Genome folding in evolution and disease

Jonas Ibn-Salem

February 14, 2018
Version: 0.0.1

Johannes Gutenberg University Mainz

Faculty of Biology
Institute of Organismic and Molecular Evolution
Computational Biology and Data Mining Group

PhD Thesis

Genome folding in evolution and disease

Jonas Ibn-Salem

1. Reviewer Prof. Dr. Miguel Andrade-Navarro
Faculty of Biology
Johannes Gutenberg University Mainz

2. Reviewer Jun.-Prof. Dr. Susanne Gerber
Faculty of Biology and Center for Computational Sciences in Mainz
Johannes Gutenberg University Mainz

Supervisor Miguel Andrade

February 14, 2018

Jonas Ibn-Salem
Genome folding in evolution and disease
PhD Thesis, February 14, 2018
Reviewers: Prof. Dr. Miguel Andrade-Navarro and Jun.-Prof. Dr. Susanne Gerber
Supervisors: Miguel Andrade and

Johannes Gutenberg University Mainz
Computational Biology and Data Mining Group
Institute of Organismic and Molecular Evolution
Faculty of Biology
Ackermannweg 4
55128 and Mainz

Contents

Abstract

The DNA of the human genome has a total length of 2 meters and fits into a 10 nm nucleus. Hi-C experiments probe the three-dimensional genome architecture by genome-wide DNA-DNA contact maps. In these maps, interphase chromosomes fold locally into regions of several hundred kb, called topologically associating domains (TADs). TADs can restrict the interactions of genes with distal regulatory sequences, such as enhancers, and might, therefore, represent important functional units of genomes.

In this thesis, I integrate genome-wide chromatin interaction maps with diverse genomic datasets to analyze the function of TADs for gene regulation during evolution and in disease genomes. More specifically, I show that duplicated gene pairs are enriched for colocalization in the same TAD, share more often common enhancer elements than expected and have increased contact frequencies over large genomic distances. The clustering of functionally related genes within TADs enables concerted gene expression and indicates evolutionary constraints in functional genome organization. Indeed, by analyzing evolutionary rearrangements between human and diverse mammal species, I show that TADs are highly conserved building blocks of genomes across millions of years of evolution. While TADs seems to conserve gene expression profiles between mouse and human, genes in rearranged TADs show divergent expression. Next, I analyze disruption of TADs in whole-genome sequenced subjects with diverse neurodevelopmental phenotypes. While the phenotypes cannot be explained by variants in protein-coding sequences, balanced chromosomal rearrangements disrupt TADs and promoter-enhancer interactions of phenotypically matching genes. This effect mechanism leads to altered gene expression in patient-derived cell lines. Finally, I present a computational tool, 7C, to predict long-range chromatin interactions by largely available ChIP-seq data. This enables the analysis of three-dimensional genome folding in many diverse tissues and conditions.

Together, these results show that TAD disruption is associated with altered gene regulation during evolution and in diseases. TADs are not only structural units of genomes but also important functional building blocks and represent regulatory environments for genes. Therefore, it will be increasingly important to take the

three-dimensional genome structure into account, not only in genomic research but also in clinical practice.

1 Introduction

Regulation of gene expression

Each cell in our body originate from the same fertilized stem cell and has therefore virtually the same genome. However, different cell types have distinct morphologies and fulfill diverse functions. This diversity is archived by expressing only a subset of genes to a specific extent for any cell type, developmental state, and environmental condition. Gene expression is therefore complex and controlled on many molecular levels (Lelli et al., 2012).

The initial sequencing of the human genome reviled a large resource of information on the genetic sequence, however, we are still far from complete understanding of the sequence itself. While functional knowledge of individual genes and its activity, evolution and associations to diseases accumulates over the last years, the non-coding parts of the genome is only very recently annotated in large collaborative efforts (Dunham et al., 2012; Roadmap Epigenomics Consortium et al., 2015; Andersson et al., 2014). These efforts lead to an increased understanding of the regulatory potential of non-coding regions and its dynamics across conditions.

The genomic DNA sequence itself encodes cis-regulatory modules (CRMs) to which transcription factors (TFs) bind by recognizing specific DNA sequence motifs. TFs often form complexes with other proteins and DNA. However, TF binding and CRM assembly require often specific epigenetic states of chromatin. Epigenetic modifications of DNA, such as methylation, influence the ability of TF to bind DNA. The chromatin structure and accessibility itself also determines if a gene can be transcribed. While so-called pioneering factors can bind closed chromatin that is wrapped around nucleosomes and remodel it to make it accessible for other TFs that require open chromatin and specific post-translational histone modifications to bind cis-regulatory regions and activate target gene expression.

Another layer in gene regulation is the three-dimensional folding structure of chromatin in the nucleus. However, most of the cell-type specific gene regulation that accounts for cell differentiation in development and morphological diversification in evolution, are driven by activation changes of enhancers (Long et al., 2016).

However, the genome is still largely considered as - From 1D to 3D

Enhancers

Enhancers were originally defined as genomic regions that enhance the expression of a reporter gene, when placed experimentally in front of a minimal promoter. (Banerji et al., 1981; Shlyueva et al., 2014). Enhancer activity can also be detected genome-wide by specific patterns of open chromatin using DNase-seq (Song and Crawford, 2010), ATAC-seq (Buenrostro et al., 2013) or posttranslational modification of histones, such as H3K27ac by ChIP-seq (Creyghton et al., 2010). Complex regulation of developmental genes is often archived by additive effects of multiple enhancers. For example, the α-globin gene locus is controlled by multiple enhancers, whereby each enhancer elements act independently and in an additive fashion without evidence of synergistic or higher-order effects (Hay et al., 2016). Also, the Indian hedgehog (Ihh) locus is regulated by multiple enhancers with individual combinations of tissue specificities that function in an additive manner (Will et al., 2017). Experimental variation for of the copy number of enhancers is associated with expression strength. Significantly reduction of the expression of the oncogene *PIMI* could not be archived by perturbing a single enhancer, but only by combinatorial repression of several weak enhancers (Xie et al., 2017). Enhancers are reviewed in more detail in Spitz and Furlong (2012), Andrey and Mundlos (2017), and Long et al. (2016).

Methods to probe the 3D chromatin archtiecture

Microscopy-based techniques to visuallize the genome in 3D

Historically, the organization of chromosomes and specific loci within the nucleus have mostly been studied using fluorescent *in situ* hybridization (FISH) experiments. FISH is limited to examine a few pre-defined loci a few hundred cells at once and is limited in spatial resolution. Novel super-resolution microscopy approaches such as STORM and PALM have enabled direct visualization of the fine-scale structures of the genome at an unprecedented resolution (Bonev and Cavalli, 2016). Labeling of specific chromatin proteins, histone marks, or genomic loci allow to analyze the dynamics of chromosomes at high resolution in living cells. However, despite spectacular technical progress, microscopy-based approaches are limited to a small number of genetic loci and do not allow a comprehensive analysis of the nuclear architecture of the complete genome. Furthermore, the specific folding patterns observed in microscopy cannot be mapped to genomic coordinates, substantially limiting the integration with other genomic data. However, a future combination of imaging-based techniques with proximity-ligation experiments together with integrative computational models might enable to study the real-time dynamics of chromatin organization with high resolution on the single cell level (Stevens et al., 2017, Flyamer et al. (2017)).

Proximity-ligation based method to quantify chromatin interactions

The frequency of interactions between different loci in the genome can be measured experimentally by proximity ligation techniques (Sati and Cavalli, 2017; Schmitt et al., 2016). These protocols are variations of the chromosome conformation capture (3C) experiment (Dekker et al., 2002). 3C works by the ability to cross-link two genomic loci that are in close physical proximity in the nuclear space by treating cells with formaldehyde leading to covalent bonds between proteins and DNA with proteins (Hoffman et al., 2015). The cross-linked chromatin is than digested by enzymes to fragment the genomic DNA. Than, re-ligation of fragmented DNA result in hybrid DNA molecules of restriction fragments that where in close physical proximity during cross-linking but originate from different regions in the linear genome sequence (Dekker et al., 2013; Andrey and Mundlos, 2017) (Fig. 1.1A).

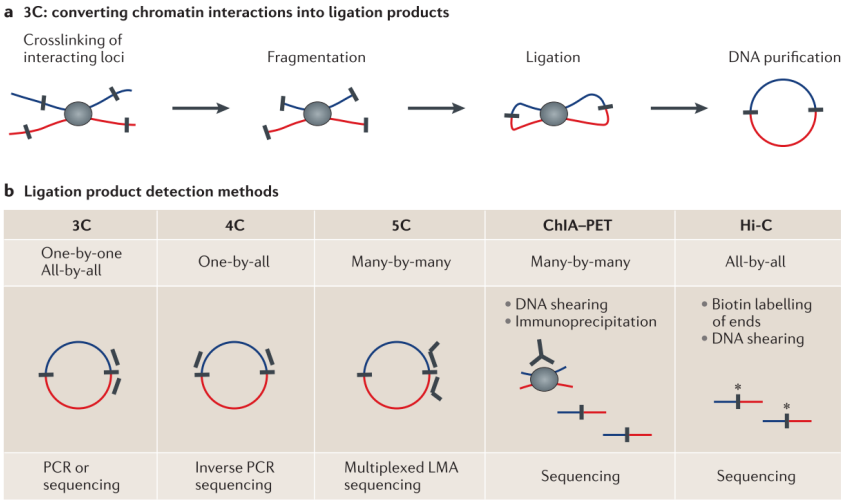

Figure 1.1.: **Proximity ligation technologies to measure chromatin interactions (A)** By treating cells with formaldehyde chromatin is crosslinked. After fragmentation with restriction enzymes, DNA from two loci in close physical proximity in the nucleus is ligated to a hybrid DNA molecules that is than made from DNA that originated from two regions distal in the linear genome (indicated in red and blue). **(B)** Different variants of the 3C experiments differ in their approaches to measure the ligation products or subsets of it in order to quantify chromatin interactions. Figure adapted from (Dekker et al., 2013).

There exist several 3C-based methods which differ by the way the ligation product, which represents and chromatin interaction, is measured and quantified (Fig. 1.1B). The classic 3C protocol allows to quantify hybrid DNA-product by quantitative PCR using specific primers to amplify the product junction (Dekker et al., 2002). In Circular chromosome conformation capture (4C) experiments, a circular PCR is used to amplify all hybrid DNA products that are ligated with a desired restriction fragment, e.g. a specific viewpoint of interest. These products are then sequences to generate an interaction profile measuring all interacting regions with

this viewpoint (Simonis et al., 2006; Noordermeer et al., 2011). Another variant of 3C, Carbon copy chromosome conformation capture (5C), combines 3C with hybrid capture approaches to identify up to millions of interactions in parallel between two large sets of loci, for example between a set of promoters and a set of distal regulatory elements (Dostie et al., 2006; Sanyal et al., 2012). Other methods combine chromatin immunoprecipitation to enrich for chromatin interactions between loci bound by specific proteins of interest or marked by post-translational histone modifications. One of these methods is chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), which allows for genome-wide analysis of long-range interactions between sites bound by a protein of interest (Fullwood et al., 2009). Therefore, ChIA-PET data represent a selected subset of all interactions, but is an efficient alternative to measure interactions at high resolution (Tang et al., 2015). The most unbiased method to quantify all pair-wise interactions genome-wide is Hi-C (Lieberman-Aiden et al., 2009). After the initial restriction enzyme step of 3C, in Hi-C, the ends are filled with a biotin-marked nucleotide and subsequently re-ligated. A streptavidin pull-down step is used to enrich for the chimeric products, which are then sequenced using paired-end sequencing technology. Each read from the resulting read-pairs is then aligned independently to the reference genome to identify the originating position of the sequenced restriction fragment. Each read pair represent a pairwise physical interaction of the corresponding regions. Interaction frequencies are usually analyzed by binning the genome into equal sized regions of several kb depending on sequencing depth. While the first Hi-C study produced genome-wide interactions at 1Mb resolution (Lieberman-Aiden et al., 2009), more recent studies could analyse folding patterns at 40kb (Dixon et al., 2012), and later up to 1kb resolution (Rao et al., 2014).

Hierarchy of chromatin 3D structure

The three-dimensonal organization of genome folding was studied extensively in recent years and is reviewd comprehensively by leading expertes in the field (Pombo and Dillon, 2015; Sexton and Cavalli, 2015; Bouwman and de Laat, 2015; Dekker and Mirny, 2016; Dixon et al., 2016; Schmitt et al., 2016; Bonev and Cavalli, 2016; Hnisz et al., 2016; Merkenschlager and Nora, 2016; Long et al., 2016; Ruiz-Velasco and Zaugg, 2017; Andrey and Mundlos, 2017).

