INVITED REVIEW

Looking beyond the genes: the role of non-coding variants in human disease

Malte Spielmann\textsuperscript{1,2,3} and Stefan Mundlos\textsuperscript{1,2,3,*}

\textsuperscript{1}Max Planck Institute for Molecular Genetics, RG Development \& Disease, 14195 Berlin, Germany,
\textsuperscript{2}Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany and
\textsuperscript{3}Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité Universitätsmedizin Berlin, 13353 Berlin, Germany

*To whom correspondence should be addressed at: Tel: +4930450569122; Fax: +4930450569131; Email: mundlos@molgen.mpg.de

Abstract

Over the past decades the search for disease causing variants has been focusing exclusively on the coding genome. This highly selective approach has been extremely successful resulting in the identification of thousands of disease genes, but ignores the functional and therefore disease relevance of the rest of the genome. Dropping sequencing costs and new high-throughput technologies such as ChIP-seq and chromosome conformation capture have opened new possibilities for the systematic investigation of the non-coding genome. These data have revealed the importance of non-coding DNA in fundamental processes such as gene regulation and 3D chromatin folding. Research into the principles of chromatin folding has revealed a domain structure of the genome, called topologically associated domains that provide a scaffold for enhancer promoter contacts. Non-coding mutations that affect regulatory elements can affect gene regulation by a loss of function, resulting in reduced gene expression, or a gain of function resulting in gene mis- or overexpression. Structural variations such as deletions, inversions or duplications have the potential to disturb normal chromatin folding. This may lead to the repositioning or disruption of topological associating domains and the relocation of enhancer elements with consecutive gene misexpression. Several recent studies highlight this as important disease mechanisms in developmental disorders and cancer. Therefore, the regulatory landscape of the genome has to be taken into consideration when investigating the pathology of human disease. In this review, we will discuss the recent discoveries in the field of non-coding variation, gene regulation, 3D genome architecture, and their implications for human genetics.

Introduction

Beyond the exome

Medical genetics is being transformed by next-generation sequencing (NGS) technologies enabling simultaneous investigation of all relevant disease genes, all protein-coding genes, and even the entire genome (1,2). Due to dropping sequencing costs and constant improvements of sequencing technologies, it is likely that in the near future whole-genome sequencing (WGS) will be feasible and affordable in a diagnostic setting (3,4). Pilot studies have shown that WGS can detect a broader range of genetic variation than other sequencing approaches, not only single nucleotide variants (SNVs) and insertion or deletions (indels), but also structural variants such as copy number variants (CNVs), inversions and translocations (5). However, if WGS is to be implemented as a "one test for all" diagnostic strategy, fundamental obstacles remain. The "regulatory code" of the non-coding genome, determining whether and how a given genetic variant affects the function of a regulatory element, remains poorly understood (5). To address this challenge, the
Encyclopaedia of DNA Elements (ENCODE) consortium was established in 2004, with the aim of systematically investigating and annotating the 98.5% of the genome that is non-coding (6). The collective findings of the ENCODE consortium suggested that around 80% of the genome contains elements linked to some biochemical function (6,7). Moreover, the space between genes is filled with cis-regulatory elements such as enhancers, silencers, promoters, and numerous previously overlooked regions containing untranslated RNA transcripts that have a regulatory role (8,9). These findings are directly relevant to human genetics, since most studies investigating the genetic basis of intellectual disability and developmental delay are focused on coding variants, but have failed to provide clear answers for over 40% of the families studied (10–12). This suggests that a large proportion of disease cases may be caused by alterations outside of the coding regions.

In this review, we will discuss recent discoveries in the field of non-coding variation, gene regulation, 3D genome architecture, and their implications for human genetics. We will focus on the latest findings and highlight the role of 3D genome architecture in human disease.

Long range control via enhancers

Gene regulation usually involves two distinct types of cis-acting elements: the promoter consisting of the core promoter and nearby regulatory elements, and more distal regulatory units, so called enhancers, silencers of locus control regions (9). While the promoter is generally located less than 1 kb from the transcription start site, enhancers can act over long distances, in some cases more than 1 Mb (Fig. 1) (13). Promoters and enhancers have overlapping functional properties and share many characteristics such that they have been considered as a single class of regulatory elements. Both represent docking sites for transcription factors (TFs) regulating the activity of the promoter and enhancer (14). In the current concept, enhancers are primed by sequence specific pioneering TFs, providing an accessible platform for the recruitment of general transcription factors. Together with RNA polymerase II, they form the preinitiation complex which in turn recruits further proteins including the mediator complex. The large distance between enhancers and promoters requires direct or indirect mediators of communication between these elements (15). One possible scenario involves physical contact enabled by looping of the intervening DNA sequence mediated by proteins such as mediator, CTCF, and cohesins, bridging enhancers and promoters together as a chromatin structure. Once the bridge is formed, expression is regulated accordingly (16–19).

