Prospects & Overviews

Disruption of regulatory domains and novel transcripts as disease-causing mechanisms

Lila Allou1,2 | Stefan Mundlos1,2,3

1RG Development & Disease, Max Planck Institute for Molecular Genetics, Berlin, Germany
2Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Berlin, Germany
3Berlin-Brandenburg Center for Regenerative Therapies, Charité-Universitätsmedizin Berlin, Berlin, Germany

Abstract

Deletions, duplications, insertions, inversions, and translocations, collectively called structural variations (SVs), affect more base pairs of the genome than any other sequence variant. The recent technological advancements in genome sequencing have enabled the discovery of tens of thousands of SVs per human genome. These SVs primarily affect non-coding DNA sequences, but the difficulties in interpreting their impact limit our understanding of human disease etiology. The functional annotation of non-coding DNA sequences and methodologies to characterize their three-dimensional (3D) organization in the nucleus have greatly expanded our understanding of the basic mechanisms underlying gene regulation, thereby improving the interpretation of SVs for their pathogenic impact. Here, we discuss the various mechanisms by which SVs can result in altered gene regulation and how these mechanisms can result in rare genetic disorders. Beyond changing gene expression, SVs can produce novel gene-intergenic fusion transcripts at the SV breakpoints.

KEYWORDS

3D genome organization, enhancer-promoter communication, gene-intergenic fusion transcript, structural variation, topologically associated domain

INTRODUCTION

Deletions, duplications, insertions, inversions, and translocations, collectively called structural variations (SVs), contribute significantly to human genome variability.1–4 Pathogenic SVs have been involved in many conditions,5 including developmental and sensory disorders, and are increasingly recognized as important causes of inherited diseases.6 They can lead to copy number changes (gains and losses of DNA), disruption of protein-coding genes, enhancer-promoter communication rewiring, and production of fusion genes.

With the emergence of novel sequencing technologies, SV identification has significantly improved.7 On average, over 27,000 SVs are identified per human genome.2 The finding that SVs can alter the communication between regulatory elements and genes by disrupting topologically associated domains (TADs) has significantly improved our interpretation of the pathogenic impact of SVs.5 Nevertheless, the medical interpretation of SVs and the prediction of their functional consequences, as well as dissecting the mechanisms by which they exert their functional impact, still need to be improved. Indeed, the vast majority of SVs occur outside the protein-coding portion of the genome.2 Thus, the incomplete functional annotation of non-coding DNA sequences and the complexity of gene regulatory landscapes...
severely hamper interpreting their pathogenic impact. Moreover, the recent advances in gene regulation and 3D genome organization indicate that additional regulatory layers (e.g., enhancer-promoter specificity) control the communication between genes and cis-regulatory elements (CREs) beyond TADs. These additional regulatory layers still need to be fully understood to fully predict the functional consequences of SVs with respect to gene regulation and human disease etiology. Additionally, even a single SV can have a vast array of consequences with simultaneous copy number changes, rewiring of TAD 3D architecture, and generation of novel fusion transcripts, which makes them harder to interpret. Finally, the incomplete functional annotation of human genes further complicates this.

In this review, we focus on rearrangements rather than aneuploidies and use the term “SV” to represent deletions, duplications, insertions, inversions, chromosomal translocations, and any combination of these events. We also exclusively focus on germline SVs. We first discuss how these SVs can alter gene expression in cis. We then discuss a new mechanism through which SVs impacting non-coding DNA sequences exert their pathogenic effect by producing a new gene-intergenic fusion transcript. Finally, we also discuss the interpretation of SV impact in light of the latest advances in gene regulation and 3D genome organization.

**Structural variation**

Human genetic variations include single-nucleotide variants (SNVs), small insertions and deletions (indels; <50 bp), and SVs. SVs can broadly differ in size, ranging from ~50 bp to several megabases, affecting more of the genome per nucleotide changes than any other sequence variant. They are highly diverse in type (Figure 1), including deletions, duplications, and insertions of genetic material, known as unbalanced copy number variants (CNVs). They also include inversions and interchromosomal and intrachromosomal translocations, known as balanced rearrangements. Unbalanced CNVs change the dosage of DNA, while balanced rearrangements do not change the DNA dosage. In addition to these described events, SVs include mobile element insertions (MEIs), tandem repeats, multiallelic CNVs of highly variable copy numbers, segmental duplications, complex genomic rearrangements (CGRs), as well as aneusomy and aneuploidy.

