An evolutionary perspective of regulatory landscape dynamics in development and disease

Martin Franke1,* and José Luis Gómez-Skarmeta1,*

1Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide/Junta de Andalucía, Sevilla, Spain.

*Correspondence to: jlgomska@upo.es, mfra2@upo.es

Key words: Gene Regulation, cis-Regulatory Elements, Regulatory Landscapes, TADs, Evolution, Development, Disease, Genome Instability

Abstract

The organization of animal genomes into topologically associating domains (TADs) provides a structural scaffold in which cis-regulatory elements (CREs) operate on their target genes. Determining the position of CREs and genes relative to TADs has become instrumental to trace gene expression changes during evolution and in diseases. Here we will review recent studies and discuss TADs as structural units with respect to their conservation and stability during genome reorganization. Furthermore, we describe how TAD restructuring contributed to morphological novelties during evolution but also their deleterious effects associated with disease. Despite considering TADs as structural units, the nested and dynamic scaffold within TADs contribute to tissue-specific gene expression, implying that such changes can also contribute to gene expression differences during evolution.

Introduction

Gene transcription is regulated by cis-regulatory elements (CREs) that can be found at large distances from the genes they control [1,2,3]. The genome territory containing all CREs regulating a particular promoter is defined as the regulatory landscape (RL) of the gene. The widespread distribution of CREs over extended RLs requires a precise coordination to bridge these long genomic distances and to ensure precise gene activation. A mechanistic framework on how cis-regulatory information is organized in the genome largely derives from chromosome conformation capture (3C) techniques, which measure physical proximity and frequency of DNA sequences in the nucleus [4]. High-throughput 3C approaches such as Hi-C [5,6] have revealed that, at the sub-megabase scale, mammalian genomes form so called topologically associating domains (TADs) [6,7]. High frequency chromatin interactions are favored within TADs and insulated between neighboring TADs (reviewed in [8]). This organization facilitates robust gene expression control by remote CREs within TADs and prevents non-specific CRE-promoter interactions between neighboring domains [3] (Figure 1a). Remarkably, functional studies using random insertions of a naive reporter gene in mouse showed that TADs overlap to a large degree with RLs [9,10]. The discovery of this topological organization of chromatin has
tremendously improved our view on how the regulatory information is organized in the genome. Here we will discuss recent work, revealing the relations between TADs and genome evolution, human diseases and gene expression during cell differentiation.

**Evolutionary conservation of TADs and regulatory landscapes**

Comparative analysis of Hi-C data in several species has demonstrated that TADs and their boundaries have been maintained in conserved syntenic regions during mammalian evolution, suggesting that TADs and even TAD boundary locations may be conserved features of animal genomes [6,11]. In support of this, it has been shown that the TAD boundary bisecting the Six gene clusters is not only conserved in all vertebrates, but is also present in echinoderms, indicating a deep evolutionary origin of some TADs [12]. However, due to the lack of Hi-C data in non-mammalian vertebrates and most invertebrates, it is not feasible to determine how general this conservation is. Research studying evolutionary conserved non-coding elements (CNEs) has provided some insight on this issue. CNEs are arranged in syntenic arrays especially around key developmental genes [13,14] and their collinearity is referred to as genomic regulatory blocks (GRBs) [15,16]. Recent work comparing GRBs between species shows that GRB boundaries appear as conserved features among all vertebrates [17-]. Strikingly, the identified GRB boundaries are highly predictive for TAD boundaries both in humans and in *Drosophila* and largely define the RLs of the corresponding developmental genes. Comparisons of GRB-TADs of homologous regions for key developmental regulators in human and *Drosophila* show similar patterns of topological constraints, suggesting that the basic chromatin structure of these loci has existed over hundreds of millions of years in bilaterian genomes [17-]. These observations strongly suggest that, at least around developmental genes, TADs are evolutionarily conserved genomic features, likely required to maintain the integrity of *cis*-regulatory relationships.

