFPE) samples. Solid tumors are routinely stored as FFPE[84]. With this technique, the DNA is crosslinked and fragmented, making it impossible to use LRS for the detection of SVs in FFPE-samples. A recent publication presents FFPE-targeted locus capture (FFPE-TLC) as a solution to this problem, a technique that ligates the crosslinked and fragmented DNA before its sequencing. This method has shown robust detection of SV breakpoints, having identified novel rearrangements missed by capture-NGS[85].

7 Conclusions

Structural variations are an important hallmark of cancer. The molecular mechanisms underlying their oncogenic potential remains incompletely understood due to the challenges of detecting SV breakpoints at a single-nucleotide resolution. The on-going improvement of sequencing technologies and the development of integrative approaches have shed more light into the important role SVs play in tumorigenesis and their considerable potential as biomarkers for early diagnosis, prognosis and as targets for effective personalized cancer treatments. However, future research is still needed in order to completely elucidate how SV are formed, their oncogenic potential and cancer sub-type specificity.

8 References

1. Escaramís, G., Docampo, E. & Rabionet, R. A decade of structural variants: Description, history and methods to detect structural variation. *Brief. Funct. Genomics* (2015). doi:10.1093/bfgp/elv014
2. Currall, B. B., Chiang, C., Talkowski, M. E. & Morton, C. C. Mechanisms for Structural Variation in the Human Genome. *Curr. Genet. Med. Rep.* **1**, 81–90 (2013).
3. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **175**, 889 (2018).
4. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* (2020). doi:10.1038/s41586-019-1913-9
5. Valton, A.-L. & Dekker, J. TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40 (2016).
6. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
7. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **10**, 1–14 (2019).
8. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nature Reviews Genetics* (2018). doi:10.1038/s41576-018-0007-0
9. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
10. Shen, M. M. Chromoplexy: a new category

of complex rearrangements in the cancer genome. *Cancer Cell* **23**, 567–9 (2013).
11. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 4283–8 (2008).
12. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
13. Fanti, L. & Pimpinelli, S. HP1: a functionally multifaceted protein. *Curr. Opin. Genet. Dev.* **18**, 169–74 (2008).
14. Trojer, P. & Reinberg, D. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Mol. Cell* **28**, 1–13 (2007).
15. Liu, J., Ali, M. & Zhou, Q. Establishment and evolution of heterochromatin. *Annals of the New York Academy of Sciences* (2020). doi:10.1111/nyas.14303
16. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
17. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. S. Role of Histone H3 Lysine 9 Methylation in Epigenetic Control of Heterochromatin Assembly. *Science (80-.).* **292**, 110–113 (2001).
18. Tamaru, H. Confining euchromatin/heterochromatin territory: jumonji crosses the line. *Genes Dev.* **24**, 1465–78 (2010).
19. Hervouet, E. *et al.* The autophagy GABARAPL1 gene is epigenetically regulated in breast cancer models. *BMC Cancer* **15**, 729 (2015).
20. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **xx**, (2021).
21. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–11 (2002).
22. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* (2006). doi:10.1038/ng1896
23. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* (2009). doi:10.1126/science.1181369
24. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–80 (2012).
25. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–80 (2014).
26. Wang, W. J., Li, L. Y. & Cui, J. W. Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. *Epigenetics and Chromatin* (2020). doi:10.1186/s13072-020-00371-7
27. Bouwman, B. A. M. & Crosetto, N. Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling. *Genes (Basel).* **9**, (2018).
28. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* **75**, 267-283.e12 (2019).
29. Numata, M., Saito, S. & Nagata, K. RAG-dependent recombination at cryptic RSSs within TEL-AML1 t(12;21)(p13;q22) chromosomal translocation region. *Biochem. Biophys. Res. Commun.* **402**, 718–24 (2010).
30. Kato, L. *et al.* Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2479–84 (2012).
31. de Lange, T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100–10 (2005).
32. Bhargava, R., Fischer, M. & O'Sullivan, R. J. Genome rearrangements associated with aberrant telomere maintenance. *Curr. Opin. Genet. Dev.* **60**, 31–40 (2020).
33. Shen, E. *et al.* POT1 mutation spectrum in tumour types commonly diagnosed among POT1-associated hereditary cancer syndrome families. *J. Med. Genet.* **57**, 664–670 (2020).
34. Zhang, C.-Z. *et al.* Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–84 (2015).
35. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50**, (2018).
36. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biol.* **16**, 72 (2015).
37. Zhang, Y. & Jasin, M. An essential role for CtIP in chromosomal translocation formation through an alternative end-joining pathway. *Nat. Struct. Mol. Biol.* **18**, 80–4 (2011).
38. Zhang, F. *et al.* The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849–53 (2009).
39. Hastings, P. J., Ira, G. & Lupski, J. R. A Microhomology-Mediated Break-Induced