Chromosomal territories and inter-chromosomal contacts

The eukaryotic genome is highly organized in the interphase nucleus. Chromosomes occupy distinct spatial regions, called chromosome territories, and intermingle less than one would expect by chance (Cremer and Cremer, 2001). This was first observed using imaging based approaches, and is reflected in Hi-C interaction maps, where inter-chromosomal contacts occur an order of magnitudes less frequent than

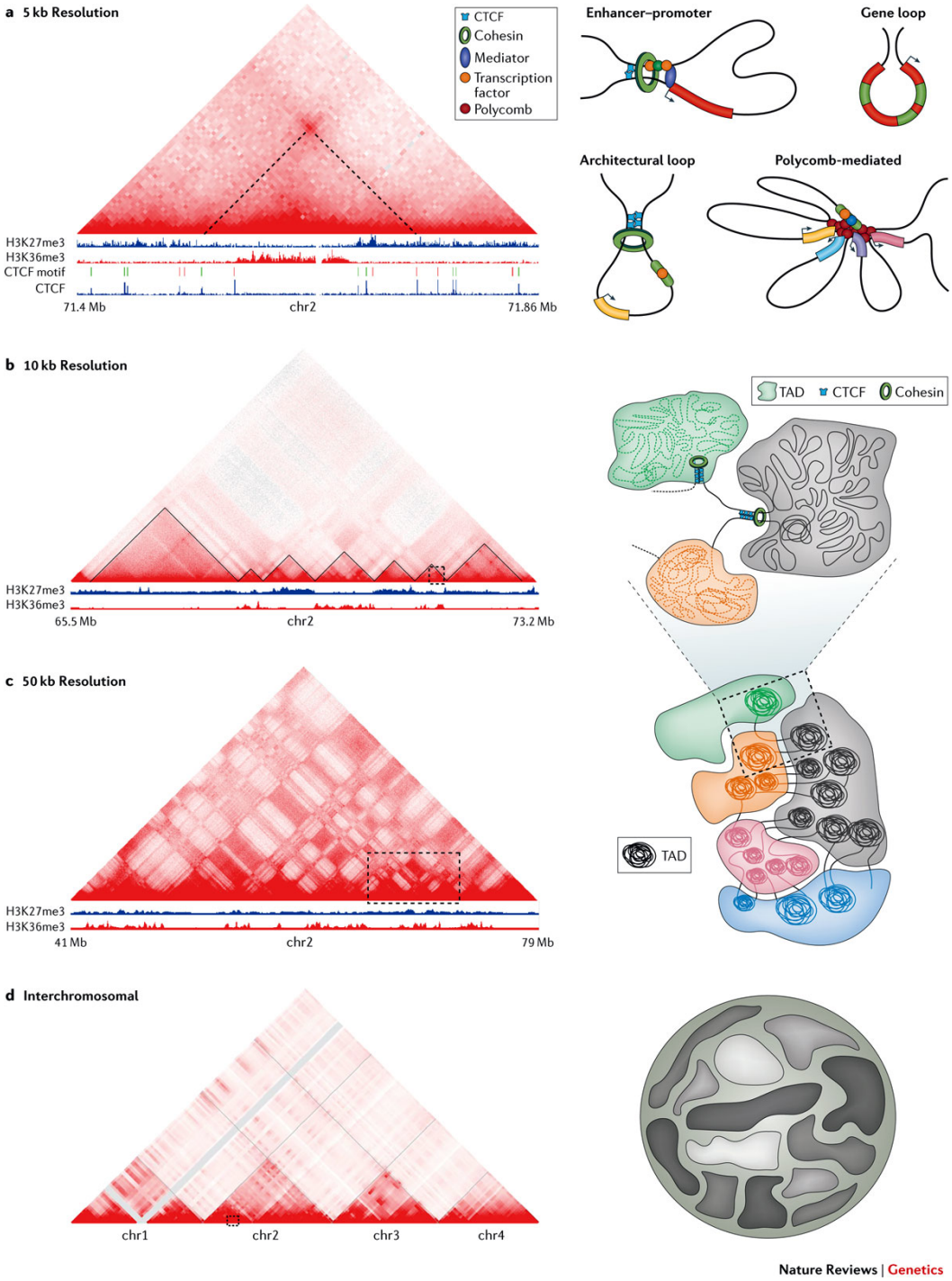

Figure 1.2.: **Hierarchical organization of chromatin three-dimensional chromatin architecture (A)** Figure source (Bonev and Cavalli, 2016) https://www.nature.com/articles/nrg.2016.112/figures/2.

intra-chromosomal contacts (Lieberman-Aiden et al., 2009). However, despite this spatial segregation of chromosome, intermingling of chromosome occurs and is associated with chromosomal translocations (Branco and Pombo, 2006; Roukos et al., 2013; Roukos and Misteli, 2014). There are also specific gene regulatory interactions between different chromosomes (de Laat and Grosveld, 2007), for example olfactory receptor genes cluster densely in the nucleus of olfactory neurons to facilitate mono-allelic expression of a single receptor gene per cell (Monahan and Lomvardas, 2015). Togther, there exists only a few but specific inter-chromosomal contacts and the genome is non-randomly organized in the nucleus by chromosomes occupying distinct spatial territories.

A/B compartments

The ability to measure genome-wide chromatin contacts using Hi-C, revealed that individual regions on chromosomes segregate by preferential interactions into two mayor clusters, referred to as A/B-compartments (Lieberman-Aiden et al., 2009). Interestingly, regions in A-compartments are associated with active histone-modifications and active transcription, whereas B-compartment is associated with heterochromatin, lamina association, and repressed genes (Bonev and Cavalli, 2016). More recnetly, higher resolution Hi-C maps further subdevided A/B compartments into six sub-compartments with prefferntial interacitons and associations to distinct chromatin features (Rao et al., 2014).

Topologically associating domains (TADs)

Compartments could be identified by clustering of long-range interactions in Hi-C maps with bin resolution of 1 Mb. In 2012 higher resolution Hi-C maps of up to 40 kb lead to the identification of genomic regions with preferential interactions within them. These genomic regions were termed topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). They are operationally defined as genomic regions with frequent interactions of loci within the domain and decreased interactions across domain boundaries. Quantitatively, Hi-C contact frequencies between loci in the same TAD are approximately two to three times those of genomic regions outside of the TAD (Merkenschlager and Nora, 2016).

TADs can be identified from Hi-C interaction maps computationally by different algorithms (Ay and Noble, 2015). The directionality index is a score for each bin in the Hi-C matrix that quantifies the number of upstream versus downstream interactions of this bin. Using hidden Markov models TAD boundaries where then identified in regions where DI is changing drastically (Dixon et al., 2012). Other algorithms compute an insulation score as the extent to which interactions cross potential TAD boundaries (Crane et al., 2015). Later the Arrowhead algorithm was in-

troduced to find "contact domains" as smaller nested structures along the diagonal of high resolution Hi-C matrices (Rao et al., 2014). Furthermore, when analyzing Hi-C interactions at different length scales, hierarchies of TADs and sub-TADs could be identified that overlap each other (Filippova et al., 2014; Fraser et al., 2015). The different algorithms and parameters used in each study is only one source of variation in reliably identifying TADs. Also, the resolution of Hi-C maps, which is mainly defined by sequencing depths but also the Hi-C protocol itself (Rao et al., 2014), as well as different normalization strategies for Hi-C contacts introduce variability (Dali and Blanchette, 2016; Forcato et al., 2017). Therefore, the number and size of TADs varies between different studies, making it difficult to compare TADs in different conditions and species across studies.

The first studies on TADs identified around 3,000 TADs with a median size of ~800 kb in human and mouse genomes (Dixon et al., 2012) and arround 1,200 physical domains of arround 100 kb in *Drosophila* genomes (Sexton et al., 2012). Analysis of 1kb or 5kb resolution Hi-C matrices resulted in nested "contact domains" with a median size of 185 kb (range 40 kb - 3 Mb) in human and mouse cells (Rao et al., 2014).

Interestingly, TADs might be equivalent to "chromatin domains" of 10 kb - 1 Mb in size detected by microscopy approaches (Cremer and Cremer, 2010; Gibcus and Dekker, 2013). Another connection of Hi-C derived interaction maps with previous microscopy observations, is that TADs in *Drosophila* correspond to bands of polytene chromosomes (Eagen et al., 2015).

The spatial positioning of TADs correlate with many genomic features measured along the linear genome (Merkenschlager and Nora, 2016). TAD boundaries are enriched for binding of "insulator proteins", such as CTCF in mammals and CP190 in Drosophila (Dixon et al., 2012; Sexton et al., 2012). Furthermore, TAD boundaries are associated with active chromatin, such as H3K4me3 and H3K27me3, DNase I hypersensitivity, active transcription, short interspersed nuclear elements (SINEs), and house-keeping genes (Dixon et al., 2012). TADs correspond to regions of early and late replication timing (Pope et al., 2014; Dileep et al., 2015) and lamina associated domains (LADs) (Dixon et al., 2012). Importantly, enhancer-promoter interactions seems to be mostly constrained within TADs (Shen et al., 2012; Ghavi-Helm et al., 2014; Symmons et al., 2014)

Altogether, there is accumulating evidence that TADs are fundamental units of chromosome organization (Dixon et al., 2016)

Chromatin loops

Recent studies provide functional insights about how chromatin loops are formed and highlight the role of architectural proteins such as CTCF and cohesin (Merken-

schlager and Nora, 2016). CTCF recognizes a specific sequence motif, to which it binds with high affinity (Kim et al., 2007; Nagy et al., 2016). Interestingly, CTCF motifs are present in convergent orientation at chromatin loop anchors (Rao et al., 2014; Tang et al., 2015; Vietri Rudan et al., 2015). Furthermore, experimental inversion of the motif results in changes of loop formation and altered gene expression (Guo et al., 2015; de Wit et al., 2015). Polymer simulations and experimental perturbations led to a model of loop extrusion, in which loop-extruding factors, such as cohesin, form progressively larger loops but stall at CTCF binding sites in convergent orientation (Sanborn et al., 2015; Fudenberg et al., 2016). According to these models, CTCF binding sites can function as anchors of chromatin loops.

In summary, these findings suggest a hierarchical organization of chromosome architecture. First, dynamic nucleosome contacts from clutches and fibers. These engage in long-range chromatin loops, some of which are stabilized by architectural proteins, such as CTCF and cohesin, and lead to the formation of TADs. TADs form cluster by their epigenomic type into A/B compartments and coalescence of compartments in the same chromosome forms chromosome territories (Bonev and Cavalli, 2016).

Molecular mechanisms of TAD and loop formation

Several initial studies analyzed the role of CTCF and cohesin in TAD formation. In two studies depletion of cohesin reduced interactions within TADs but did not alter TAD strucures completley (Seitan et al., 2013; Zuin et al., 2014). One study report a significant increase in interactions between differnt TADs after cohesin depletion (Sofueva et al., 2013). Depletio of CTCF had a similar effect with increased inter-domains interactions (Seitan et al., 2013; Zuin et al., 2014). These result suggest a functinoal role of cohesin and CTCF in promoter-enhancer interactions and TAD formation (Pombo and Dillon, 2015).

Dynamics of chromatin structure

Proximity ligation experiments like Hi-C measure contact frequencies as average over millions of cells used in the experiments. Therefore, identified contacts might not be present in each individual cells and can be can be dynamic over short timescales in each individual cell. It is important to keep in mind how chromosome folding change during cell division. Spatial organization is generally studied in non-synchronous cells, of which interphase cells make up the biggest proportion (Bouwman and de Laat, 2015). In interphase, chromosomes are decondenced and hierarchically organized into the territories, compartments, and TADs as described above. To prepare for cell division chromosomes untangle and condense, while transcription ceases almost entirely. Mitotic chromosomes do not show preferential organization, such as compartments or TADs (Naumova et al., 2013). Enhancer-

promoter looping might be absent as well (Dekker, 2014). After cell division, chromosomes decondense and fold into the interphase hierarchy. While individual genes are relatively mobile during early G1 phase, they become quickly constrained to a small nuclear sub-volume, after which genome folding is relatively stable for the rest of the interphase (Chubb et al., 2002; Walter et al., 2003). These dynamics during cell-cycle raise the question of how the pattern of 3D organization is re-established with each cell division.

Variation of chromatin structure across cell types

The primary domain architecture of chromatin is largely preserved in different cell types and even across species (Dixon et al., 2012; Rao et al., 2014). However, chromatin dynamics specify distinct gene expression programs and biological functions (Bonev and Cavalli, 2016). One example of dynamic chromatin organization is dosage compensation, in which the X chromosome is transcriptionally inactivated in human female cells. Whereas normal TAD structures were observed on the active X chromosome, only two very large domains were identified on the inactive X chromosome in *Drosophila* and human (Deng et al., 2015; Rao et al., 2014). Other examples include differences in terminally differentiated post-mitotic cells. For example, rod photo-receptor cells in nocturnal mammals have an unusual, inverted nuclear architecture, in which heterochromatin is enriched in the center of the nucleus and is absent from the periphery (Solovei et al., 2013). Further biological processes related to cell cycle exit strongly affect chromatin three-dimensional organization. These are quiescence in yeast, where intrachromosomal contacts increase (Rutledge et al., 2015), and senescence, where heterochromatin relocalizes from the nuclear periphery to the interior (Chandra et al., 2015). However, A/B compartments and TADs seems to be largley unaffected in these processes (Criscione et al., 2016).