**Mechanism of TAD reorganization during evolution**

During the course of evolution, animal genomes undergo extensive reorganization, changing the relative order of syntenic blocks and TADs. However, this reorganization is biased as evolutionary-associated breakpoints are non-randomly distributed across different lineages and overlap more often than expected [18,19]. Reconstruction of ancestral genomes and statistical modeling of several chromatin features link the putative breakpoints to open chromatin and chromatin regions that physical interact [20]. It is therefore plausible that in addition to a negative selection against TAD disruption, TAD boundaries themselves could be hotspots for recombination [20,21]. Recent advances in disentangling the molecular processes involved in TAD formation in mammals demonstrate how genomes can be reorganized while preserving the modular nature of TADs. High resolution Hi-C has revealed that TAD boundaries flanking TADs strongly interact with each other, forming anchors of chromatin loops. These loop
anchors have been found to depend on the binding of architecture proteins such as CTCF and Cohesin [22,23]. The depletion of either of these two components in mammalian cells disrupts these loops, reduces internal TAD contacts and elevates interactions between insulated TADs, resulting in alteration of gene expression [24,25]. The recently proposed loop extrusion model helps to explain these observations [22,26] (Figure 1b). The model proposes that ring-shaped Cohesin complexes bind to chromatin and extrude the chromatin fiber to form chromatin loops. The extrusion is transiently stabilized, when Cohesin reaches CTCF proteins bound at distally located binding sites in head-to-head configuration. The process is likely to be dynamic, whereby the Cohesin complex extrudes chromatin, terminates at CTCF, falls off and starts again. The loop extrusion model can interpret both the high internal interaction frequency and the insulation property of TADs, as the dynamic process of extrusion increases the probability of CRE-promoter contacts on the extruded chromatin within TAD [22,26].

Moreover, loop extrusion likely causes knots and torsional stress ahead of the Cohesin complex as chromatin progressively extrudes [27]. Topoisomerase II beta (TOP2B) physically interacts with CTCF and Cohesin and it has been hypothesized that TOP2B-mediated double strand breaks (DSBs) helps solving topological stress associated with chromosome architecture [28]. A recent report shows that TOP2B-mediated DSBs coincide with the strength of interactions at loop anchors and with binding strength of CTCF/Cohesin [27]. Intriguingly, DSBs preferentially localize in open chromatin, few nucleotides adjacent to the CTCF-binding sites outside the loop and, in the majority of the cases, on both sides of the loop anchor (Figure 1b). This suggests that the forces that act on the chromatin fiber during TAD formation can be released by TOP2B, while preserving TAD integrity. This observation provides a rationale for the evolutionary constraints in respect to CTCF-binding sites at TAD boundaries and in contrast to the binding-sites located within TADs. The former are more conserved at the sequence level, they exhibit a stronger binding strength and are constitutively bound by CTCF throughout development [11,23]. Therefore, these features could be a direct reflection of forces acting on TAD boundaries to preserve cis-regulatory integrity within TADs. TAD boundaries are vulnerable to DSBs, which can lead to a modular reshuffling of TADs during genome reorganization [27] (Figure 1c). Indeed, Hi-C analysis in different mammalian species has shown that all identified synteny breaks map to boundaries between TADs and do not compromise TADs [11].

**TAD restructuring in disease**

In contrast, CTCF/Cohesin-associated DSBs occurring within TADs could also be the molecular substrate for human pathologies [27]. Genomic rearrangements affecting the TAD structure reshuffle RLs, leading to ectopic interactions that expose CREs to new target genes and consequently causing instant gene regulation changes (reviewed in [29]) (Figure 1d). For
instance, deletions that overlap with boundary regions are frequently associated with T-cell acute lymphoblastic leukemia’s. Deletions of these boundaries in non-malignant cells induce ectopic interactions of previously insulated genomic regions, which is sufficient for proto-oncogene activation [30]. Similarly, translocations in patients with alveolar rhabdomyosarcoma are associated with the fusion of PAX3 and FOXO1 genes [31]. The translocation also fuses the regulatory landscapes of both genes, generating a novel TAD. This configuration leads to ectopic interaction of FOXO1 regulatory sequences with the PAX3 promoter, which likely results in ectopic expression of PAX3:FOXO1 in cell lineages associated with alveolar rhabdomyosarcomas [32]. Ectopic expression of PAX3 is also associated with limb malformations [33]. The Epha4 TAD comprises a large genomic landscape including a cluster of CREs regulating limb expression of Epha4. The deletion of large parts of the Epha4 TAD and the telomeric TAD boundary lead to ectopic interactions of the Epha4 limb CREs with Pax3 in the neighboring TAD. This results in Pax3 mis-expression and brachydactyly phenotype in mouse mutants. Inversions encompassing the opposite Epha4 TAD boundary and the same set of limb CREs, misplaces them in front of the Wnt6 gene, causing its mis-expression and another limb phenotype [33]. In addition, it has recently been shown that tandem duplications that encompass the RLs of two neighboring TADs and a TAD boundary can restructure TADs by forming new chromatin domains or so called neo-TADs [34••]. The pathogenicity of duplications at the Sox9 locus is determined by the gene and the cis-regulatory content incorporated within the neo-TAD. The duplicated Kcnj2 gene adopts the regulatory activity of Sox9-specific CREs included within the neo-TAD, resulting in ectopic expression of Kcnj2 in a Sox9-like limb expression pattern and a limb malformation characterized by nail aplaisa and shorter digits [34••]. A similar mechanism has also been observed in colorectal cancer, where recurrent tandem duplications create new chromatin domains that drive ectopic IGF2 expression by lineage-specific CREs, which are normally inaccessible to the IGF2 promoter [35].