SVs arise by homologous recombination (HR) between repeated sequences (recurrent SVs) or by non-homologous mechanisms that occur throughout the genome (non-recurrent SVs). Indeed, SVs are usually caused by erroneous DNA replication and damage repair, as well as the activities of repetitive elements. Several mechanisms are known to form SVs in germline and somatic cells. Such mechanisms include non-allelic homologous recombination (NAHR), non-homologous end-joining (NHEJ), double-strand break repair (DSBR), synthesis-dependent strand annealing (SDSA), break-induced replication (BIR), single-strand annealing (SSA), microhomology-mediated end joining (MMEJ), breakage-fusion-bridge (BFB) cycle, replication slippage, fork stalling and template switching (FoSTeS), microhomology-mediated break-induced replication (MMBR), and so on. We direct the reader to some excellent reviews for more details on the mechanisms underlying SV formation.

Various experimental techniques have been developed to identify SVs. Karyotyping was routinely performed to diagnose and screen genetic diseases. Nowadays, it is rarely used for SV discovery due to its low sensitivity and precision. Instead, array comparative genomic hybridization (array CGH) is commonly used in routine clinical diagnostics for genetic disorders. However, this technique...
has limitations, including its relatively low resolution, the inability to detect balanced rearrangements, and the low efficacy in detecting mosaics.

Due to its limitations, array CGH is now mainly replaced by second-generation short-read sequencing technologies, which became available more than 10 years ago. However, due to the short-read length limitation and the human genome’s repetitive nature, SVs in repetitive and segmental duplication regions remain difficult to identify.

The third-generation long-read sequencing and imaging technologies can overcome the limitations of second-generation short-read sequencing techniques. We refer the reader to another review for an in-depth discussion of strategies and algorithms for SV detection.

Current estimates suggest that a human genome may harbor >27,000 SVs compared to the reference genome. These SVs primarily impact non-coding DNA sequences. Therefore, completing the functional annotation of non-coding regions and understanding the basic principles underlying gene regulation and the genome’s spatial 3D organization is crucial to comprehend their pathogenic potential fully. In the following section, we provide a brief overview of the current concepts regarding the hierarchical structures of the 3D genome. For more details on the organization and function of the 3D genome and the methods for understanding how chromatin is folded in the nucleus, we refer the reader to some excellent reviews.

### 3D genome organization

Cytological and microscopy techniques revealed that individual chromosomes occupy separate chromosome territories in the nucleus during interphase. Chromosome conformation capture (3C) technologies such as HiC rely on quantifying interaction frequencies between loci that lie in close spatial proximity in the nucleus independently of their linear genomic distance. HiC technology uncovered that the genome is hierarchically organized into multiple layers in the nucleus (Figure 2). On a course scale, the genome is compartmentalized into A and B compartments corresponding to the spatial segregation of open and closed chromatin. The A-compartment largely correlates with accessible, transcriptionally active euchromatin. It is mainly present in the interior nuclear space. Conversely, the B-compartment largely correlates with compacted, transcriptionally silent heterochromatin predominantly located in the periphery of the nucleus as lamina-associated domains (LADs) or nucleoli as nucleolar-associated domains (NADs). The separation of A and B compartments has been proposed to reflect phase separation between active and inactive chromatin based on their inherent physical properties.

Compartment boundaries are subdivided into regions called TADs. TADs are megabase-long genomic regions that interact with themselves at a higher frequency than with the rest of the genome. The regions preventing contact between neighboring TADs are called boundaries or borders (Figure 2). Boundaries are mainly formed by the site-directed divergent binding of the CCCTC-binding factor (CTCF). TAD formation and function are mainly achieved by the combined action of the cohesin complex and CTCF through a process called loop extrusion.

During this process, cohesin is loaded into chromatin and progressively extrudes a loop until a roadblock factor stalls it or the cohesin unloading factor Wap1 releases it.

Although transcription start sites (TSSs) and transcription termination sites (TTTs), as well as other proteins such as RNA polymerase II, might act as cohesin barriers, the most common cohesin roadblock factor is formed by convergent CTCF binding sites (CBSs) in vertebrates.

TADs constrain the genomic regions an enhancer can act upon, and most enhancer-gene interactions occur within TADs through chromatin loops that connect distant DNA fragments. Thus, TADs represent fundamental regulatory units facilitating the communication between enhancers and their target genes in the same domain. At the same time, TAD boundaries act as insulators that prevent enhancers from contacting non-target genes in other domains. Indeed, several observations support the relevance of TADs for enhancer function. For example, enhancers and their target genes usually lie within the same TAD and disruption of TAD boundaries by SVs or mutations in CBSs can result in novel enhancer-promoter combinations and ectopic gains in gene expression.
However, this does not seem to apply to all genome regions. First, there is a significant fraction of the genome wherein TADs cannot be clearly detected, such as near Wnt6 (Wingless-type MMTV integration site family, member 6), which is located in a gene-dense region composed of weakly defined, less insulated structures.\textsuperscript{[48,72]} Second, TAD fusion was observed following boundary elimination at the Sox9-Kcnj2 locus, and TADs loss was observed following CTCF or cohesin depletion. However, this did not cause significant gene expression changes.\textsuperscript{[8,54,57]} Third, using mouse embryonic stem cells (mESCs), a recent study found that enhancer-promoter interactions and transcription are largely insensitive to acute depletion of CTCF, cohesin, WAPL, or YY1 (Ying Yang 1).\textsuperscript{[73]} Fourth, genes within a TAD are not always co-regulated, and the degree of promoter responsiveness to enhancers can vary based on position within a TAD.\textsuperscript{[66,74–76]} These data and others (see section “Enhancer adoption: rare or overlooked?”) indicate that additional regulatory layers control the responsiveness between genes and enhancers beyond inclusion in the same domain. Overall, this delineates the organizational and functional complexity of the gene regulatory landscapes, hampering the interpretation of SVs impacting non-coding DNA sequences and, thus, our understanding of human disease etiology.