Subtler effects of changes in chromatin reorganization are observed during biological processes such as cell differentiation and signaling (Bonev and Cavalli, 2016). During the transition of embryonal stem cells (ESC) from ground-state of pluripotency to a primed state for differentiation, a gradual and reversible establishment of long-range contacts was observed between bivalent gene promoters (Joshi et al., 2015). These changes depended on Polycomb repressive complex 2, underscoring its role in establishing 3D genome organization in early development, as was previously shown for *Drosophila* (Bantignies et al., 2011). To address the question, how nuclear architecture change during lineage specification, a recent Hi-C study produced Hi-C interaction maps in ESCs and four ESC-derived lineages representing early developmental stages (Dixon et al., 2015). Interestingly, TADs are mostly unchanged during lineage specification, but intra-TAD interactions in some domains were strongly altered and these changes correlate with active chromatin state (Dixon et al., 2015). Furthermore, often entire TADs relocate from one compartment to another, which was also associated with transcriptional changes of

genes in these TADs. Also in B cell differentiation, several regions change compartment identity and relocate from the nuclear periphery to the interior (Lin et al., 2012).

The dynamics of chromatin architecture were also studied in response to stimuli, such as hormone signaling by progestin or estradiol. Despite large changes in the transcriptional activity, only small changes were observed in the domain organization of chromatin (Le Dily et al., 2014). However, often the entire TAD responded as a unit by changing histone modifications and switching between A and B compartment. This suggests, that transcription status is coordinated within TADs.

- Add percentage of shared boundaries between cell types (Dixon et al., 2012; Rao et al., 2014).
- Summary of dynamics

Evolution of chromatin organization

While TADs were initially described for mammalian genomes, a similar domain organization was found in the genomes of non-mammalian species such as *Drosophila* (Sexton et al., 2012), zebrafish (Gómez-Marín et al., 2015) *Caenorhabditis elegans* (Crane et al., 2015) and yeast (Hsieh et al., 2015; Mizuguchi et al., 2014).

However, in *C. elegans* TADs were only observed on the X chromosome of XX hermaphrodites and not on autosomes (Crane et al., 2015). The small and compact genome with most of the cis-regulatory information within 10kb from the TSS might not need long-range domains for gene regulation (Long et al., 2016). In contrast, plants have long-range cell-type specific enhancers (Zhu et al., 2015), but TAD like structures could not be observed in Hi-C experiments in *A. thaliana*. Interestingly, both *C.elegans* and *A.thaliana* do not encode a CTCF homolog (Heger et al., 2012), suggesting alternative mechanisms of genome organization and segmentation in these species (Long et al., 2016). TAD-like structures are therefore not required for eukaryotic interphase chromosome folding. Nevertheless, Hi-C experiments in bacteria and yeast suggest that self-interacting domains may be an ancient feature of chromosome organization. Hi-C data in *Caulobacter* cells revealed so-called chromosomally interacting domains (CIDs) of 30 to 420 kb in size (Le et al., 2013). In fission yeast *S. pombe*, globule structures at the 40-100 kb scale were identified and depend on cohesin complex (Mizuguchi et al., 2014). Furthermore, in very short CIDs of around 2-10 kb were detected in *S.cerevisiae* and bounded by highly transcribed genes (Hsieh et al., 2015). Further high-resolution experiments in more diverse species are needed to understand the evolutionary origin of genome segregation into TAD-like structures.

The presence of domain-like structures in diverse species across the tree of life leads to the question if not only the genomic sequence but also its folding structure

is conserved between species. Interestingly, TADs are not only largely stable across different cell-types (Dixon et al., 2012; Rao et al., 2014) and during differentiation (Dixon et al., 2015), but also remarkably similar between homologous regions in mouse and human (Dixon et al., 2012). More than 54% of TAD boundaries in human hESC cells occur at homologous genomic positions in mouse ESCs (Dixon et al., 2012). Similarly, 45% of contact domains called in mouse B-lymphoblasts were also identified at homologous regions in human lymphoblastoid cells (Rao et al., 2014). A single TAD boundary at the Six gene loci could be traced back in evolution to the origin of deuterostomes (Gómez-Marín et al., 2015). However, these analyses focused only on the subset of syntenic regions that can be mapped uniquely between genomes. Initial comparative Hi-C studies identified several evolutionary breakpoints at TAD boundaries and hihglights a major role of conserved CTCF binding sites in facilitating conservation of TADs (Vietri Rudan et al., 2015). However, it remains to be investigated systematically if TAD regions as a whole might be stable or disrupted by rearrangements during evolution and how this is associated to conserved or divergent transcriptional regulation between species.

Disruption of chromatin architecture in disease

The insulating function of TADs restricting promoter-enhancer interactions (Symmons et al., 2014) suggests that alternations of TAD structures may induce ectopic interactions between regulatory elements and genes in neighboring TADs, leading to gene dysregulation (Fig. #ref(fig:EnhancerAdoption)).

Several experiments of genetic manipulation of specific TAD boundaries can change the surrounding interaction patterns and thus affect the expression of nearby genes. After experimental deletion of a 58-kb region encompassing a TAD boundary in mouse ESC, interactions between adjacent TADs significantly increase and genes in neighboring TADs were upregulated (Nora et al., 2012). More precise deletion or inversion of CTCF binding site at loop anchors in TAD boundaries altered local chromatin architecture and nearby gene expression (Dowen et al., 2014; Guo et al., 2015; Narendra et al., 2015).

The alternation of chromatin contacts upon TAD boundary disruptions leads to the question whether disrutpion of TADs could also be and mechansim of disease pathology. This hypothesis were initilay dirven by studying the etiology of the Liebernberg syndrome, in which an upper-limb malformation phenotype is casued by deletion in the vinicity of *PITX1* gene. *PITX1* becomes therby accessible to an enhancer with specific activity in lower-limb and was missexpressed in a corresponding mouse model (Spielmann et al., 2012). An overlap of thes deletion with later identifed TAD boundaries lead to the hypothesis of an regulatory effect mechanism of boundary deletions referred to as enhancer adoption (Spielmann and Mundlos, 2013).

An association of TAD disrutpions with disease was first shown by computational analysis integrating TAD structures, with tissue specific enhancers, and large crhomosomal deletions associated with clinical phenotypes (Ibn-Salem et al., 2014).

In this study, the Human Phenotype Ontology database (Köhler et al., 2014) was used to relate the phenotypes of 922 deletion cases from the DECIPHER database (Firth et al., 2009) to monogenic diseases associated with genes in or adjacent to the deletions. This was used to differntiate for each deletion case between two possible pathogenic effect mechanism that best explains the phenotypes observed in the patients. Many deletions could be explained by a gene dosage mechansim and haploinsufficiency of genes located within the deletion. However, about 12 percent of cases could be best explained by a TAD boundary disruption and a specific combinations of tissue-specific enhancers and genes in adjacent TADs. Importantly, randomization of phenotype data showed that this enhancer adaption mechanism was significantly more frequent then expected by chance (Ibn-Salem et al., 2014). Therfore, this study shows for the first time an association of TAD disrutpiton with genetic diseases and highglight an regulatory effect mechansims of TAD disruption that need to be further investigated and has to be considered when interpreting the gentic variations in patient genomes.

Intersingly, these initial results are largly confirmed by very recent studies investigating structural variants that disrupt TADs and lead to ectopic gene expression in genetic diseases (Lupiáñez et al., 2015; Franke et al., 2016; Redin et al., 2017) and cancers (Northcott et al., 2014; Hnisz et al., 2016; Weischenfeldt et al., 2016). This is discussed in more detial in secion #ref(thesis-discussion).

Aims of this thesis

The recent advances of genome-wide ligation proximity mapping revealed that interphase chromosomes are highly organized and structurally segregate into TADs. TADs were already shown to associate with diverse genomic functions such as histone modifications (Dixon et al., 2012; Sexton et al., 2012), replication timing (Pope et al., 2014), and gene expression correlation (Nora et al., 2012; Le Dily et al., 2014). We previously showed a significant association of TAD disruptions by large chromosomal deletions with clinical phenotypes, likely caused by an enhancer adaption mechanism (Ibn-Salem et al., 2014; Lupiáñez et al., 2015). However, it is still unclear how exactley the genome folds into TADs and what consequence this has for gene regulation during evolution and in genetic diseases. For example, it is not clear to which extends genes within the same TAD are expressed and regulated in a coordinated manner. Despite initial evidence of evolutionary conservation of TADs in homologous regions between human and mouse (Dixon et al., 2012; Vietri Rudan et al., 2015), there was no systematic analysis of the stability of TADs during evolution and the consequences of TAD disruptions on gene expres-

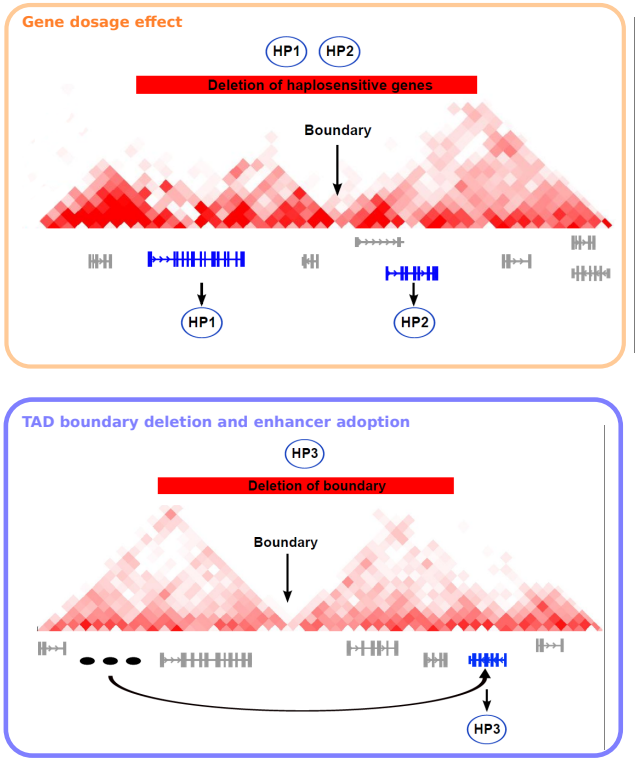

Figure 1.3.: **Models of pathomechanisms by chromosomal deletions involving TAD disruption** In each panel, an exemplary deletion is shown as a red bar, a TDB is indicated with a black arrow, and genes associated with the phenotypes of the CNV patient are shown in blue, other genes in gray. Phenotypic abnormalities are represented as exemplary HPO terms (HP1, HP2 and HP3). Three tissue-specific enhancers are shown in (B) as black ovals. **(A)** Gene-dosage effect. A deletion leads to a reduction in the dosage of haplosensitive genes located within the CNV. The individual with the deletion has two phenotypic abnormalities (HP1, HP2) resulting from deletion of two haplosensitive genes. AMendelian disease related to mutations in the first gene is associated with HP1, and aMendelian disease related to mutations in the second gene is associated with HP2. **(B)** TAD boundary deletion and enhancer adoption. Removal of the topological domain boundary allows the tissue-specific enhancer inappropriately to activate a phenotypically relevant gene located adjacent to the deletion, a phenomenon that we refer to as enhancer adoption. In this case, the individual with the deletion has a phenotypic abnormality (HP3) that is also seen in individuals with a Mendelian disease related to a mutation in the gene adjacent to the deletion. Figure adopted from (Ibn-Salem et al., 2014).

sion divergence between species. The accumulating evidence of the important gene regulatory function of TADs leads to the question if TADs play an important role in genetic diseases (Spielmann and Mundlos, 2016). More specifically, TADs might be used together with other genomic annotations, such as enhancers and their interactions with regulated genes to interpret structural variations in patient genomes. Furthermore, increasing mechanistic understanding of chromatin looping interactions and TAD formation could be used to improve genome-wide contact maps and predict long-range interactions in diverse tissues and conditions. Therefore, this thesis addresses the question whether TADs represent only structural units of the genome or also important functional building blocks in which gene regulation is coordinated. By computationally integrating genome-wide chromatin interaction maps with diverse genomic datasets, including sequence conservation, regulatory activity, protein binding, gene expression and clinical phenotypes, I address the following questions.

Is the three-dimensional folding structure of genomes associated with co-regulation of functionally related genes?

- How are paralog genes distributed in the linear genome and in the three-dimensional genome architecture?
- Are paralogs co-regulated with shared enhancers and located in the same TAD in genomes of human and other species?
- Can we learn about the evolutionary history of genes and how they are created within regulatory environments of TADs?

Are TADs functional building blocks of genomes and subjected to selective pressure during evolution?

- Are human TAD regions conserved during evolution or disrupted by rearrangements when compared to other vertebrate genomes?
- Have genes within TADs a more conserved expression profile across different tissues?
- Are disruptions of TADs during evolution associated with changes in gene expression profiles?