**TAD restructuring during evolution**

Despite the deleterious effects of CREs and gene rewiring during TAD restructuring in disease, few reports suggest that such alterations have contributed to gene expression changes important for evolutionary novelties. In vertebrate HoxA and HoxD clusters, the 5’ and 3’ genes are located in different TADs [1,36–38]. The partitioning of the clusters by a TAD boundary contributes to keep the regulatory information acting on 5’ and 3’ Hox genes largely separated. This separation is critical for proper development of different structures and organs, among them the vertebrate appendages (reviewed in [39]). In contrast, the single ancestral Hox cluster in the cephalochordate amphioxus is not partitioned by a TAD boundary [40••]. Instead, a single TAD contains the whole cluster and the 3’ flanking neighborhood that harbors several long-
distance CREs. Accordingly, the synteny between vertebrate and amphioxus around the cluster is conserved in the 3’ region but not in the 5’ genomic territory. Therefore, it is likely that a genomic reorganization at the 5’ region of the Hox cluster at the base of the vertebrates was important for the evolution of critical 5’ CREs essential for vertebrate novelties. Most likely, the appearance of these CREs required the generation of a new TAD to separate the regulatory information associated with 3’ and 5’ genes [40–]. Similarly, the genomic region around the Sonic hedgehog (Shh) gene emerged by a chromosome rearrangement at the base of the vertebrate lineage, which allowed the recruitment and incorporation of the Lmbr1 gene to the Shh TAD [3,41]. The intron 5 of Lmbr1 became a substrate for the emergence of the Shh long-distance ZRS enhancer, a critical CRE for unpaired and paired fin evolution [42]. Interestingly, a recent study investigating hindbrain boundary formation in different fish species provides a more detailed picture about the molecular events leading to novel morphological structures [43–]. In zebrafish, the rhombomeric boundaries are characterized by actomyosin cables that refine hindbrain segmentation, an adaptive trait specific to the Ostariophysi lineage. Their formation requires the activity of the rac3b/rfng/sgca regulatory cluster that emerged by a lineage-specific chromosomal rearrangement, fusing the ancestral RLs of rac3b/rfng and sgca. The rac3b/rfng/sgca genes in this new RL share the same chromatin interaction domain and several CREs with redundant activities that control the hindbrain boundary specific expression. However, only one CRE located in the ancestral block of sgca harbors a conserved sequence that predates the chromosomal rearrangement in the Ostariophysi lineage, suggesting that ectopic interactions and adoption of hindbrain boundary expression in the rac3b/rfng/sgca regulatory cluster accompanied the rearrangement. This pioneering regulatory interaction and novel gene expression domain of the rac3b/rfng/sgca genes likely facilitated the emergence of additional and redundant CREs within the Ostariophysi lineage, providing further robustness in gene expression [43–].