**SVs ALTERING GENE REGULATION IN Cis**

The most studied and best-understood mechanisms through which SVs can cause disease are those directly affecting coding sequences (i.e., disrupting genes, changing gene dosage, and creating fused genes). Nevertheless, the latest technological advances allowed us to decipher new complex mechanisms through which SVs can cause disease. Indeed, detailed studies demonstrated that SVs can exert their pathogenic effect by rewiring 3D genome organization and enhancer-gene communication. Following, we will summarize the mechanisms by which SVs can alter gene regulation in cis, considering 3D chromatin architecture and TAD organization.

**Leaving the TAD structure intact**

SVs that occur within TADs and affect CREs can result in a tissue-specific loss of function or gain of function of their endogenous target genes, which can be located several hundred kilobases away\textsuperscript{[77–84]} (Figure 3A–C). These SVs usually cause phenotypic features resembling selective phenotypic sub-features of coding mutations. For example, the engrailed-1 (En1) gene loss of function mutations cause severe limb and brain malformations. In contrast, intra-TAD homologous deletions encompassing its limb-specific CRE, the long non-coding RNA (lncRNA) locus Maenli (Master activator of engrailed-1 in the limb), cause a partial phenotype restricted to the limb\textsuperscript{[85]} (Figure 3D). Of note, duplications occurring within TADs and affecting CREs can also result in tissue-specific misexpression of their endogenous target genes and disease\textsuperscript{[77,79,86–89]} (Figure 3C). This is, for example, the case for duplications involving regulatory elements within the Ihh (Indian hedgehog) regulatory domain. These duplications are associated with craniosynostosis and synpolydactyly. In this case, the craniosynostosis strictly associates with Ihh overexpression, whereas the observed synpolydactyly results from Ihh tissue-specific misexpression.\textsuperscript{[89]}

One major prerequisite to interpreting and deciphering the pathogenicity of intra-TAD SVs is a complete functional annotation of non-coding DNA sequences. For example, the Maenli lncRNA locus was not annotated as an En1 limb-specific CRE.\textsuperscript{[85]} This hampered the clinical interpretation of the pathogenic impact of the 28-kb non-coding homozygous deletions identified in the vicinity of the En1 gene, especially since En1 was not involved in the disease until then.\textsuperscript{[85]} Another example is the deletions/duplications identified downstream of the SHOX (Short Stature Homeobox) gene. CNVs affecting one or more of the seven SHOX annotated non-coding regulatory elements represent one of the most frequent causes of SHOX-haploinsufficiency.\textsuperscript{[90]} Recently, two partially overlapping small deletions of ~12- and ~8-kb, approximately 15-kb downstream SHOX, were identified in four unrelated individuals with clinical features strongly suggestive of SHOX-haploinsufficiency. However, these deletions do not include previously annotated SHOX regulatory elements, suggesting they may affect yet unidentified SHOX-specific CREs. Thus, the incomplete functional annotation of SHOX-specific CREs hampers fully comprehending the pathogenic potential of the identified CNVs.\textsuperscript{[90]}

Moreover, it is essential to consider that tissue-specific transcription is generally controlled by multiple CREs that work totally or partially redundant to achieve their target genes’ precise tissue-specific expression patterns. This confers phenotypic robustness to loss of function mutations in individual enhancers.\textsuperscript{[91,92]}