The physical presence of looped chromatin has been demonstrated by fluorescent in situ hybridization (FISH) experiments, in which the distance between two given chromosome fragments can be measured. The development of chromosome conformation capture (3C) and its high-throughput derivatives (e.g. 4C, 5C, ChIA-PET and Hi-C) (19,20), have made it possible to explore in more detail the 3D architecture of the genome by quantifying chromatin looping via a proximity ligation assay. For example, it was shown at the SHH locus that, even when its limb enhancer element is deleted, loop formation is still observed (21). In another interesting study, Deng and colleagues showed that chromatin looping can also be induced using artificial zinc fingers, with subsequent activation of developmentally silenced genes (22).

Topologically associating domains (TADs) and the 3D structure of chromatin

The human genome is often depicted in a linear fashion, despite the fact that the two meters of DNA needs to be tightly folded in order to be packed into the nucleus. Recent data indicate that chromatin folding, and thus the 3D organization of the genome in the nucleus, is directly linked to central aspects of gene regulation. Major insights have been gained by Hi-C, an expansion of the 3C technologies, using a purification of ligation products followed by massively parallel sequencing (17,23).

Deep sequencing of Hi-C libraries has revealed sub-orders of chromosome organization at the megabase scale, designated as topologically associating domains or TADs (24–26) (Fig. 2). Interestingly, these structures are conserved among species, cell types, and tissues (24,25). They appear to form a regulatory scaffold for the genome restricting the contacts an enhancer may have, thereby preventing promiscuous enhancer activity (24,26–28). An example is shown in Figure 2 (Fig. 2).

TADs appear to be structures that both promote contacts within a domain and at the same time prevent contacts between neighbouring domains. Important for the separation of neighbouring activities are so called boundary regions, initially inferred from Hi-C data sets by measuring abrupt changes in the directionality of contacts (24). Boundaries are strongly enriched in the architectural proteins CTCF (CCCTC-binding factor) and cohesin (29). CTCF seems to be crucial for boundary function, as its ablation affects TAD organization by decreasing intr-domain and increasing inter-domain contacts (30), although its presence is not exclusive to boundaries (24). Furthermore, a striking correlation was found between CTCF motif orientation and looping, taking place in more than 90% of the cases in a convergent manner (27,31). CRISPR-Cas9 genome editing experiments have further supported the role of CTCF in TAD boundary organization (25,32).
These data demonstrate that TADs and their boundaries play a key role in controlling gene expression. Structural variations such as deletions, inversions or duplications thus have the potential to interfere with the TAD structure by disrupting or repositioning boundaries. Recent studies highlight this as an important disease mechanism (33–35).

Loss of function mutations in regulatory elements

Similar to coding mutations, single nucleotide variants in enhancer elements can result in a loss of function or a gain of function of the target gene (Fig. 3). Given the temporal and spatial specificity of enhancers and the complexity of expression patterns of many developmental genes, the loss of a specific regulatory element might result in the loss of a specific expression domain, provided that it is not compensated by other regulatory elements. Such a specific loss can result in a phenotype consisting of a subset of features observed with a complete loss of the gene’s function, e.g. by a coding mutation. An example is isolated pancreatic agenesis which is caused by mutations in a highly conserved pancreas specific enhancer located 25 kb downstream of PTF1A (36). Mutations in PTF1A itself result in syndromic form of pancreatic agenesis, featuring severe neurological symptoms, whereas PTF1A enhancer mutations result in an isolated pancreatic anomaly (37). A similar example was described for PAX6 in patients with aniridia (Fig. 4A) (38). Regulatory loss of function point mutations have also been reported for SOX9. Mutations in SOX9 cause campomelic dysplasia, a lethal skeletal dysplasia with male to female sex reversal (39,40). Deletions of the proposed testis enhancer of SOX9 result in sex reversal but no skeletal phenotype, whereas deletions and point mutations further upstream cause Pierre-Robin syndrome, a condition with growth defects of the cranial skeleton, but normal sexual development (Fig. 4B). Further examples have been reported for SHH (41) and TBX5 (42).

Often, deletions remove coding sequence, but the phenotype can better be explained by the loss of a non-coding cis-regulatory element. A study by Birnbaum and co-workers shows that the deletion of coding exons can also result in a loss of function and a gain of function of the target gene. A loss of a regulatory element by a point mutation or a deletion is expected to result in the loss of a specific expression domain, while a gain of function mutation or a duplication of an enhancer element can increase the frequency of enhancer-promoter interactions causing tissue specific gain of expression and misexpression.