Intra TAD connections fine-tune gene expression during differentiation and evolution
In contrast to the conserved and invariant nature of TAD boundaries during evolution and throughout development, intra-TAD chromatin interactions undergo extensive changes during cell differentiation. In addition to the role of CTCF at TAD boundaries, CTCF binding patterns show a high degree of cell variability. Genome-wide CTCF occupancy patterns have revealed approximately 110000 CTCF-binding sites across 40 human cell types with 80% of them being cell-type specific [44]. These variable sites have been shown to be engaged in architectural loops within TADs and to mediate interactions between promoters and active CREs [23,45, see also 46]. High resolution Hi-C in human cell lines has shown that TADs are further sub-divided into nested structures consisting of smaller loop domains [23]. However, and in contrast to larger TADs, these sub-structures are more variable across cell types and possibly contribute to
tissue-specific gene expression [8]. For instance, Hi-C analyses across several human cell lines and primary tissues have identified frequently interacting regions (FIREs) that are restricted within smaller sub structures. The local changes are associated with active CREs near genes with tissue specific functions [47]. These observation and previous studies suggest that CRE-promoter interactions can be further restrained in a more cell-type-specific organization within TADs [47–49]. Cell-type-specific interactions involve transcription factors bound at CREs and protein complexes, such as Mediator and Cohesin, which physically and functionally connect CREs with the transcriptional machinery of promoters [49,50]. Additionally, the architectural protein Yin Yang1 (YY1) has also been recently described to mediate specific CRE-promoter interactions within TADs [51,52]. However, the dynamics and hierarchical relationship of tissue-specific and CTCF-mediated loop interactions remain poorly understood. Recently evolved CTCF binding sites in primary liver cells of mouse, dog or macaque can mediate local chromatin changes within TADs that are frequently associated with differential gene expression [11]. It is therefore appealing to speculate that changes in intra-TAD interactions, such as CTCF-mediated looping, contribute to gene expression changes and consequently to inter-species differences.

Concluding remarks

Despite the current more realistic picture about topological organization of RLs and how alterations of TAD structure can modify gene regulation in disease and during evolution, we still lack precise knowledge of how CREs operate within the structural context of TADs. A detailed understanding of the mechanisms that facilitate functional interactions within the TAD scaffold and how they influence each other will undoubtedly shed light on how gene expression is controlled during differentiation and development. Deciphering this functional relationship will be important from an evolutionary perspective as changes within CREs or their scaffolds could result in gene expression changes. Future research will benefit from technical improvements in identifying CREs and capturing chromatin conformations in high resolution and in small cell populations from broader taxonomic samplings.

Acknowledgements

We apologize to those researchers whose work could not be discussed as well as to all the authors whose original work could not be cited due to space limitations. We thank I. Maeso, J. J. Tena and R. D. Acemel for fruitful discussions. This work was supported by the Spanish government [grants BFU2016-74961-P and BFU2016-81887-REDT], the Andalusian Government [grant BIO-396] and the European Union’s Horizon 2020 research and innovation programmes European Research Council (ERC) [grant agreement # 740041] to J.L.G.S. and the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement [#800396] to M.F.
References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

· of special interest
· of outstanding interest


[*] The authors investigate the role of the Shh TAD in mediating the long-distance regulatory relationship of the ZRS enhancer that controls Shh limb expression. Using numerous reporter gene insertions they show that ZRS acts pervasively within the TAD. A series of engineered intra-TAD deletions and duplications, reducing or increasing the distance, does not affect Shh expression. In contrast, TAD disruption prevents regulatory contacts in a distance-dependent manner. This study demonstrates that the Shh TAD structure buffers against long genomic distances.


[•••] This study investigates the relationship of genomic regulatory blocks (GRB) with the TAD organization from human and *Drosophila*. GRB boundaries are highly predictive for TAD
boundaries, revealing a conserved regulatory architecture. In vertebrates, GRB-TADs exhibit a stronger insulation and significant enrichment for constitutively bound CTCF sites than non-GRB-TADs, suggesting chromatin topology as evolutionarily conserved feature to maintain long-range *cis*-regulatory relationships.


[*] Using the auxin-inducible degron system, the authors remove CTCF in mouse embryonic stem cells and show that CTCF is essential for looping between CTCF binding sites and insulation of TADs, whereas active and inactive genome compartments are not affected.


[*] Deletion of the Cohesin-loading factor Nipbl in mouse liver leads to a loss of TAD structure
and loop interactions, while active and inactive compartmental segregation is not affected.