**Creating a new gene-regulatory landscape**

**Enhancer adoption**

SVs occurring between TADs can rearrange the 3D chromatin organization of a locus (Figure 4A–D). Indeed, deletions encompassing TAD boundaries will fuse the two adjacent TADs (TAD fusion) (Figure 4B).\textsuperscript{[63,93–96]} Duplications that include TAD boundaries will create new TADs (‘neo-TAD’ formation) (Figure 4C).\textsuperscript{[68,71,94,96–101]} and inversions encompassing TAD boundaries will result in the exchange of regulatory material between TADs (TAD shuffling) (Figure 4D). Finally, chromosomal translocations can result in the fusion of TADs (TAD fusion), or relocate TAD boundaries, thereby swapping regulatory material between TADs (TAD shuffling).\textsuperscript{[63,100–102]} 3D chromatin rewiring may lead to gene misexpression through a process termed enhancer adoption or enhancer ‘hijacking,’ in which ectopic expression of a gene is driven by an enhancer that normally regulates another gene located in a different regulatory domain. Thus, inter-TAD SVs can create a new gene-regulatory landscape, causing gene misexpression and disease. Several examples have been reported in the literature, such as deletions at the Epha4-Pax3 locus. These deletions rewrite 3D chromatin organization at the locus (TAD fusion) and bring enhancers from the Epha4 (Ephrin type-A receptor 4) TAD to the vicinity of Pax3.
SVs leaving TADs’ structure intact. Structural variants that occur within TADs (intra-TAD) do not alter the TAD 3D architecture. Nevertheless, they can change enhancer dosage, resulting in loss or gain of function of their endogenous target gene. (A) Schematic representation of the wild-type genomic locus showing gene A expressed in the developing brain (blue) under the control of its tissue-specific enhancer (blue ellipse) and gene B expressed in the developing limb bud (red) under the control of its tissue-specific enhancers (red ellipses). Genes A and B and their respective tissue-specific enhancers are located in two TADs (blue and green triangles) separated by a TAD boundary (black polygon). (B) The intra-TAD deletion of an enhancer element can result in tissue-specific loss of function of gene A in the brain (blue). (C) The intra-TAD duplication of an enhancer element can result in tissue-specific overexpression and possibly misexpression of gene B in the limb (red). (D) In the wild-type embryo, En1 is expressed in the brain (blue), limb (red), and somites (green). The En1 knockout (En1\(^{-/-}\)) results in a complete loss of En1 expression in all three tissues and severe congenital limb and brain malformations. The Maenil IncRNA locus controls En1 expression, specifically in the limb. Its loss (Maenil\(^{-/-}\)) results in a complete loss of En1 expression, specifically in the limb at developmental stage E9.5, and severe congenital limb malformations. The question mark (?) indicates that the enhancers controlling En1 expression in the brain and the somites are not yet characterized.

(Paired box 3). In this case, Pax3 is under the control of Eph4 enhancers leading to a pathological gain of Pax3 expression in an Eph4-like pattern and disease.\(^{[63]}\) Another example is the duplications at the Sox9-Kcnj2 locus. In this case, a region including the Kcnj2 (Potassium inwardly-rectifying channel, subfamily J, member 2) gene and Sox9 (SRY (sex determining region Y)-box 9) enhancers, normally located in two distinct TADs, is duplicated, resulting in the formation of a new TAD ("neo-TAD"), in which Kcnj2 is under the control of the duplicated Sox9 enhancers leading to gene misexpression and disease.\(^{[68]}\) Moreover, translocations involving the PAX3 (Paired Box 3) locus on chromosome 2 and the FOXO1 (Forkhead Box O1) locus on chromosome 13, bringing together the PAX3 and FOXO1 genes, have been recently reported to rewire 3D chromatin organization (TAD shuffling). The PAX3 promoter of the fused PAX3-FOXO1 gene is thus controlled by FOXO1 regulatory elements, causing PAX3-FOXO1 oncogene activation in non-PAX3 territories and disease.\(^{[102]}\)

**Enhancer disconnection**

SVs occurring between TADs can also rearrange the 3D chromatin organization of a locus without resulting in enhancer adoption. This is, for example, the case for an 89-Mb heterozygous inversion on chromosome 6, causing branchio-culo-facial syndrome (BOFS).\(^{[103]}\)

Using patient-specific human induced pluripotent stem cells (hiPSCs), the authors demonstrated that the inversion shuffles enhancers controlling TFAP2A (Transcription Factor AP-2 Alpha) gene expression with novel genes within the same TAD. However, this 3D chromatin rewiring does not result in enhancer adoption. In this case, the disease is caused by a loss of TFAP2A expression due to the disconnection of the gene from its cognate enhancers.\(^{[103]}\) Thus, 3D chromatin rewiring by SVs may lead to gene loss of function through enhancer disconnection (Figure 4D). This mechanism is increasingly recognized as the cause of developmental disorders. For example, using whole genome
SVs creating a new gene-regulatory landscape. Structural variants that occur between TADs can rewire the TAD 3D architecture causing enhancer adoption and/or enhancer disconnection, resulting in gene misexpression and/or gene loss of function and disease. (A) Schematic representation of the wild-type genomic locus showing gene A expressed in the developing brain (blue) under the control of its tissue-specific enhancer (blue ellipse) and gene B expressed in the developing limb bud (red) under the control of its tissue-specific enhancers (red ellipses). Genes A and B and their respective tissue-specific enhancers are located in two TADs (blue and green triangles) separated by a TAD boundary (black polygon). The inter-TAD rearrangements represented in (B), (C), and (D) are indicated using double arrows. (B) Deletions involving TAD boundaries can cause TAD fusion and enhancer adoption, in which gene A is under the control of gene B enhancers, resulting in gene A misexpression in the limb (red). Chromosomal translocations can also cause TAD fusion and enhancer adoption. (C) Duplications involving TAD boundaries can create a new chromatin domain ("Neo-TAD" formation) and cause enhancer adoption, in which gene A is under the control of gene B enhancers, resulting in gene A misexpression in the limb (red). (D) Inversions involving TAD boundaries can cause TAD shuffling. This can lead to enhancer adoption, in which gene A is under the control of gene B enhancers, resulting in gene A misexpression in the limb (red). It can also lead to enhancer disconnection, in which gene B is separated from its cognate enhancers, resulting in gene B loss of function in the limb. Chromosomal translocations relocating TAD boundaries can also cause TAD shuffling, resulting in enhancer adoption and/or enhancer disconnection.