Can clinical phenotypes be explained by rearrangements affecting TADs and the regulation of relevant genes?

- Can TADs be used to interpret gene regulatory effects of balanced chromosomal rearrangements in whole-genome sequenced patients?
- How can we quantify the similarity of phenotypes observed in patients and phenotypes associated with genes in order to prioritize candidate genes?
- Can we provide a computational tool to integrate functional genomic elements, chromatin interaction data, and TADs with phenotype data of patients to predict pathomechanism of structural variations?

Can ChIP-seq signals and sequence features predict chromatin looping interactions?

- Does the cross-linking effect in ChIP-seq provide specific signals at interacting chromatin loop anchors?
- Does the genomic sequence encode features that are predictive for long-range chromatin interactions?
- Can we provide a computational method to predict chromatin looping interactions in specific cell-types and conditions of interest?
- Which transcription factors are most predictive and eventually functionally involved in chromatin looping?

Structure of this thesis

The following four chapters address and discuss the questions raised above. First, the focus is on duplicated genes in the human genome. Because of their related sequence and function, shared evolutionary history, and close colocalization in the genome they represent an interesting model to study how genome folding is related to regulation of gene expression during evolution (Chapter 2). Next, whole-genome alignment between human and other vertebrate genomes are used to systematically investigate whether TADs represent conserved building blocks of genomes and whether rearrangements of TADs lead to altered gene expression programs (Chapter 3). The next chapter address disruptions of chromatin organization by analyzing disease associated rearrangement breakpoints from whole-genome sequenced patients of various genetic diseases to explain their phenotypes by miss-regulation of genes in disrupted TADs (Chapter 4). Finally, I make use of recent insights in chromatin loop formation to provide a computational tool for predicting long-range chromatin contacts from largely available ChIP-seq data, with the aim to facilitate three-dimensonal folding analyisis in diverse tissues and conditions for which Hi-C like data is not available (Chapter 5). The overall findings are then discussed together with further research perspectives (Chapter 6).

2 Paralog genes in the 3D genome architecture

Preamble

This chapter was published as a first-author paper in the journal Nucleic Acids Research:

Ibn-Salem J, Muro EM, Andrade-Navarro MA. ***Co-regulation of paralog genes in the three-dimensional chromatin architecture***. Nucleic Acids Res. 2017;45(1):81-91. doi:10.1093/nar/gkw813.

The publication is availible online: https://doi.org/10.1093/nar/gkw813. My contributions to this publication is indicated in Table E.1. The source code of the complete analysis is available at GitHub: https://github.com/ibn-salem/paralog_regulation. Supplementary figures are shown in Appendix A.

Abstract

Paralog genes arise from gene duplication events during evolution, which often lead to similar proteins that cooperate in common pathways and in protein complexes. Consequently, paralogs show correlation in gene expression whereby the mechanisms of co-regulation remain unclear. In eukaryotes, genes are regulated in part by distal enhancer elements through looping interactions with gene promoters. These looping interactions can be measured by genome-wide chromatin conformation capture (Hi-C) experiments, which revealed self-interacting regions called topologically associating domains (TADs). We hypothesize that paralogs share common regulatory mechanisms to enable coordinated expression according to TADs. To test this hypothesis, we integrated paralogy annotations with human gene expression data in diverse tissues, genome-wide enhancer–promoter associations and Hi-C experiments in human, mouse and dog genomes. We show that paralog gene pairs are enriched for co-localization in the same TAD, share more often common enhancer elements than expected and have increased contact frequencies over large genomic distances. Combined, our results indicate that paralogs share common regulatory mechanisms and cluster not only in the linear genome but also in the three-dimensional chromatin architecture. This enables concerted expression of paralogs over diverse cell-types and indicate evolutionary constraints in functional genome organization.

Introduction

Paralog genes arise from gene duplication events during evolution. The resulting sequence similarity between paralog pairs might lead to similar structure and function of encoded proteins (Koonin, 2005). Since paralogs often form part of the same protein complexes and pathways, it is advantageous for the cell to coordinate their expression (Makova and Li, 2003).

In eukaryotes, genes are regulated in part by binding of transcription factors to promoter sequences and to distal regulatory regions such as enhancers. By chromatin looping, enhancer bound proteins can physically interact with the transcription machinery at the promoter of genes (Ptashne, 1986; Deng et al., 2012; Carter et al., 2002; Tolhuis et al., 2002; Spitz and Furlong, 2012). These chromatin looping events can be measured by chromatin conformation capture (3C) experiments (Dekker et al., 2002), which use proximity-ligation, and more recently high-throughput sequencing (Hi-C) to measure DNA-DNA contact frequencies genome-wide (Lieberman-Aiden et al., 2009).

These interaction maps revealed tissue-invariant chromatin regions, named topologically associating domains (TADs), which have more interactions within themselves than with other regions (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). TADs seem to be stable across cell types and conserved between mammals (Dixon et al., 2012; Rao et al., 2014; Vietri Rudan et al., 2015). Regions within TADs show concerted histone chromatin signatures (Dixon et al., 2012; Sexton et al., 2012), gene expression (Le Dily et al., 2014; Nora et al., 2012), and DNA replication timing (Pope et al., 2014). Furthermore, disruption of TAD boundaries is associated to genetic diseases (Ibn-Salem et al., 2014; Lupiáñez et al., 2015).

We wondered if the Hi-C data could reveal evolutionary pressure driving paralogous expansion to favour the clustering of paralogs in the three-dimensional chromatin architecture and their regulation by common enhancer elements to enable the cell to fine-tune and coordinate their expression. To do this, we collected Hi-C data from a number of studies profiling contacts in several cell types from human (Dixon et al., 2012; Rao et al., 2014), mouse and dog (Vietri Rudan et al., 2015), and we compared the properties of these data with respect to paralog genes. Our results pinpoint that pairs of paralog genes tend to be co-regulated and co-occur within TADs more often than equivalent control gene pairs. When placed in different TADs, paralogs still tend to co-occur in the same chromosome and have more contacts than control gene pairs. In contrast, close paralogs in the same TAD have significantly less contacts with each other than comparable gene pairs, which could indicate that these pairs of paralogs encode proteins that functionally replace each other.

These observations have relevance for the study of the evolution of chromatin structure and suggest that tandem duplications generating paralogs are under selection

according to how they contribute or not to the fine structure of the genome as reflected by TADs. Thus TADs provide a favorable environment for the co-regulation of duplicated genes, which is likely followed by the evolutionary generation of additional regulatory mechanisms allowing the separation of paralogs into different TADs in the same chromosome but connected, and eventually their migration into different chromosomes.

Materials and methods

Selection of pairs of paralog genes

All human genes and human paralog gene pairs were retrieved from Ensembl GRCh37 (Ensembl 75) database by using the `biomaRt` package (Durinck et al., 2009, 2005) from within the statistical programming environment R. For each gene we downloaded the Ensembl gene ID, HGNC symbol, transcription sense, transcription start site (TSS) coordinates, and gene length. We only considered protein coding genes with "KNOWN" status that are annotated in the 22 autosomes or the 2 sexual chromosomes. For each gene we used the earliest TSS coordinate. Within this set of genes, all pairs of human paralog genes were downloaded from Ensembl (Vilella et al., 2009). This resulted in a total of 19,430 human genes; more than half of those had at least one human paralog gene (Fig. A.1A).

However, many human genes have more than one paralog (Fig. A.1B). To avoid overrepresentation of genes, we filtered the pairs such that each gene occurred only once. Thereby we selected the pairs by minimizing the rate of synonymous mutations (dS) between them using a maximum-weighted matching graph algorithm implement in the python package `NetworkX` (Galil, 1986). The number of synonymous mutations between paralogs has been used to approximate the duplication age (Lan and Pritchard, 2016). Therefore our implementation favours the selection of young paralog pairs for larger paralog families and guaranties that each gene occurs only once. This filtering strategy resulted in 6256 unique paralog pairs for downstream analysis (Table 2.1). We observed that modications of this strategy to select unique paralog genes did not affect essentially the results of our study (e.g. by selecting pairs while maximising dS; Fig. A.2).

Analogously to the human data we downloaded all pairs of protein coding paralog genes from the *Mus musculus* (GRCm38.p2) and *Canis lupus familiaris* (CanFam3.1) genomes from Ensembl. The numbers of filtered gene pairs are shown in Table 2.1 . Furthermore, we related human paralog genes to orthologs in mouse and dog only if there was a unique one-to-one orthology relationship reported in the Ensembl database.

Table 2.1.: Filtering of human paralog gene pairs

Paralog pairs	Human	Mouse	Dog
All paralog pairs	46546	110490	28293
One pair per gene	6256	7323	4959
On the same chromosome	1560	2397	658
Close pairs (TSS distance $\leq$ 1 Mb)	1114	1774	455
Distal pairs (TSS distance $>$ 1 Mb)	446	623	203

Enhancers to gene association

Human enhancer annotations, including their genome locations and the corresponding genes they regulate, were obtained from the supplementary data of a recent CAGE analysis (Andersson et al., 2014). In this study, the activity of enhancers and genes was correlated within 500kb over hundreds of human cell types to provide a regulatory interaction map between 27,451 enhancers and 11,604 genes consisting of 66,942 interactions.

Topological associating domains

We obtained topological associating domain (TAD) calls from two recently published Hi-C studies in human cells (Dixon et al., 2012; Rao et al., 2014). TAD locations mapped to the hg18 genome assembly were converted to hg19 using the UCSC liftOver tool (Hinrichs et al., 2006). A/B-compartment and sub-compartment annotations were obtained from high-resolution Hi-C experiments in human GM12878 cells (Rao et al., 2014).

Hi-C interaction maps

Individual chromatin-chromatin contact frequencies from IMR90 cells at 5 kb resolution were retrieved from (Rao et al., 2014)(NCBI GEO accession: GSE63525). We used only reads with mapping quality $\geq$ 30 and normalized the raw contact matrices applying the provided normalization vectors for KR normalization by the matrix balancing approach (Knight and Ruiz, 2013). We only considered pairwise gene interactions if the TSSs of the two genes were located in different bins of the Hi-C matrix with normalized contacts $\geq$ 0. Capture Hi-C data between promoter regions in human GM12878 cells were downloaded from ArrayExpress (accession: E-MTAB-2323) (Mifsud et al., 2015).

Randomization

We analysed the distribution of paralog pairs over chromosomes depending on the linear distance between them. For doing so, we sampled gene pairs from all human

genes with equal and independent probability and refer to them as random gene pairs.

For strand analysis, co-localisation in TADs, and Hi-C contact quantification between paralog pairs, we constructed a carefully sampled control set of gene pairs as null-model. Thereby we accounted for the linear distance bias observed for paralog pairs. First, we calculated all possible non-overlapping pairs of human genes on the same chromosome. From the resulting set of gene pairs we randomly sampled pairs according to the linear distance distribution of paralog gene pairs. Therefore, we assigned to each gene pair a sampling weight that is proportional to the probability to sample the pair. The sampling weight $w(g_i, g_j)$ for a given pair of genes g_i and g_j with absolute distance $d_{i,j}$ is defined as

$$w(g_i, g_j) = \frac{f_{\text{paralogs}}(d_{i,j})}{f_{\text{all}}(d_{i,j})}$$

where f_{paralogs} is the observed frequencies of distances in the paralog genes and $f_{\text{all}}(d_{i,j})$ the frequency of pairwise distances in the population of gene pairs from which we sample. We computed the observed frequencies by dividing the distances into 90 equal-sized bins after log_{10} distance transformation and counted occurrences of gene pairs for each bin. The resulting sampling weights for all gene pairs are normalized to sum up 1 and were then used as probabilities for sampling:

$$p_{\text{dist}}(g_i, g_j) = \frac{w(g_i, g_j)}{\sum_{i,j} w(g_i, g_j)}$$

Next, for comparison of shared enhancers we slightly modified the sampling of gene pairs to account for the observation that paralogs tend to be associated to more enhancers than non-paralogs (Fig. A.1D). Assuming that the number of enhancers associated to genes is independent from the distance, we computed sampling probabilities by

$$p_{\text{dist+eh}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j)$$

whereby n_i and n_j are the number of enhancers associated to g_i and g_j, respectively and $p_{\text{eh}}(n)$ is the probability to sample a gene associated to n enhancers:

$$p_{\text{eh}}(n) = \frac{w_{\text{eh}}(n)}{\sum_{i=0}^{N} w_{\text{eh}}(i)}$$

and

$$w_{\text{eh}}(n) = \frac{f_{\text{paralogs}}(n)}{f_{\text{all}}(n)}$$

where $f_{\text{paralogs}}(n)$ and $f_{\text{all}}(n)$ gives the frequency of genes associated to n enhancers observed in the paralog pairs and all gene pairs, respectively.