Replication Model for the Origin of Human Copy Number Variation. *PLoS Genet.* **5**, e1000327 (2009).
40. Michelini, F. *et al.* Damage-induced lncRNAs control the DNA damage response through interaction with DDRNAs at individual double-strand breaks. *Nat. Cell Biol.* **19**, 1400–1411 (2017).
41. Wang, H. *et al.* An intragenic long noncoding RNA interacts epigenetically with the RUNX1 promoter and enhancer chromatin DNA in hematopoietic malignancies. *Int. J. cancer* **135**, 2783–94 (2014).
42. Zhang, Y. *et al.* Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3070–5 (2002).
43. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170**, 507-521.e18 (2017).
44. Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–19 (2011).
45. Bohlander, S. K. & Kakadia, P. M. *DNA Repair and Chromosomal Translocations.* **200**, (Recent Results Cancer Res, 2015).
46. Li, J. *et al.* Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS Genet.* **8**, e1002692 (2012).
47. Goodier, J. L. Restricting retrotransposons: a review. *Mob. DNA* **7**, 16 (2016).
48. Iden, M. *et al.* Multi-omics mapping of human papillomavirus integration sites illuminates novel cervical cancer target genes. *Br. J. Cancer* (2021). doi:10.1038/s41416-021-01545-0
49. Xu, M. *et al.* Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. *Theranostics* **9**, 1115–1124 (2019).
50. Zhao, K., Liu, A. & Xia, Y. Insights into Hepatitis B Virus DNA Integration-55 Years after Virus Discovery. *Innov. (New York, N.Y.)* **1**, 100034 (2020).
51. Rahal, Z., Abdulhai, F., Kadara, H. & Saab, R. Genomics of adult and pediatric solid tumors. *Am. J. Cancer Res.* (2018).
52. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–81 (2015).
53. Shtivelman, E., Lifshitz, B., Gale, R. P. & Cananni, E. Fused transcript of abl and bcr genes in CML. *Nature* (1985).
54. Robinson, D. R. *et al.* Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing. *Nat. Genet.* **45**, 180–5 (2013).
55. Demichelis, F. *et al.* TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* **26**, 4596–9 (2007).
56. Northcott, P. A. *et al.* Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49–56 (2012).
57. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148–59 (2008).
58. Popovic, R. *et al.* Histone methyltransferase MMSET/NSD2 alters EZH2 binding and reprograms the myeloma epigenome through global and focal changes in H3K36 and H3K27 methylation. *PLoS Genet.* **10**, e1004566 (2014).
59. Lhoumaud, P. *et al.* NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains. *Nat. Commun.* **10**, 4843 (2019).
60. Zhang, Y. *et al.* Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol.* (2019). doi:10.1186/s13059-019-1818-9
61. Allen, B., Pezone, A., Porcellini, A., Muller, M. T. & Masternak, M. M. Non-homologous end joining induced alterations in DNA methylation: A source of permanent epigenetic change. *Oncotarget* **8**, 40359–40372 (2017).
62. Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* **27**, 246–258 (2017).
63. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–47 (2013).
64. Glodzik, D. *et al.* A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat. Genet.* **49**, 341–348 (2017).
65. Mikulasova, A. *et al.* Microhomology-mediated end joining drives complex rearrangements and overexpression of MYC and PVT1 in multiple myeloma. *Haematologica* **105**, 1055–1066 (2020).
66. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
67. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).
68. Pócza, T. *et al.* Germline Structural Variations in Cancer Predisposition Genes. *Front. Genet.* **12**, 1–14 (2021).
69. Davies, J. J., Wilson, I. M. & Lam, W. L.

Array CGH technologies and their applications to cancer genomes. *Chromosome Res.* **13**, 237–48 (2005).
70. JR, D. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, (2018).
71. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**, 1586–92 (2009).
72. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–9 (2008).
73. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–6 (2007).
74. Cantsilieris, S., Baird, P. N. & White, S. J. Molecular methods for genotyping complex copy number polymorphisms. *Genomics* (2013). doi:10.1016/j.ygeno.2012.10.004
75. Hajirasouliha, I. *et al.* Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26**, 1277–83 (2010).
76. Xia, L. *et al.* Multiplatform discovery and regulatory function analysis of structural variations in non-small cell lung carcinoma. *Cell Rep.* (2021). doi:10.1016/j.celrep.2021.109660
77. Mitchell, T. J. *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* (2018). doi:10.1016/j.cell.2018.02.020
78. BS, C. *et al.* Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat. Genet.* **41**, (2009).
79. Xing, R. *et al.* Whole-genome sequencing reveals novel tandem-duplication hotspots and a prognostic mutational signature in gastric cancer. *Nat. Commun.* (2019). doi:10.1038/s41467-019-09644-6
80. Anderson, N. D. *et al.* Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* **361**, (2018).
81. Thandapani, P. Super-enhancers in cancer. *Pharmacol. Ther.* **199**, 129–138 (2019).
82. Fong, C. Y. *et al.* BET inhibitor resistance emerges from leukaemia stem cells. *Nature* **525**, 538–42 (2015).
83. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).
84. N, B. Tissue preparation: Tissue issues. *Nature* **448**, (2007).
85. Allahyar, A. *et al.* Robust detection of translocations in lymphoma FFPE samples using targeted locus capture-based sequencing. *Nat. Commun.* (2021). doi:10.1038/s41467-021-23695-8