sequencing (WGS) in patients with congenital anomalies, a study systematically mapped the breakpoints of 273 balanced translocations at nucleotide resolution. A >7% of the balanced translocations disrupted TADs encompassing known haploinsufficiency disease-related genes.[104]

**Enhancer adoption: Rare or overlooked?**

Several studies have reported enhancer adoption as a mechanism that results in the ectopic expression of a gene as the underlying molecular mechanism of rare genetic disorders.[63,68,96,97,99,101,105,106] Since its discovery, it has been tempting to speculate that this mechanism could explain a significant fraction of unsolved cases. Although it is currently difficult to accurately estimate the prevalence of this disease mechanism due to our limited understanding of the basic mechanisms underlying enhancer-promoter communication, several studies suggest that this mechanism may not be as frequent as expected. Indeed, a recent study reported that gene expression is generally not altered when shuffling or fusing TADs, indicating a certain genome robustness against these events and that enhancer adoption is rare.[107] This study used highly rearranged chromosomes spanning around 75% of the *Drosophila* genome, including eight large nested inversions, smaller inversions, and thousands of deletions. Another study analyzed 288,457 somatic SVs using WGS data from 2658 cancers across 38 tumor types to understand their effects across TADs. This study revealed that only 1.4% of the boundary deletions fusing TADs, resulted in a change in expression in nearby genes more than twofold.[108] Moreover, a recent study investigated 11 individuals with CGRs, including germline chromothripsis, by combining short- and long-read genome sequencing with HiC and phased RNA-sequencing (RNA-seq). This study reported that while the majority of breakpoints reshuffle
TADs, most genes located 200-kb around a breakpoint do not show significant expression changes. These data suggest that the enhancer adoption mechanism may be a rare cause of disease.

INTERPRETING THE IMPACT OF SVs: 3D AND MORE

Overall, the discovery that SVs can rewire TAD 3D architecture and the communication between regulatory elements and genes without disrupting the gene or regulatory elements sequences has dramatically improved our interpretation of the pathogenic impact of SVs. Nevertheless, when assessing the pathogenic impact of SVs, it is important to consider additional critical factors. First, TADs are highly dynamic structures, and TAD boundary positions are not strictly conserved among cell types. Thus, cell-type-specific or tissue-specific HiC maps related to the diseased tissue should be considered. It is also essential to consider that TAD boundaries are not strictly impenetrable. Indeed, inter-TAD communication can also occur. This is exemplified by the inter-TAD communication involving the Lbx1 (Ladybird homeobox 1) and Fgf8 (Fibroblast growth factor 8) regulatory domains, although Lbx1 and Fgf8 exhibit divergent expression patterns. Interestingly, the communication between the two regulatory domains contributes at least partially to the disease mechanism underlying split-hand/foot malformation type 3 (SHFM3), associated with tandem duplications at the FGF8 locus.

When assessing the pathogenic impact of SVs, especially those disrupting TAD 3D architecture with TAD fusion, neo-TAD formation, or TAD shuffling, it is important to consider the cell-type-specific and/or tissue-specific expression profiles of the genes at the rearranged loci. This is particularly true when the candidate genes are not known to cause disease and/or their function is still unknown. Of note, around 20% of human proteins lack physiologically informative descriptions, which constitutes another major limiting factor in assessing the pathogenic impact of SVs and genetic variants as a whole. Moreover, it is essential to remember that even a single SV can have many consequences. For example, it has been shown that non-coding SVs can disrupt long-range gene regulation, resulting in a combination of both gene loss of function due to enhancer disconnection and gene misexpression due to enhancer adoption (Figure 4D). This is, for example, reported for the inversion involving the Eph4 and Wnt6 neighboring TADs causing congenital limb malformations. This inversion causes ectopic interactions between Wnt6 and the Eph4 enhancers, resulting in Wnt6 misexpression in the developing limb bud. At the same time, this inversion causes a disconnection between Eph4 and its cognate enhancers, resulting in Eph4 silencing in the developing limb bud.