Analogously, we sampled sets of pairs accounting additionally for the observed bias in paralog pairs to be in the same strand.

$$p_{\text{dist+eh+strand}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j) \cdot p_{\text{strand}}(s_{i,j})$$

whereby $s_{i,j}$ is 1 if both genes, g_i and g_j, are transcribed from the same strand and 0 otherwise. The probability $p_{\text{strand}}(s_{i,j})$ is computed in the same way as the probability by number of enhancers $p_{\text{eh}}(n)$ in equation (2.2.5).

Lastly, we sampled a set of gene pairs by taking additionally the gene length into account and computed sampling probabilities by

$$p_{\text{dist+eh+len}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j) \cdot p_{\text{len}}(l_i) \cdot p_{\text{len}}(l_j)$$

whereby $p_{\text{len}}(l)$ for gene length l is computed in the same way as for distances between gene pairs (equation (2.2.5)) and by dividing gene lengths into 20 equal sized binds after log_{10} transformation of gene lengths in bp.

For each paralog pair on the same chromosome within 1 Mb distance, we sampled 10 random gene pairs with this procedures each resulting in $n = 156,000$ sampled gene pairs that served as background in our statistical analysis. These sampling approaches resulted in similar distribution of linear distances (Fig. A.3), associated enhancers of each gene (Fig. A.4), same strand (Fig. A.5), and gene lengths (Fig. A.6).

Statistical tests

We compared observed fractions of gene pairs, on the same chromosome, with the same transcription sense, within the same TAD or compartment, and with at least one shared enhancer between pairs of paralogs and random or sampled pairs using the Fisher's exact test. Hi-C contact frequencies and genomic distances between TSS of gene pairs were compared using a Wilcoxon rank-sum test. All analyses were carried out in the statistic software R version 3.2.2.

Results

Distribution of paralog genes in the human genome

Paralogs are homologous genes that arise from gene duplication events. Their common ancestry and replicated sequence often leads to similar structure and function in related pathways and protein complexes. We therefore hypothesised that the transcription of paralogs should have a tendency for co-regulation, which could correspond to their position in the genome and within TADs. To test this hypothesis, we first focused on the positions of paralogs in the linear genome.

From all $19,430$ protein coding genes in the human genome, $13,690$ (70.5%) have at least one paralog (Fig. A.1A). However, many human genes have several paralogs (Fig. A.1B). From all $46,546$ paralog gene pairs we filtered for only one pair per gene $(n = 6,256)$ and further for non-overlapping pairs on the same chromosome $(n = 1,560)$ (see). We will refer to close paralogs if their transcription start sites (TSSs) are within 1 Mb of each other $(n = 1,114)$ and to distal pairs for paralogs with TSSs separated by more than 1 Mb $(n = 446)$ (Table 2.1).

We first compared basic properties between genes that have at least one paralog copy and genes without human paralogs. Paralogs have significantly larger gene length than non-paralog genes ($p = 1.7 \times 10^{-53}$, Wilcoxon rank-sum test, Fig. A.1C), which fits the observation from (He and Zhang, 2005) in yeast. Furthermore, paralogs tend to be associated to more enhancers compared to non-paralog genes (on average 3.8 vs. 2.5 enhancers per gene, $p = 2.89 \times 10^{-70}$, Fig. A.1D) and the distance to the nearest associated enhancer is significantly shorter ($p = 2.71 \times 10^{-22}$,Fig. A.1E).

Since most genome duplication events in humans emerge through tandem duplications (Newman et al., 2015), we expected some co-localization among pairs of paralog genes. Indeed 24.9% of paralog pairs are located on the same chromosome. We compared this to random expectation by sampling random gene pairs from all protein coding human genes and found only 5.3% of randomly sampled gene pairs on the same chromosome ($p < 10^{-16}$, Fig. 2.1A).

We further analysed whether paralog pairs tend to be located in close genomic distance on the same chromosomes. We compared the distance between paralog gene pairs to the distance of completely random genes on the same chromosome. As expected there is a strong bias of genomic co-localization among paralog gene pairs that is not observed for random gene pairs ($p = 4.3 \times 10^{-32}$, Fig 2.1B).

We also observed that close paralog genes show more often than expected the same transcription orientation. From all paralog pairs within 1 Mb on the same chromosome 66.1% have the same sense. This is significantly more than for randomly sampled genes with the same distance (52.6%, $p = 3.2 \times 10^{-18}$, Fig. 2.1C).

Furthermore, we observed that paralogs in the same strand are closer to each other on the chromosome than pairs in opposite strands ($p = 3.48 \times 10^{-8}$, Fig. 2.1D).

Together, this shows that paralogs tend to be located within short linear distance on the same chromosome and same transcription sense, which might enable coordinated regulation by shared regulatory mechanisms.

Co-expression of paralog gene pairs across tissues

To assess whether paralog genes tend to be indeed co-regulated we compared gene expression of paralog gene pairs over several human tissues and cell lines.

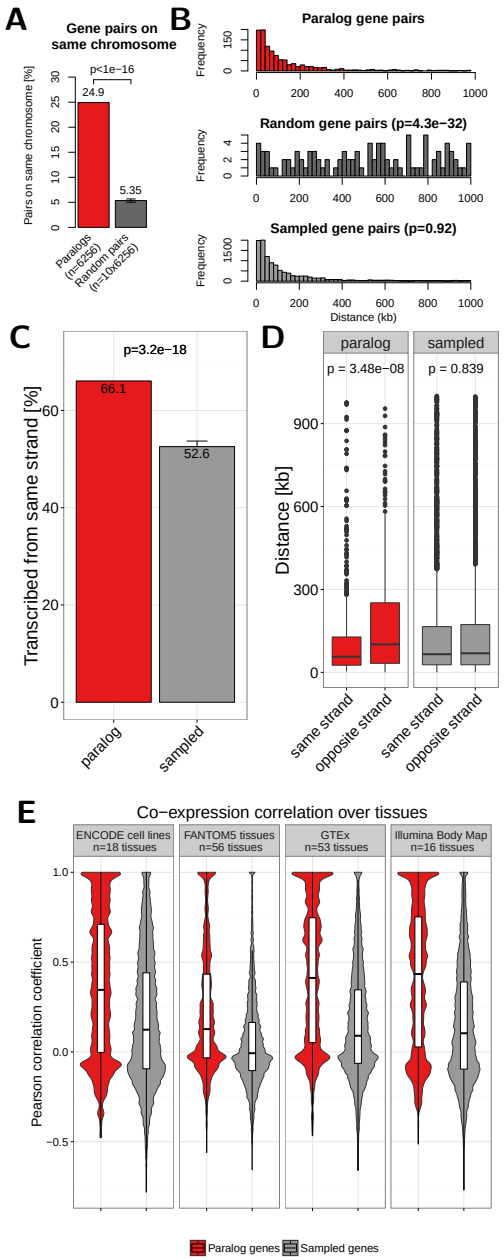

Figure 2.1.: **(A)** Percent of paralog (red) and random (dark grey) gene pairs that are located on the same chromosome. The error bar indicates the standard deviation observed in 10 times replicated random sampling of gene pairs. **(B)** Genomic distance distribution of paralogs gene pairs (top), random gene pairs (center) and gene pairs sampled according to distance distribution of paralogs (bottom). Distances are measured in kilo base pairs (kb) between TSS of genes in pairs. P-values are calculated using Wilcoxon rank-sum test. **(C)** Percent of paralog (red) and sampled (grey) gene pairs that are transcribed from the same strand. Only pairs on the the same chromosome within 1 Mb are considered here. Error bars indicate the standard deviation observed in 10 times replicated sampling of gene pairs. **(D)** Boxplot of the genomic distance between paralogs and sampled gene pairs with the same or opposite strands. **(E)** Distribution of Pearson correlation coefficients of gene expression values in four independent data sets between paralog gene pairs (red) and sampled control gene pairs (grey). White boxes show 25th, 50th and 75th percent quantile of the data and the filled areas indicate the density distribution..

We compared the Pearson correlation coefficient (PCC) of gene expression values over $n = 18$ cell-lines analysed by the ENCODE consortium by RNA-seq (Djebali et al., 2012). The distribution of PCC among paralog genes is bimodal with one peak around -0.1 and another at nearly 1.0, which indicates that there exists a group of paralog pairs without expression correlation and that the expression of other paralogs is highly positively correlated. Notably, we did not find the latter signal for positive correlation in our control set of carefully sampled gene pairs (Fig. 2.1E).

We repeated the analysis with three other independent gene expression data sets form FANTOM5 ($n = 56$ tissues) (Forrest et al., 2014), GTEx ($n = 53$ tissues) (Ardlie et al., 2015) and the Illumina Body Map ($n = 16$ tissues), which we retrieved from the EBI Expression Atlas (Petryszak et al., 2015). In all data sets we found more positively correlated paralog pairs compared to the sampled gene pairs (Fig. 2.1E). This shows that many paralogs are expressed with high coordination in a tissue specific manner.

Paralog genes share enhancers

We hypothesised that common gene regulation of close paralog genes is likely to be facilitated by shared enhancer elements. Indeed we found that paralog gene pairs within 1 Mb on the same chromosome are associated to the same enhancer elements more often than expected by chance (Fig. 2.2). We estimated the expected background distribution of shared enhancers by carefully sampling gene pairs with the same distributions as paralogs in distances and associated enhancers to single genes (Fig. A.4, section 2.2).

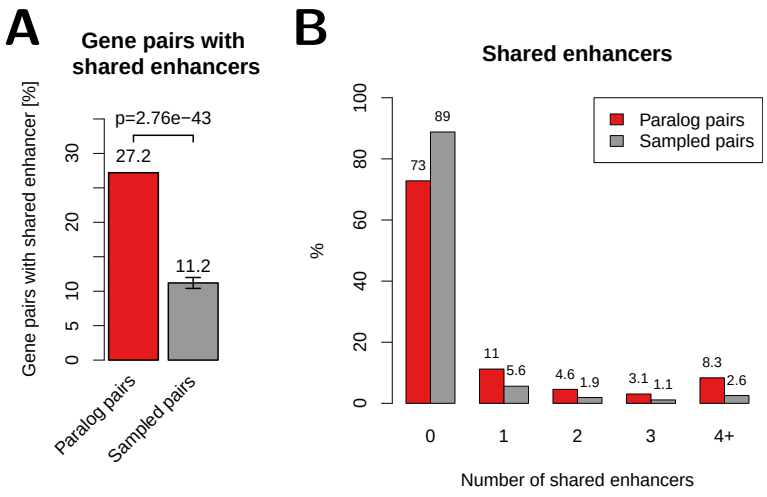

Figure 2.2.: **Shared enhancers among paralog gene pairs.** **(A)** Percent of close paralog (red) and sampled control (grey) gene pairs with at least one shared enhancer. **(B)** Percent of gene pairs versus number of shared enhancers for paralog and sampled control gene pairs.

While 27.2% of the paralog gene pairs have at least one enhancer in common, we observed this for only 11.7% of the sampled gene pairs ($p = 4.2 \times 10^{-40}$, Fig. 2.2A). This could be replicated when comparing against sampled gene pairs where in addition to distance and number of enhancers linked to single genes, also the transcription sense and gene length were taken into account during sampling of control gene pairs ($p = 3.4 \times 10^{-41}$ and $p = 5 \times 10^{-30}$, respectively; Fig. A.7). Next, we compared the percent of gene pairs with shared enhancers as a function of the number of shared enhancers between paralogs and sampled gene pairs. We observed that paralog pairs are enriched for higher number of shared enhancers compared to the sampled gene pairs (Fig. 2.2B). Together, these results indicate that paralog genes are more often co-regulated by common enhancer elements than other genes.

Co-localization of paralogs in TADs

To facilitate their function in gene regulation, distal enhancer elements need to interact physically via chromatin looping with promoter elements at the TSS of their target genes. These looping interactions occur frequently within so called topological associating domains (TADs). These are regions of hundreds of kb that show high rates of self-interactions and few interactions across domain boundaries in genome-wide Hi-C experiments (Dixon et al., 2012; Rao et al., 2014). Genes within the same TAD are therefore likely to have common gene regulatory programs (Le Dily et al., 2014; Nora et al., 2012).

We used TADs from Hi-C experiments in eight different human cell-types (HeLa, HUVEC, K562, KBM7, NHEK, IMR90, GM12878, and hESC) from two recent studies (Dixon et al., 2012; Rao et al., 2014). Notably, the called TADs differ in size between the two publications due to different resolution of Hi-C experiments and different algorithms used to call them from Hi-C contact matrices (Fig. A.8). TADs from (Rao et al., 2014) have a median size of around 240 kb and are nested, so that several small domains can occur within one ore more larger domains. In contrast TADs from (Dixon et al., 2012) are of 1 Mb on average and are defined as non-overlapping genomic intervals.