Figure 2. The topological-associated domain (TAD) architecture of the EphA4 locus. The genomic landscape of the EphA4 locus: Hi-C interactions from mouse ES cell are shown in a heat map in which each dot reflects two interaction pairs of DNA. The resulting interaction profile shows the formation of triangles (dotted lines) that represent individual TADs. There is a high degree of interaction within each TAD but little contact between TADs. The EphA4 locus appears as one large TAD with transitions on either side demarcating boundary regions where the interactions diminish and orientate to the other direction. The binding profile of CTCF transcription factor is shown below. There is an enrichment of CTCF binding at the boundaries and at gene promoters. The 4C-seq profiles of the viewpoints Ihh, EphA4, and Pax3 are depicted below. Note that the interaction profiles are restricted to the respective TADs (24,34,76).

Figure 3. The effects non-coding mutations on gene expression. Similar to coding mutations, single nucleotide variants (SNV) in enhancer elements can result in a loss of function and a gain of function of the target gene. A loss of a regulatory element by a point mutation or a deletion is expected to result in the loss of a specific expression domain, while a gain of function mutation or a duplication of an enhancer element can increase the frequency of enhancer-promoter interactions causing tissue specific gain of expression and misexpression.
Gain of function mutations in regulatory elements

An example for regulatory gain of function mutations has been described at the Sonic hedgehog (SHH) locus (Fig. 4C). Hedgehog genes encode for a developmentally important family of signaling molecules. In the developing limb, SHH is expressed at a restricted site, the zone of polarising activity, which is tightly regulated by the ZRS, a regulatory region approx. 1 Mb 5’ of SHH (46–48). Several conditions with various types of polydactyly, including triphalangeal thumb polysyndactyly syndrome, Haass type polysyndactyly, and Werner syndrome, have been associated with point mutations in the ZRS. As shown in transgenic mice, these point mutations result in a misexpression of SHH in the anterior part of the limb bud, which in turn results in additional digits (polydactyly) and fusion of digits (syndactyly) (48). It was shown that ZRS mutations alter the balance of transcription factor binding sites by either creating new sites or inactivating existing sites (49). Interestingly, very similar phenotypes have been described for duplications encompassing the ZRS, indicating that some enhancer elements may be dose sensitive (46,50).

However, as shown for the point mutations, duplications of the ZRS do not only result in an increase of expression, but also in misexpression, as indicated by the polydactyly phenotype. Still, the precise mechanism of these mutations needs to be proven, as no animal models exist so far.

A similar example was described in patients with female-to-male sex reversal carrying tandem duplications upstream of SOX9 (51). These duplications are thought to result in a gonad specific gain of function of SOX9, while overlapping deletions have been identified in individuals with male-to-female sex reversal (52). Our own lab has also described several examples of tandem duplications of non-coding cis-regulatory elements associated with congenital limb malformation: Duplications of a 5 kb limb enhancer element within the TAD BMP2 cause brachydactyly type A2, a shortening of the digits (53,54). Several similar tandem duplications of non-coding cis-regulatory elements were described 5’ of IHH in patients with syndactyly and craniosynostosis (55).
Non-coding variants in complex traits and diseases

Disease-associated nucleotide variants identified in genome-wide association studies (GWAS) are rarely found in coding regions. Instead, most disease-associated index SNPs are located in non-coding regions of the genome (56). The recent introduction of CRISPR/Cas9 genome editing has opened new opportunities to investigate also common non-coding variants located in cis-regulatory elements (57, 58). Claussnitzer and colleagues recently showed that the FTO allele, which shows the strongest genome-wide association signal for obesity, acts as a gain of function (59). Using CRISPR/Cas9 genome editing, they showed that the disease-associated single-nucleotide variant rs1421085 T-to-C disrupts a conserved motif for the ARID5B repressor, which unleashes a preadipocyte enhancer, leading to a doubling of IRX3 and IRX5 expression during early adipocyte differentiation.

Another recent study identified a common Parkinson disease-associated risk variant in a non-coding distal enhancer element, regulating the expression of α-synuclein (SNCA), a key gene implicated in the pathogenesis of Parkinson disease (60). They demonstrated that the disease-associated risk variant prevents the binding of two repressive transcription factors EMX2 and NKX6-1 to an enhancer element, resulting in transcriptional upregulation of SNCA.

Another interesting example is a non-coding variant at the 8q24 locus that associates with cleft lip with or without cleft palate (61). Recent mouse studies have shown that this interval contains very remote cis-acting enhancers that control Myc expression in the developing face (62).

TAD disruption and position effects

The term “position effect” has long been used for balanced translocations and other larger structural rearrangements that could not be explained by the content of the variants alone. The discovery of the topological domain architecture of the genome and our increased knowledge about enhancers and chromatin-folding, enables us to better understand the mechanisms underlying potential “position effects”. It is now becoming clear that structural variations have the potential to alter the topological domain architecture of the genome by deleting, or misplacing TAD boundaries, thereby allowing enhancers from neighbouring domains to ectopically activate genes causing misexpression and disease (33). This mutational mechanism was termed enhancer adoption (35).