Additional regulatory layers contribute to the specific and functional communication between genes and CREs beyond being in the same TAD. Deciphering these additional regulatory layers and understanding their underlying molecular mechanisms is expected to improve our interpretation of SVs' pathogenic impact significantly. For instance, striking examples in the literature illustrate the para-

**FIGURE 5** Inter-TAD communication contributes to the disease mechanism underlying SHFM3. (A) Schematic representation of the wild-type genomic locus showing the Lbx1 TAD in blue and the Fgf8 TAD in orange. The two TADs are separated by a TAD boundary (black polygon). A dashed triangle indicates communication between the Lbx1 and Fgf8 TADs. The red ellipses within the Fbxw4 gene represent the Fgf8 limb-specific enhancers (indicated with red arrows). The specific expression profiles of Lbx1, Btrc, and Fgf8 during embryonic limb development are also shown. Lbx1 is expressed in the muscle progenitor cells (blue), and Fgf8 is expressed in the Apical Ectodermal Ridge (AER) (orange). Btrc is weakly expressed in the developing limb bud (light green). (B) Schematic representation of the impact of the tandem duplications causing SHFM3 on 3D genome organization and enhancer-promoter communication. The intact Lbx1 TAD is duplicated, and a new TAD is generated between the two Lbx1 TADs. This “neo-TAD” corresponds mainly to the Fgf8 TAD (orange) and a duplicated genomic region centromeric of the Lbx1 TAD boundary (brown). The dashed red arrows indicate the inter-TAD communication, resulting in ectopic interactions between the Fgf8 limb-specific enhancers within the “neo-TAD” and the Lbx1 and Btrc genes within one or both of the neighboring Lbx1 TADs. This causes misexpression of Lbx1 and Btrc in the developing limb bud in an Fgf8-like pattern, likely contributing to the disease. Of note, Fgf8 expression is unaffected while Lbx1 expression also increases in the muscle progenitor cells due to gene copy number change.

dox in promoter responsiveness to non-cognate enhancers. Among them are the MEF2C (Myocyte Enhancer Factor 2C) and PAX3 promoters. Several translocations involving different chromosomes distal to MEF2C have been reported in the literature. They all disrupt the MEF2C TAD 3D architecture creating a new MEF2C-regulatory
landscape specific to each rearrangement. Nevertheless, all reported SVs cause a similar haploinsufficiency phenotype due to haploinsufficient MEF2C expression. These data suggest that the MEF2C promoter is impervious to the enhancer activity of non-cognate CREs. Conversely, the PAX3 promoter seems permeable to the enhancer activity of non-cognate enhancers. Indeed, deletions and translocations involving the PAX3 locus lead to ectopic interactions between the PAX3 promoter and non-cognate enhancers. This results in gene misexpression or oncogene activation, causing disease. These examples suggest that understanding which types of promoters are permeable to the enhancer activity of non-cognate CREs and which are not, as well as whether this is absolute, would help improve our predictions of the impact of SVs on gene expression.

In addition to enhancers, several studies indicate that silencers are abundant in mammalian genomes and can repress gene expression by physically interacting with their target genes. It is thus tempting to speculate that SVs disrupting silencer elements or silencer-gene communication can also cause disease. For example, a recent study reported a mechanism termed “silencer hijacking,” whereby SVs relocate a repressive element (H3K27me3-rich region) close to the tumor suppressor gene 1IKZF2 (IKAROS Family Zinc Finger 2), resulting in its expression downregulation in Leukemia.

Taken together, the latest advances in understanding the basic principles underlying gene regulation and the genome’s spatial organization highlight the challenging task of interpreting and predicting the pathogenic impact of SVs in clinical diagnostics. The newly reported mechanism through which SVs can cause disease, which we will discuss in the following section, undoubtedly corroborates this.

**SVs Creating a Novel Gene-Intergenic Fusion Transcript**

Somatic SVs are a common driver of fusion gene formation. Indeed, gene fusions have been extensively studied in cancer. They typically involve one or more coding sequences resulting in chimeric mRNA transcripts that can form oncogenic “neo-antigens.” Several SVs driving gene fusion formation have also been involved in rare genetic disorders, including congenital anomalies, autism, and intellectual disability. Recently, a study described a complex SV driving a novel gene-intergenic fusion transcript formation as the underlying disease mechanism of an autosomal dominant hereditary motor neuropathy DHMN1. Using WGS, the authors identified a 1.35-Mb duplication of chromosome 7q42.3 inserted in the reverse orientation at chromosome 7q36.2. The inserted sequence fragment contained four protein-coding genes and their regulatory elements, as well as the upstream regulatory elements and a partial copy of the UBE3C gene (Figure 6A).