[••] This study shows for the first time that TAD formation by loop extrusion is accompanied by topological stress relief and genome instability. Topoisomerase II beta (TOP2B) induces double strand breaks (DSBs) at evolutionary conserved loop anchors bound by CTCF and Cohesin. Importantly, DSBs are enriched on both sides of topological domain borders, which largely preserves TAD integrity.


34. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, Kraft K,

[•] The authors show by capture Hi-C and 4C-seq in patient cells and genetically modified mice that genomic duplications can result in the formation of new chromatin domains (neo-TADs), when duplicating a TAD boundary. Neo-TADs are functionally and spatially separated from their genomic environment and thereby determining the duplications’ pathogenesis. Mouse models of a duplication including *Kcnj2* show neo-TAD formation that allow ectopic activation of *Kcnj2* in *Sox9* expression domains, causing limb malformation resembling the Cooks syndrome.


37. Berlivet S, Paquette D, Dumouchel A, Langlais D, Dostie J, Kmita M: **Clustering of Tissue-Specific Sub-TADs Accompanies the Regulation of HoxA Genes in Developing Limbs.** *PLoS Genet* 2013, **9**:e1004018.


39. Lonfat N, Duboule D: **Structure, function and evolution of topologically associating domains (TADs) at HOX loci.** *FEBS Lett* 2015, **589**:2869–2876.


[•] Using 4C-seq and 3D chromatin modeling, the authors profile the chromatin organization of the Hox cluster in the invertebrate amphioxus. In contrast to the bipartite TAD structure at the HoxA and HoxD cluster of jawed vertebrates, the amphioxus Hox cluster is organized into a single chromatin domain. They propose a stepwise evolution of the Hox bipartite regulatory system in the chordate ancestor by the incorporation of a new TAD neighboring posterior Hox
This study provides a causal link between a genomic rearrangement, the evolution of gene expression patterns and morphological novelties. The rac3b/rfng/sgca cluster is organized in a novel chromatin domain that emerged by a rearrangement in the Ostariophysi lineage. The rearrangement was accompanied by a gain of long-range CREs interaction driving rhombomeric boundary expression. Subsequently, redundant CREs evolved, providing robustness for the segregation of neuronal progenitors in the zebrafish hindbrain.


[·] This study, together with [52], describes the transcription factor YY1 as an important mediator of cell-type specific CRE-promoter interactions. The authors demonstrate that YY1 enhances DNA interactions in vitro and that deletion of YY1 binding sites or depletion of YY1 protein disrupts loop interactions and gene expression.


[·•] This study, together with [51], describes the transcription factor YY1 as an important mediator of cell-type specific CRE-promoter interactions. The authors investigate how CTCF and YY1 orchestrate chromatin looping during early neural lineage commitment. CTCF occupancy and pluripotency interactions are predominantly lost in the transition from ES-cells to neural progenitor cells (NPCs). In contrast they observe a gain of NPC-specific interactions mediated by YY1, connecting NPC-specific genes and CREs. These interactions are nested within larger interactions anchored by constitutively bound CTCF sites.
Figure 1: **Genome organization into TADs during evolution and TAD restructuring in disease.**

a) Genes (black rectangles) and corresponding CREs (black ovals) of a given RL are organized into TADs that are characterized by high internal chromatin interactions. CTCF-associated TAD boundaries are characteristic signatures in vertebrates.

b) TAD formation by the loop extrusion model and associated DSBs (schematic adapted from [22,26,27]).
1. Ring-shaped Cohesin complex (orange circle) extrudes chromatin forming loops.
2. Genes and CREs are brought into proximity, which can mediate gene expression (blue arrow).
3. Chromatin extrusion causes torsional stress ahead of Cohesin and is transiently stabilized at CTCF-bound sites in convergent orientation (red and blue triangles).
4. Topoisomerase II beta (yellow circle) can induce DSBs to relieve torsional stress outside of loop anchors (TAD boundaries) and on both strands.

c) DSBs at TAD boundaries (dotted vertical lines) preserve TAD integrity and allow modular TAD reshuffling, which is typically observed in genome reorganizations in evolution.

d) Chromosomal rearrangements and structural variations (e.g. inversions) that restructure RLs and chromatin interactions frequently cause gene mis-expression and diseases.