An early example suggestive of such a disease mechanism is Liebenberg syndrome, an autosomal-dominant condition in which the arms partially acquire morphological characteristics of the legs (63). We identified several deletions of a cis-regulatory boundary element close to PITX1, an important gene for hindlimb development. Without this boundary or insulator
element, an enhancer element with fore- and hindlimb activity from a neighbouring regulatory domain is free to act on PITX1, thereby inducing ectopic expression of PITX1 in the forelimb causing the phenotype.

The relevance of TADs for genomic integrity and disease was exemplarily demonstrated at the EPHA4 locus (34). A deletion encompassing EPHA4 at the telomeric side of the TAD was shown to result in brachydactyly (short digits) (Fig. 5A), whereas an inversion and a duplication on the centromeric side involving part of the EPHA4 TAD were shown to be associated with a complex form of syndactyly (Fig. 5C). In addition, a family with a duplication was investigated, whose breakpoints partially overlapped with a deletion present in a mouse mutant where 800 kb between the Ihh (Indian hedgehog) gene and the most centromeric part of the Epha4 TAD are deleted. These two rearrangements, despite being of a different nature, both resulted in severe polydactyly (7 and more fingers). The results indicated that different structural variations can result in similar, but distinct malformation syndromes, and thus likely to be independent of the gene in this locus, i.e. EPHA4. CRISPR/Cas9 genome editing was used to re-engineer (recapitulate) the human deletions and inversions in mice (58). We could show that an enhancer cluster located within the Epha4 TAD, which normally regulates Epha4 expression in the limb bud, was now ectopically activating different genes, depending on the breakpoint, i.e. Pax3 (Paired box3) in brachydactyly (Fig. 5A), Wnt6 (Wingless-type MMTV integration site family, member 6) in syndactyly (Fig. 5C), and Ihh in polydactyly (54). All mutant phenotypes were caused by gene misexpression, due to the ectopic interaction of the enhancers with the target gene. However, this interaction was dependent on the disruption of one of the TAD boundaries (telomeric for Pax3, centromeric for Ihh and Wnt6), since shifting the deletions so that they did not include the boundary, resulted in no ectopic interaction and normal mice. Furthermore, a similar pathomechanism was identified causing autosomal dominant adult-onset demyelinating leukodystrophy (ADLD) (Fig. 5B) (64,65) and mesomelic dysplasia (severe shortening of the middle segment of the lower limbs) (66). It was suggested that up to 11% of all deletions in the DECIPHER database results in enhancer adoption (35).

Recent data indicate that TAD reorganization and disruption play a key role in the pathogenesis of cancer, which is not surprising since misregulation of genes is a common feature in cancer (67–69). Some examples are AML/CLL, gliomas, and medulloblastoma, where enhancer activation or enhancer hijacking has been identified as the major mechanism of disease (70–72). In fact, recent studies show that TAD boundary-associated CTCF loops are frequently mutated in colon cancer, allowing promiscuous activity of genes (73,74).

These results suggest that non-coding mutations may contribute to a substantial number of human disease phenotypes and should thus be taken into account for the medical interpretation of mutations and copy number variants.

**Conclusion for the Clinic**

Recent findings show that mutations and structural variations outside of the coding genome can interfere with the normal gene regulation by regulatory loss or gain of function mutations. Structural variations have also the potential to interfere with the TAD structure of the genome by shifting regulatory elements between domains and/or by interfering with the position of boundaries. For example, deletions that include a boundary element can result in ectopic enhancer-promoter interactions thereby inducing gene misexpression and disease. Therefore, the cis-regulatory landscape and TAD architecture of the genome have to be taken into consideration when investigating the pathology of human disease.

The next challenge will be the medical interpretation of single nucleotide variants from clinical WGS data. The sheer number of non-coding variants in each individual and generation make classical functional work-up strategies impossible (3). Further knowledge about the non-coding genome and more experimentally validated non-coding variants are needed to develop computational prediction tools for the medical interpretation of non-coding mutations.

**Acknowledgements**

We thank all members of the Mundlos laboratory for helpful discussions. We thank Dario G. Lupiáñez for his comments and providing the 4C-seq data. We also thank Thomas Spletstoesser (www.scistyle.com) for assistance in figure design.

**Conflict of Interest statement.** None declared.

**Funding**

Work in SMs laboratory is funded by the Deutsche Forschungsgemeinschaft (DFG), the Berlin Institute for Health (BIH), and the Max Planck Foundation (MPF). MS was supported by the Deutsche Forschungsgemeinschaft (SP1532/2-1) and by a fellowship of the Berlin-Brandenburg School for Regenerative Medicine.

**References**