We hypothesised that paralog gene pairs might be located more often in the same TAD than expected by chance. Indeed, we found that, depending on cell-type and study, between 35% and 73% of close paralog pairs are located in the same TAD (Fig. 2.3A). In seven out of nine data sets this difference was significant ($p < 0.05$) with respect to the sampled control gene pairs with the same linear distance. We also calculated a set of $n = 2,624$ stable TADs that are found in more than 50% of cell types analysed in (Rao et al., 2014). Notably, we found for paralog pairs a 1.25 fold enrichment to be located in the same stable TADs compared to sampled gene pairs ($p = 0.00013$, stable_TADs in Fig. 2.3A).

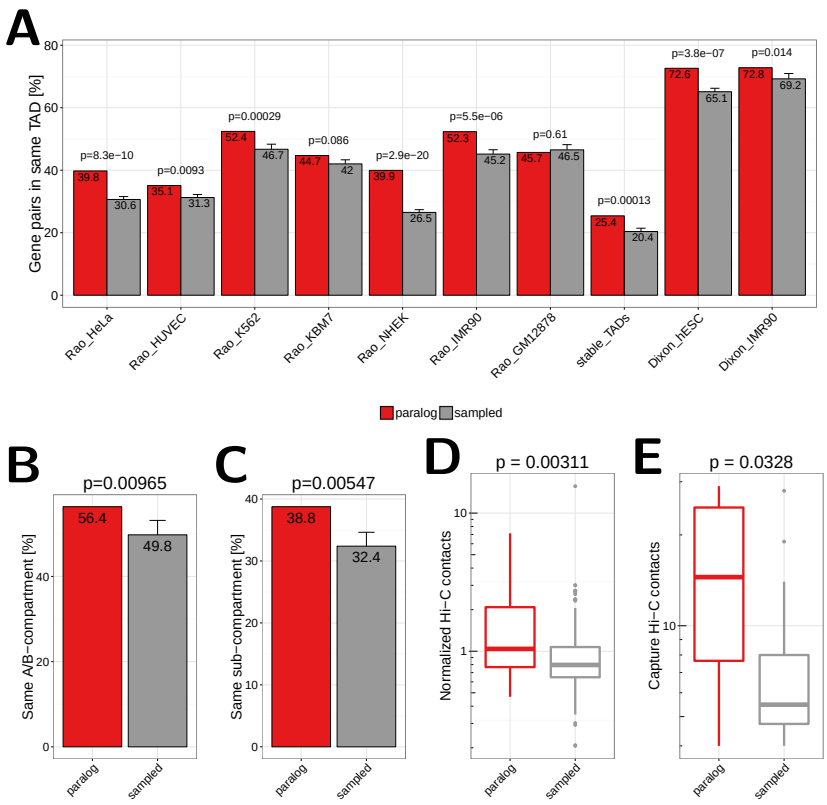

Figure 2.3.: **(A)** Co-localization of close paralog genes within the same TAD compared against sampled gene pairs for TAD data sets from different cell types and studies. The first seven bars show values for TADs called in HeLa, HUVEC, K562, KBM7, NHEK, IMR90, and GM12878 cells by (Rao et al., 2014). The eighth bar shows the value for stable TADs across cell types form this study (at least 90% reciprocal overlap in 50% of cells). The last two bars show data for TADs called in hESC and IMR90 cells by (Dixon et al., 2012). Error bars indicate standard deviation in 10 times replicated sampling of gene pairs. P-values are computed using Fisher's exact test. **(B)** Percent of gene pairs annotated to same A/B compartment according to Hi-C data in GM12878 cells from (Rao et al., 2014). Pairs located in the very same compartment interval were excluded. **(C)** Percent of gene pairs annotated to same sub compartment (A1, A2, B1, B2, B3, B4) according to (Rao et al., 2014). Pairs located in the same subcompartment interval were excluded. **(D)** Normalized Hi-C contact frequencies between TSSs of distal paralog gene pairs and sampled background gene pairs. **(E)** Promoter capture-C contact frequencies between distal paralog gene pairs and sampled background gene pairs.

Beside TADs, Hi-C interaction maps have revealed interaction patterns of two compartments (A and B) that alternate along chromosomes in intervals of several Mb. Thereby loci in A compartment preferentially associate with other loci in A and loci in B with others in B (Lieberman-Aiden et al., 2009; Rao et al., 2014; Dekker et al., 2013). We therefore asked whether pairs of paralogs from the same chromosome are preferentially located within the same compartment (both A or both B) whereby we excluded pairs that are in the same compartment interval. We found that 56.4% of paralogs on the same chromosome but not in the same compartment interval are in compartments of the same type. This was only observed for 49.2% of sampled pairs ($p = 0.0046$, Fig. 2.3B). Next we tested the same for recently distinguished sub-compartment types from high-resolution Hi-C interaction maps (Rao et al., 2014). Again, paralogs are enriched to be located within the same sub-compartment type (38.9% vs. 31.6%, $p = 0.0046$, Fig. 2.3C).

These results show that close paralogs are enriched to be located in the same regulatory unit of the genome as defined both by TADs and compartments.

Distal paralog pairs are enriched for long-range chromatin contacts

Since it was shown that actively transcribed genes are localized in the same active spatial compartments and tend to contact each other frequently in the nucleus (at their promoters (Cremer et al., 2015; Mifsud et al., 2015)) we hypothesised that this might be the case for distal paralogs on the same chromosome too. As spatial proximity can be approximated by Hi-C contact frequencies (Lieberman-Aiden et al., 2009) we compared the number of normalized Hi-C contacts between TSS of distal paralog genes (that have promoters separated by more than 1 Mb) to the sampled gene pairs with the same linear distances distribution. We used recently published in situ Hi-C data from IMR90 cells at 5kb bin-size resolution (Rao et al., 2014) and observed significantly more normalized chromatin interactions between paralog genes compared to sampled control gene pairs ($p = 0.0022$, Fig. 2.3D). We furthermore used an independent data set of high resolution promoter-promoter interactions measured by capture Hi-C (Mifsud et al., 2015). Again, we observed a strong enrichment of promoter-promoter interactions between distal paralogs compared to control genes pairs ($p = 0.027$, Fig. 2.3E). This shows that also distal paralogs are enriched for long-range interactions, indicating that they tend to be in closer spatial proximity than other genes.

Close paralogs have fewer contacts than expected

The observed enrichment of Hi-C contacts of paralogs is distance dependent. We observe for close paralogs fewer Hi-C contacts than for equally distant sampled gene pairs (Fig. 2.4A). To analyse this in more detail we focused on only those pairs

on the same chromosome that have a TSS distance of at least 10kb but less than 1Mb. This is the distance range of most paralog pairs and allows to separate genes in Hi-C interaction maps and TADs (Fig. A.9A). Consequently, we observe paralogs more often in the same TAD in eight out of nine data sets for this distance range (Fig. A.9B). For these pairs we observe significant lower Hi-C contact frequencies if pairs are within the same IMR90 TAD (Rao et al., 2014) as compared to sampled genes ($p = 0.00094$) but not if pairs are in different TADs ($p = 0.81$, Fig. 2.4B). We got comparable results when analysing the Capture Hi-C data the same way (Fig. A.9C). Next, we tested whether this can be explained by the nested sub-TAD structure of TADs called from high-resolution Hi-C in IMR90 (Rao et al., 2014). We divided pairs into four groups, namely, 'no TAD', if both pairs are not in any TAD, 'different TAD', if pairs do not have at least one TAD in common, 'different sub-TADs', if they have at least one TAD in common but are in different sub-TADs, and 'same sub-TAD', if they overlap exactly the same set of TADs. While we saw that paralogs are more often in the no TAD group ($p = 1.4 \times 10^{-20}$), we found that they were highly depleted from the different TAD group ($p = 1.6 \times 10^{-40}$) and highly enriched to be located within the same sub-TAD ($p = 4.2 \times 10^{-9}$, Fig. 2.4C). However, although not always significant, paralogs have fewer Hi-C contacts than sampled gene pairs in all of these groups (Fig. 2.4D). In addition, close paralogs within the same TAD share more enhancers than close paralogs not being in the same TAD (Fig. 2.4E). However, the positive correlation of gene expression over different tissues is not significantly higher for paralogs whether they are in the same TAD or not (Fig. A.10).

In summary, we observed that while close paralogs (situated at less than 1Mb) have more shared enhancers if they are in the same TAD than not, these within TAD paralog pairs have fewer contacts compared to other within TAD pairs of genes.

Paralogs in mouse and dog genome

Next, we asked whether the co-localization and co-regulation of paralogs is conserved in other species. For this, we conducted an analogous analysis with paralog gene pairs from mouse (*M. musculus*) and dog (*C. familaris*) genomes. Similar as for human data, we found that more than two third of the genes had at least one paralog copy (Fig. A.11A,D), paralog pairs clustered on the same chromosome (Fig. A.11B,E), and had close linear distances (Fig. A.11C,F).

We sampled control gene pairs with the same distance distribution as paralogs for both species separately (Fig. A.11C,F). We used TADs from recently published Hi-C data in liver cells of mouse and dog (Vietri Rudan et al., 2015), which have a size distribution comparable to TADs from human cells (Fig. A.8). We computed the fraction of paralog pairs that are located in the same TAD for both species. Consistent with the observation in human, we found that paralogs tend to colocalize more

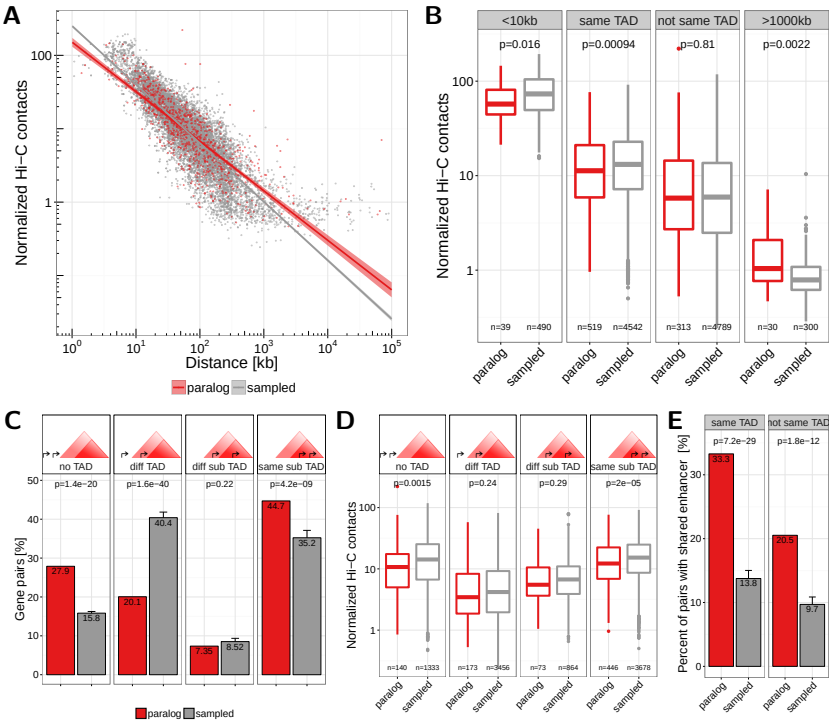

Figure 2.4.: **(A)** Normalized Hi-C contacts by genomic distance between paralog (red) and sampled (grey) gene pairs. Lines show linear regression fit separately for paralogs (red) and sampled (grey) pairs with 95% confidence intervals in shaded areas. **(B)** Normalized Hi-C contacts between pairs of paralogs (red) and sampled gene pairs (grey) for the groups: $<$10kb genomic distance, located in the same TAD, not in the same TAD, and with genomic distance $>$1000kb. **(C)** Number of gene pairs located either in no TAD, in different TADs (or only one pair member in a TAD), both in a TAD but in different sub-TADs, or within the same sub-TAD, for paralogs (red) and sampled (grey) pairs. TADs from IMR90 cells from (Rao et al., 2014) were used, which nested in contrast to TAD calls from (Dixon et al., 2012). **(D)** Normalized Hi-C contacts between pairs of paralogs (red) and sampled gene pairs (grey) for the four groups of pairs in sub-TAD structures shown in (C). **(E)** Percent of gene pairs with at least one shared enhancer for paralog genes (red) and sampled control genes (grey) separated for pairs in the same IMR90 TAD (left) or not (right).

frequently within the same TAD in mouse ($p = 7.2 \times 10^{-22}$) and dog ($p = 0.00064$) than expected by chance (Fig. 2.5A). We also quantified directly the contact frequencies between promoters of distal paralogs on the same chromosome and found them significantly more frequently in contact than sampled gene pairs with the same genomic distance for paralogs in mouse ($p = 7 \times 10^{-7}$) and dog ($p = 0.008$) (Fig. 2.5B). Together, these results indicate that enriched long-range interactions between paralogs are not human specific but rather a general evolutionary conserved feature of genome organization.