As other genes and their respective regulatory elements (Figure 6B). Interestingly, in addition to rewiring 3D chromatin architecture, the authors demonstrated that the SV produces a novel gene-intergenic fusion transcript in which the UBE3C partial copy is transcribed from the reverse strand and incorporates a terminal pseudo-exon from the adjacent non-coding DNA sequences within the DHMN1 locus (Figure 6B). Although the novel gene-intergenic fusion transcript does not undergo nonsense-mediated decay, it is unclear whether it encodes for a protein. Nevertheless, by overexpressing the novel gene-intergenic fusion transcript in Caenorhabditis elegans, the authors demonstrated that this transcript likely contributes to the disease mechanism through a dominant-negative reduction of the wild-type UBE3C protein. This novel disease mechanism may be overlooked and could explain the pathogenic impact of other SVs reported in the literature. For example, it is tempting to speculate that the 78-kb insertion from chromosome 8 associated with Charcot-Marie-Tooth neuropathy CMTX3 causes the disease through a similar mechanism. This complex insertion corresponds to a duplication of chromosome 8q24.3, containing a partial copy of the ARHGAP39 (Rho GTase Activating Protein 39) gene, inserted into the SOX3 (SRY-Box Transcription Factor 3) TAD at chromosome Xq27.1. In this case, the partial copy of the ARHGAP39 gene may be under control of SOX3 enhancers generating a novel ARHGAP39-intergenic fusion transcript expressed in SOX3 territories and causing disease.

Fusions involving intergenic/non-coding DNA sequences that result in cryptic (pseudo) exon formation (Figure S1) have only been reported by a few studies. Nevertheless, they may be more prevalent than expected. Indeed, most genomic breakpoints in fusion genes are intronic or intergenic. However, they are not typically present in mRNA or protein-coding sequences, making the detection of gene-intergenic fusions difficult. Thus, to improve the interpretation of the pathogenic impact of SVs and the identification of disease-causing gene-intergenic fusions in rare genetic disorders, it is essential to use a combination of WGS and RNA-seq patients’ data, as well as develop novel algorithms for the detection of novel gene-intergenic fusion transcripts. Interestingly, a recent study reported in the literature demonstrated that using rRNA-minus RNA-seq datasets is advantageous for identifying novel gene-intergenic fusion transcripts. This study revealed, nevertheless, that only a fraction of all genomic rearrangements (~7%) is transcribed, and an even smaller fraction is causal for fusion genes. These data suggest that gene-intergenic fusion transcripts may be a rare cause of genetic disorders.

**Conclusions**

The latest technological developments have significantly improved our interpretation of the pathogenic impact of SVs. For many SVs, however, the underlying disease mechanism still needs to be fully resolved, and for many others, the prediction of their pathogenic impact still needs to be improved. Indeed, several challenges remain, hampering the medical interpretation of SVs, predicting their phenotypic consequences, and characterizing their underlying disease mechanism.
FIGURE 6  SVs producing a novel gene-intergenic fusion transcript. Structural variants that disrupt protein-coding sequences and simultaneously impact intergenic non-coding DNA sequences can produce novel gene-intergenic fusion transcripts at the SV breakpoint. (A) Schematic representation of the genomic region of chromosome 7q36.2-q36.3 (Wild type) involving the formation of the CGR associated with DHMN1 (DHMN1). The blue segment represents the DHMN1 linkage region. The orange segment represents the DHMN1 insertion sequence. The black arrow indicates the orientation of the DHMN1 insertion sequence. The 7q36.3 duplicated sequence (orange) is inserted in an inverted orientation within the DHMN1 locus. This duplicated sequence contains four protein-coding genes (RNF32, LMBR1, NOM1, and MNX1), their regulatory landscapes, and a partial copy of the UBE3C gene and its regulatory landscape. Letters indicate the proximal (a) and distal (b) ends of the insertion site, and the proximal (c) and distal (d) ends of the 7q36.3 duplicated sequence. (B) The DHMN1 insertion sequence impacts 3D genome organization. It creates a new TAD containing the four protein-coding genes, their regulatory landscapes, and the partial copy of the UBE3C gene and its regulatory landscape (orange). This “Neo-TAD” also contains non-coding DNA sequences adjacent to the insertion breakpoint at the 7q36.2 locus (blue). The CGR results in the generation of a novel gene-intergenic fusion transcript (UBE3C-IF). The UBE3C partial copy is transcribed from the reverse strand and incorporates a pseudo-exon from the intergenic non-coding DNA sequences adjacent to the insertion breakpoint (a).