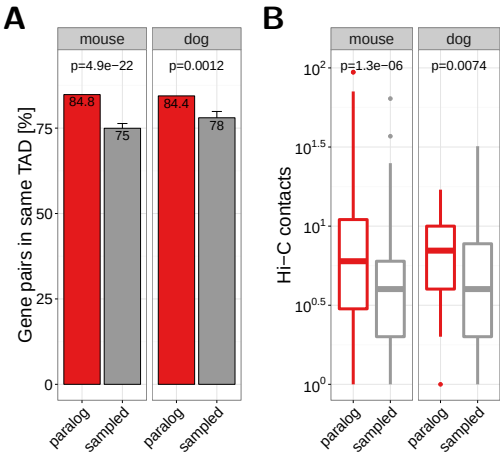

Figure 2.5.: **(A)** Co-occurrence of close paralog genes with the same TAD in mouse (left panel) and dog (right panel). **(B)** Hi-C contacts between promoter of distal gene pairs in Hi-C experiments in liver cells from mouse (left panel) and dog (right panel). Hi-C data and TAD calls were taken from (Vietri Rudan et al., 2015).

Orthologs of human paralogs show conserved co-localization

Next, we wanted to test more directly whether the spatial co-localization of human paralogs is indeed conserved during evolution. In cases where the gene duplication event occurred before the separation of human and mouse (or human and dog) we can eventually assign each human gene of a pair of paralogs to one ortholog in mouse (or dog genomes) (Fig. A.12).

We could map 37.1% ($n = 579$) and 34.6% ($n = 540$) of the close human paralogs to one-to-one orthologs in mouse and dog, respectively (Fig. A.13A,D). We hypothesised that the two one-to-one orthologs of human paralog pairs would also be close in the mouse and dog genomes. Indeed, we found that the orthologs of human paralogs tend to cluster on the same chromosome (Fig. A.13B,E) and are biased for close linear distances (Fig. A.13C,F).

We further investigated how many one-to-one orthologs of the human paralog pairs were located in the same TAD in mouse and dog genomes. Although not significant, we found that mouse orthologs of close human paralogs share more often the same TAD in mouse than orthologs of sampled human gene pairs (80% vs. 76%, $p = 0.11$; Fig. 2.6A). Significant enrichment was observed with orthologs in the dog genome (85% vs. 77%, $p = 0.0016$; Fig. 2.6A).

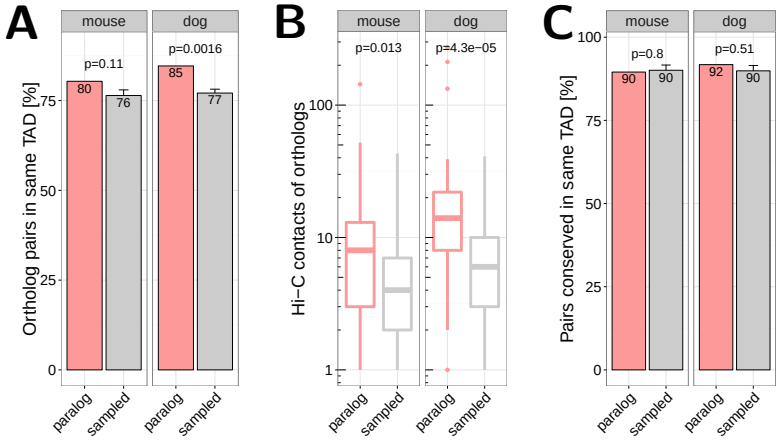

Figure 2.6.: ** One-to-one orthologs of human paralog genes in mouse and dog genome.** **(A)** Percent of mouse (left) and dog (right) orthologs of human paralog pairs that are in the same TAD in the mouse and dog genome, respectively. **(B)** Normalized Hi-C contacts between promoters of one-to-one orthologs of human distal paralogs in the mouse (left) and dog (right) genome. **(C)** Percent of gene pairs with conserved co-localization. Orthologs in the same TAD in mouse (left) and dog (right) as percent of all orthologs of human paralog pairs that are in the same TAD in human. For human TADs from IMR90 cells from (Rao et al., 2014) were used.

For distal human paralogs we quantified the promoter contacts of their orthologs in mouse and dog and found enriched Hi-C contacts in mouse ($p = 0.011$) and dog ($p = 2.4 \times 10^{-5}$; Fig. 2.6B).

These results show that both the co-localization of paralogs in TADs and the contacts between distal paralogs are only weekly conserved at the evolutionary distances examined here. For example, we see that given a pair of human genes in the same TAD the likelihood of their orthologs being in the same TAD in mouse or dog is the same whether they are paralogs or not (Fig. 2.6C).

All together, our results support the notion that tandem duplications generate paralog gene pairs that are selected if they accommodate in TADs but following evolutionary events allow their reorganization outside TADs. While within oganisms distal paralog genes are coordinated, such coordination can be eventually erased by evolution.

Discussions

The generation of large datasets of gene expression across multiple tissues allowed the observation of clusters of pairs and triplets of co-expressed genes in higher eukaryotes (e.g. in *Drosophila* (Boutanaev et al., 2002) or in mammals (Purmann et al., 2007)) and it was previously suspected that the structure of chromatin would have to do with this (Sproul et al., 2005), particularly cis-acting units (Purmann et al., 2007). The discovery and characterization of topologically associating domains (TADs) has finally brought to the light the chromatin structure that could be responsible for this co-regulation.

To study the interplay between TADs, gene co-regulation and evolution in the human genome, we decided to focus on pairs of paralogs because they have a tendency to be produced by tandem duplication (Newman et al., 2015) and, because of homology, result in proteins with related functions. However, the particular emergence and evolution of paralogs are probably responsible for special properties that distinguish them from non-paralog genes as we described: greater gene length, more enhancers, as well as a shorter distance to the next enhancer. These differences, which could be partially explained by the observation that paralogs are more often tissue specific (Fig. A.1F), complicated the methodology for choosing meaningful control pairs (see section 2.2).

Once we ensured the generation of the appropriate backgrounds, we could study the position of pairs of paralogs respect to TADs. This allowed us to test, on the one hand, the resilience of TADs to genome shuffling and, on the other hand, the rate of accommodation and gain of functionally related genes. Possibly, the generation of paralogs by tandem duplication might continuously impose a strain in the pre-existing genomic and regulatory structure, but also a chance for the evolution of new functionality.

On the one hand, we observed many pairs of paralogs within TADs. On the other hand, pairs of paralogs in different TADs, however distant from each other, tend to have more contacts than control gene pairs. This suggests a many-step mechanism where first tandem duplication fits TAD structure but then subsequent chromosomal rearrangements relocate paralogs at larger distances (while keeping contacts) and eventually reorganization of regulatory control allow their increased independence being eventually placed even in different chromosomes where contact is no longer necessary. Thus, TADs are units of co-regulation but do not have a strong preference for keeping co-regulated genes within during evolution. This model agrees with the recent work from Lan and Pritchard reporting that young pairs of paralogs are generally close in the genome (Lan and Pritchard, 2016).

A second effect that we observed was the existence of fewer contacts between close pairs of paralogs than in comparable pairs of non-paralog genes, particularly if they

are in the same TAD (Fig. 2.4B), while sharing more enhancers (Fig. 2.4E). This result could reflect the existence of pairs of paralogs encoding proteins that replace each other, for example sub-units of a complex that occupy the same position in a protein complex but are expressed in different cells. One such case is exemplified by CBX2, CBX4 and CBX8, which occupy neighbouring positions within the same TAD in human chromosome 17 and encode replaceable subunits of the polycomb repressive complex 1 (PRC1) complex involved in epigenetic regulation of cell specification (Becker et al., 2015). The expression of such groups of paralogs require active coordination to ensure exclusive expression of only one gene or a subset of genes per condition, resulting in patterns of divergent expression. Since there might be also conditions where none of these genes are expressed, such divergent expression patterns are different from negative correlation.

Previous work studying gene expression of duplicated genes already studied how after gene duplication paralogs tend to diverge in their expression (Makova and Li, 2003; Huminiecki, 2004; Rogozin et al., 2014) but it was observed that while some paralogs are co-expressed some others have negative correlation across tissues (Makova and Li, 2003). Our interpretation of these observations together with our results is that the initial tandem duplication event forming a paralog is advantageous to situate the new copy in an environment that allows its controlled regulation, ideally under the same regulatory elements than the original copy, and this can be attained by duplicating both gene and surrounding regulatory elements. This would preclude the duplication of genes with very entangled regulatory associations. Once this happens, if the new protein evolves into a replacement, then the regulatory constraints on its coding gene are strong and there would be a tendency to keep it in the vicinity of the older gene so that a divergent pattern of expression can be ensured.

To support this hypothesis, we contrasted our data with the data collected in the HIPPIE database of experimentally verified human protein-protein interactions (Schaefer et al., 2012). We observed the well-known fact that paralog pairs generally encode for proteins that interact more often than non-paralog proteins (Fig. A.14). But, most importantly, we observed that the chances of close pairs of genes to encode for interacting proteins raise 2.3-fold if they are in the same TAD, while, in contrast, if these genes are paralogs the difference is much smaller (1.2-fold, Fig. A.14). We interpret this result as evidence for a significant population of within TAD paralog pairs encoding for non-interacting proteins, which supports our hypothesis that paralog pairs within the same TAD would have a tendency to encode for proteins replacing each other.

Conclusion

We propose that paralog genes generated by tandem duplication start their life coregulated within TADs, then are moved outside to other places in the chromosome and eventually to different chromosomes. TADs would then fit genomic duplications situating the new copy in a duplicated regulatory enviroment. Subsequent genomic rearrangements would create divergent regulatory circuits eventually allowing their disentanglement. An exception would be genes that precise to be strongly co-regulated with the original copy, for example, to produce a replacement protein.

TADs would thus act as protective nests for evolving newcomer genes. This seems to be a reasonable evolutionary mechanism, much simpler than creating from nothing a complete new regulatory environment for a new gene.

Acknowledgements

The authors thank all members of the CBDM group for fruitful discussions.

Stability of TADs in evolution

3

Preamble

This chapter is submitted as a corresponding-author paper to BMC Biology and is currently under review. A preprint is published at bioRxiv:

Krefting J, Andrade-Navarro MA, **Ibn-Salem J**. ***Evolutionary stability of topologically associating domains is associated with conserved gene regulation***. bioRxiv. 2017. doi:doi.org/10.1101/231431.

The preprint is availible online: https://doi.org/10.1101/231431. My contributions to this publication is indicated in Table E.1. The source code of the complete analysis is available at GitHub: https://github.com/Juppen/TAD-Evolution. Supplementary figures and links to supplementary tables are shown in Appendix B.

Abstract

Background: The human genome is highly organized in the three-dimensional nucleus. Chromosomes fold locally into topologically associating domains (TADs) defined by increased intra-domain chromatin contacts. TADs contribute to gene regulation by restricting chromatin interactions of regulatory sequences, such as enhancers, with their target genes. Disruption of TADs can result in altered gene expression and is associated to genetic diseases and cancers. However, it is not clear to which extent TAD regions are conserved in evolution and whether disruption of TADs by evolutionary rearrangements can alter gene expression.

Results: Here, we hypothesize that TADs represent essential functional units of genomes, which are selected against rearrangements during evolution. We investigate this using whole-genome alignments to identify evolutionary rearrangement breakpoints of different vertebrate species. Rearrangement breakpoints are strongly enriched at TAD boundaries and depleted within TADs across species. Furthermore, using gene expression data across many tissues in mouse and human, we show that genes within TADs have more conserved expression patterns. Disruption of TADs by evolutionary rearrangements is associated with changes in gene expression profiles, consistent with a functional role of TADs in gene expression regulation.

Conclusions: Together, these results indicate that TADs are conserved building blocks of genomes with regulatory functions that are often reshuffled as a whole instead of being disrupted by rearrangements.

Keywords

Genome rearrangements; Topologically associating domains; TAD; Chromatin interactions; 3D genome architecture; Hi-C; Evolution; Selection; Gene regulation; Structural variants

Introduction

The three-dimensional structure of eukaryotic genomes is organized in many hierarchical levels (Bonev and Cavalli, 2016). The development of high-throughput experiments to measure pairwise chromatin-chromatin interactions, such as Hi-C (Lieberman-Aiden et al., 2009) enabled the identification of genomic domains of several hundred kilo-bases with increased self-interaction frequencies, described as topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012). Loci within TADs contact each other more frequently and TAD boundaries insulate interactions of loci in different TADs. TADs have also been shown to be important for gene regulation by restricting the interaction of cell-type specific enhancers with their target genes (Nora et al., 2012; Symmons et al., 2014; Zhan et al., 2017). Several studies associated disruption of TADs to ectopic regulation of important developmental genes leading to genetic diseases (Ibn-Salem et al., 2014; Lupiáñez et al., 2015). These properties of TADs suggested that they are functional genomic units of gene regulation.

Interestingly, TADs are largely stable across cell-types (Dixon et al., 2012; Rao et al., 2014) and during differentiation (Dixon et al., 2015). Moreover, while TADs were initially described for mammalian genomes, a similar domain organization was found in the genomes of non-mammalian species such as *Drosophila* (Sexton et al., 2012), zebrafish (Gómez-Marín et al., 2015) *Caenorhabditis elegans* (Crane et al., 2015) and yeast (Hsieh et al., 2015; Mizuguchi et al., 2014). Evolutionary conservation of TADs together with their spatio-temporal stability within organisms, would collectively imply that TADs are robust structures.