First, although high-throughput technologies have produced a vast amount of data in terms of protein validation, protein-protein interactions, genetic variants, gene/protein expression, and 3D structure, contributing to developing a more precise picture of the human proteome, many human proteins still have only a vague or speculative function annotated in UniProtKB/Swiss-Prot[151] and about 10% of human protein-coding genes have no function annotated at all, either predicted or experimentally confirmed[152] The functional characterization of human proteins is a huge challenge currently undertaken using systems biology approaches combining high-throughput omics technologies and bioinformatics. However, we estimate that there were between 8 and 10 papers describing newly characterized human proteins published each month in the last five years, corresponding to a decrease in the number of uncharacterized proteins by only 25% in 5 years[152] Thus, the scientific community should encourage collaborative projects aiming at functionalizing uncharacterized proteins to expedite the missing functional annotation of protein-coding genes. The Human Proteome Project (HPP), an international consortium dedicated to cataloguing the 20,230 proteins list, understanding the complexity of the human proteome, and making human proteomics an integrated complement to genomics and other “omics” across clinical, biomedical, and life sciences, has just committed to move in this direction. In its first phase, which was focused on protein validation, the HPP project federated mostly proteomics experts. To succeed in its second phase, it will need to recruit specialists in various other human and model organism biology fields.

Second, to detect and analyze structural variation in non-coding regulatory elements, localizing these regions in the genome and validating them experimentally is necessary. Online databases are the most comprehensive and convenient method to identify regulatory regions throughout the genome because these integrate data from many assays based on biochemical properties (e.g., transcription factor (TF) binding, open chromatin, and histone modifications) to locate functional elements accurately. The main limitation of these databases is that most of the regulatory elements for which they provide information, especially enhancers, are simply putative, predicted regulatory regions, of which only a tiny portion has been experimentally validated. Examples include the Encyclopedia of DNA Elements (ENCODE database), the Functional Annotation of the Mammalian Genome project (FANTOMS), the PsychENCODE Consortium, and the
machine learning tool RefMap.\(^{153,154}\) While we expect new techniques for regulatory element discovery to emerge in the near future, the scientific community should encourage the development of high-throughput projects aiming at experimentally validating in vitro and/or in vivo the cell-type and tissue-specific predicted regulatory regions.

Third, as illustrated by the examples discussed above, disease-causing SVs do not necessarily disrupt enhancer or gene sequences but instead affect the ability of enhancers to activate their target genes specifically. Despite the tremendous progress in the field of enhancer biology during the past few decades, the biochemical processes whereby enhancers communicate specifically with their target genes still need to be clarified.\(^{155}\) Thus, we anticipate that elucidating the molecular and genetic principles controlling enhancer specificity will help us to better predict the effect of genetic variation in human diseases.

Fourth, the interpretation of the pathogenic impact of SVs may, in some cases, only be achieved by using a multi-omics approach, such as the case for the SVs producing novel gene-intergenic fusion transcripts. This combines genomic, epigenomic, transcriptomic, and functional genomic information. Such an approach has been proposed to unify variant detection and interpretation.\(^{156}\) It should be considered by clinicians, clinical geneticists, and scientists to improve patients’ diagnosis and understanding the basic mechanisms of disease. Indeed, while whole genome and exome sequencing have increased the identification of causal variants compared to single gene testing alone, with diagnostic rates of approximately 50% for inherited diseases,\(^{157}\) integrated multi-omics analysis may further increase diagnostic yield. Moreover, advances in omics technologies have begun to enable personalized medicine at an extraordinarily detailed molecular level. These technologies have contributed medical advances that have begun to enter clinical practice. However, each technology individually cannot capture the entire biological complexity of most human diseases. Integration of multiple technologies has thus emerged as an approach to provide a more comprehensive view of biology and disease. We direct the reader to an excellent review discussing the importance of integrative omics for health and disease.\(^{158}\)

Overall, substantial obstacles must be overcome to improve our prediction of the pathogenic impact of SVs and their medical interpretation. Overcoming these challenges becomes crucial. Indeed, recent studies have shown that SVs alter the expression of nearby genes with larger effect sizes than SNVs or small indels, indicating that SVs are more likely to be pathogenic.\(^{159,160}\) Since these obstacles are not overcome, it is essential to experimentally validate the functional consequences of disease-associated SVs to assign pathogenicity. This is not only true for de novo SVs but also inherited variants. Experimentally validating the functional consequences of disease-causing SVs paves the way for characterizing novel disease mechanisms by which SVs exert their pathogenic impact, which is crucial for developing accurate computational prediction tools for SV interpretation.

ACKNOWLEDGMENTS

The authors thank Giulia Cova and Juliane Glaser for information, discussions, and corrections during the preparation of this manuscript. This study was supported by a grant from the Deutsche Forschungsgemeinschaft (MU 880/16-1) to S.M. L.A. was supported by the FOXG1 Foundation (grant PS.7905).

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were generated or analyzed in this study.

ORCID

Stefan Mundlos https://orcid.org/0000-0001-7824-1945

REFERENCES

Science, 330, 641–646.
Science, 330, 641–646.


SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.