This motivated the first studies comparing TAD structures across different species, which indeed suggested that individual TAD boundaries are largely conserved along evolution. More than 54% of TAD boundaries in human cells occur at homologous positions in mouse genomes (Dixon et al., 2012). Similarly, 45% of contact domains called in mouse B-lymphoblasts were also identified at homologous regions in human lymphoblastoid cells (Rao et al., 2014). A single TAD boundary at the Six gene loci could be traced back in evolution to the origin of deuterostomes (Gómez-Marín et al., 2015). However, these analyses focused only on the subset of syntenic regions that can be mapped uniquely between genomes and do not investigate systematically if TAD regions as a whole might be stable or disrupted by rearrangements during evolution.

A more recent study provided Hi-C interaction maps of liver cells for four mammalian genomes (Vietri Rudan et al., 2015). Interestingly, they described three examples of rearrangements between mouse and dog, which all occurred at TAD boundaries. However, the rearrangements were identified by ortholog gene adjacencies, which might be biased by gene density. Furthermore, they did not report the total number of rearrangements identified, leaving the question open of how many TADs are actually conserved between organisms. It remains unclear to which extent TADs are selected against disruptions during evolution (Nora et al., 2013). All these studies underline the need to make a systematic study to verify if and how TAD regions as a whole might be stable or disrupted by rearrangements during evolution.

To address this issue we used whole-genome alignment data to analyze systematically whether TADs represent conserved genomic structures that are rather reshuffled as a whole than disrupted by rearrangements during evolution. Furthermore, we used gene expression data from many tissues in human and mouse to associate disruptions of TADs by evolutionary rearrangements to changes in gene expression.

Results

Identification of evolutionary rearrangement breakpoints from whole-genome alignments

To analyze the stability of TADs in evolution, we first identified evolutionary rearrangements by using whole-genome alignment data from the UCSC Genome Browser (Kent et al., 2003, 2002) to compare the human genome to 12 other species. These species where selected to have genome assemblies of good quality and to span several hundred million years of evolution. They range from chimpanzee to zebrafish (Fig 3.1). The whole-genome data consists of consecutive alignment blocks that are chained and hierarchically ordered into so-called net files as fills (Kent et al., 2003). To overcome alignment artifacts and smaller local variations between genomes we only considered top-level fills or non-syntenic fills and additionally applied a size threshold to use only fills that are larger than 10 kb, 100 kb, or 1000 kb, respectively. Start and end coordinates of such fills represent borders of syntenic regions and were extracted as rearrangement breakpoints for further analysis (see Methods for details).

First, we analyzed the number and size distributions of top-level and non-syntenic fills between human and other species (Fig 3.1). As expected, closely related species such as chimpanzee and gorilla have in general fewer fills but larger fill sizes (mean length 1 kb), whereas species which are more distant to human, such as chicken

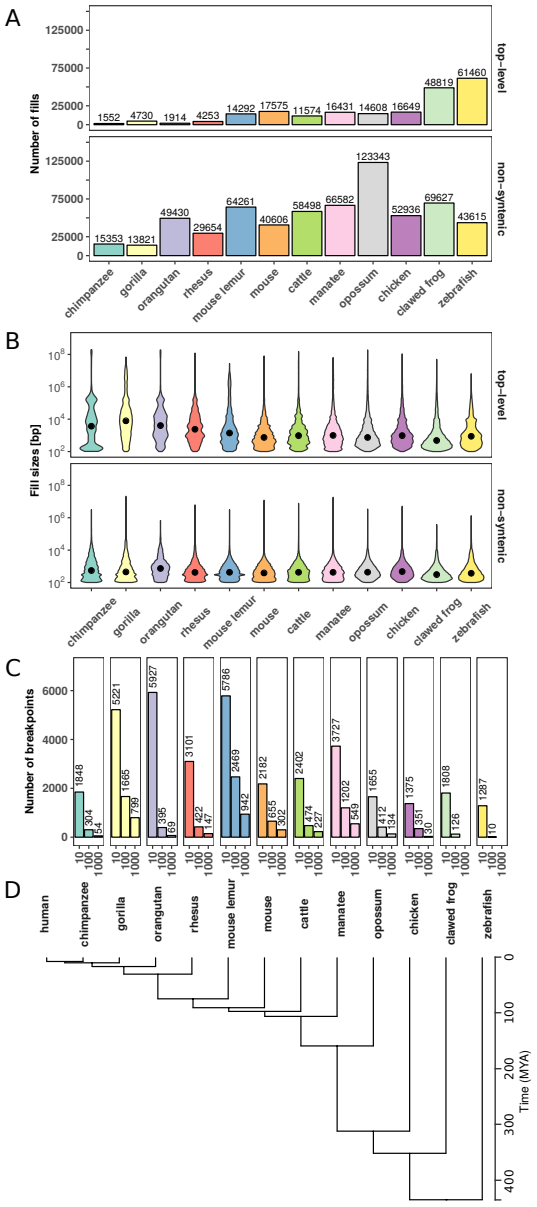

Figure 3.1.: **Number and size distributions of fill sizes of whole-genome alignments between human and 12 other species.** **(A)** Number of syntenic alignment blocks (fills) between human (hg38) and 12 other species. Top-level fills are the largest and highest scoring chains and occur at the top level in the hierarchy in net files (top panel). Non-syn fills map to different chromosomes as their parent fills in the net files (bottom panel). **(B)** Size distribution of top-level (top panel) and non-syntenic (bottom panel) fills as violin plot. **(C)** Number of identified rearrangement breakpoints between human and 12 other species. Breakpoints are borders of top-level or non-syn fills that are larger or equal than a given size threshold (x-axis). **(D)** Phylogenetic tree with estimated divergence times according to *http://timetree.org/*.

and zebrafish, tend to have more but smaller fills (mean length 1 kb, Fig 3.1A,B). However, we also observe many small non-syntenic fills in closely related species, likely arising from transposon insertions (Mills et al., 2006). As a consequence of the number of fills and size distributions, we identify different breakpoint numbers depending on species and size threshold applied. For example, the whole-genome alignment between human and mouse results in 2182, 655, and 302 rearrangement breakpoints for size thresholds, 10 kb, 100 kb, and 1000 kb, respectively (Fig 3.1C). Together, the number and size distributions of syntenic regions reflect the evolutionary divergence time from human and allow us to identify thousands of evolutionary rearrangement breakpoints for enrichment analysis at TADs.

Rearrangement breakpoints are enriched at TAD boundaries

Next, we analyzed how the identified rearrangement breakpoints are distributed in the human genome with respect to TADs. We obtained 3,062 TADs identified in human embryonic stem cells (hESC) (Dixon et al., 2012) and 9,274 contact domains from high-resolution *in situ* Hi-C in human B-lymphoblastoid cells (GM12878) (Rao et al., 2014). To calculate the number of breakpoints around TADs, we enlarged each TAD region by +/-50% of its size and divided the region in 20 equal sized bins. For each bin we computed the number of overlapping rearrangement breakpoints. This results in a size-normalized distribution of rearrangement breakpoints along TAD regions.

First, we analyzed the distribution of breakpoints at different size thresholds between human and mouse at hESC TADs (Fig. 3.2A). Rearrangement breakpoints are clearly enriched at TAD boundaries and depleted within TAD regions. Notably, this enrichment is observed for all size thresholds applied in the identification of rearrangement breakpoints. Next, we also analyzed the breakpoints from chimpanzee, cattle, opossum, and zebrafish (Fig 3.2B) at the 10 kb size threshold. Interestingly, we observed for all species a clear enrichment of breakpoints at TAD boundaries and depletion within TAD regions. To quantify this enrichment, we simulated an expected background distribution of breakpoints by placing each breakpoint 100 times at a random position of the respective chromosome. We than calculated the fraction of observed and expected breakpoints that are closer than 40 kb to a TAD boundary. For all size thresholds and analyzed species, we computed the log-fold-ratio of actual breakpoints over random breakpoints at domain boundaries (Fig 3.2C). For virtually all species and size thresholds analyzed, we found breakpoints significantly enriched at boundaries of TADs and contact domains (Fig 3.2C, B.1). Depletion was only observed for some combinations of species and size thresholds which have only very few breakpoints (see Fig 3.1C). Furthermore, we compared the distance of each breakpoint to the closest TAD boundary and observed nearly always significantly shorter distances for actual breakpoints compared to random controls (Fig B.2). Overall, the enrichment was stronger for TADs in hESC compared to the

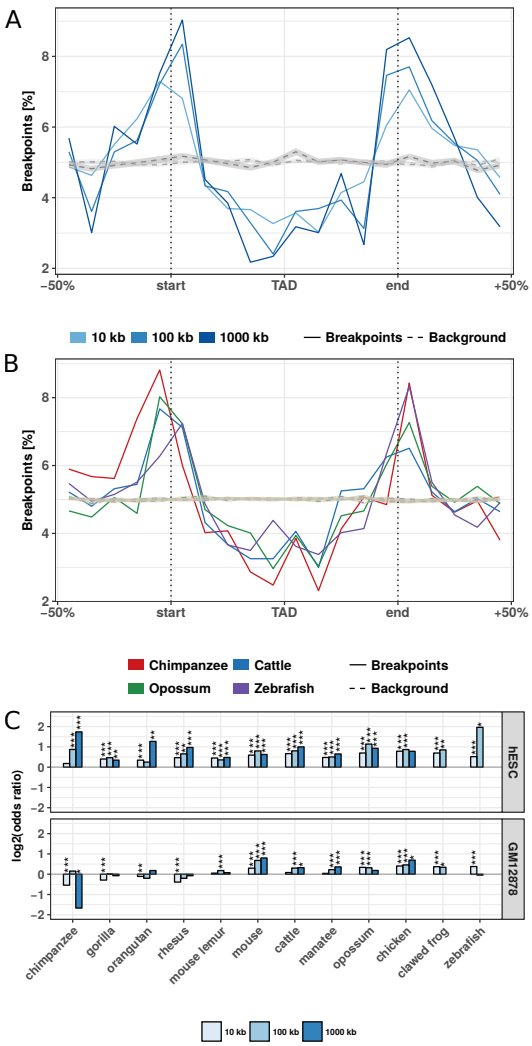

Figure 3.2.: **Evolutionary rearrangements are enriched at TAD boundaries. (A)** Distribution of evolutionary rearrangement breakpoints between human and mouse around hESC TADs. Each TAD and 50% of its adjacent sequence was subdivided into 20 bins of equal size, the breakpoints were assigned to the bins and their number summed up over the corresponding bins in all TADs. Blue color scale represents breakpoints from different fill-size thresholds. Dotted lines in gray show simulated background controls of randomly placed breakpoints. **(B)** Distribution of rearrangement breakpoints between human and: chimpanzee, cattle, opossum, and zebrafish, at 10 kb size threshold around hESC TADs. Dotted lines in gray show simulated background controls of randomly placed breakpoints. **(C)** Enrichment of breakpoints at TAD boundaries as log-odds-ratio between actual breakpoints at TAD boundaries and randomly placed breakpoints. Enrichment is shown for three different fill size thresholds (blue color scale) and TADs in hESC from (Dixon et al., 2012) (top) and contact domains in human GM12878 cells from (Rao et al., 2014) (bottom), respectively. Asterisks indicate significance of the enrichment using Fisher's exact test (*p <= 0.05; **p <= 0.01; ***p <= 0.001).

contact domains in GM12878. However, these differences were likely due to different sizes of TADs and contact domains and the nested structure of contact domains, which overlap each other (Rao et al., 2014). Rearrangements between human and both closely and distantly related species are highly enriched at TAD boundaries and depleted within TADs. These results show (i) that rearrangements are not randomly distributed in the genome, in agreement with (Farré et al., 2015), and (ii) strong conservation of TAD regions over large evolutionary time scales, indicating selective pressure against disruption of TADs, presumably because of their functional role in gene expression regulation.

Clusters of conserved non-coding elements are depleted for rearrangement breakpoints

Another interesting feature that can be extracted from whole-genome alignments are highly conserved non-coding elements (CNEs) (Polychronopoulos et al., 2017). CNEs are defined as non-protein-coding sequences of at least 50 bp with over 70% sequence identity between distantly related species such as human and chicken (Polychronopoulos et al., 2017). In the human genome, CNEs cluster around developmental genes in so-called genomic regulatory blocks (GRBs) (Kikuta et al., 2007). It has been shown recently that many GRBs coincide with TADs in human and *Drosophila* genomes (Harmston et al., 2017). Therefore, we asked whether evolutionary breakpoints are also enriched at boundaries of GRBs. This would support the idea of a conserved regulatory environment around important developmental genes. Indeed we saw a strong enrichment around GRBs (Fig 3.3A). This is consistent with previous studies in *Drosophila* and Fish where CNE arrays often correspond to syntenic blocks (Engström et al., 2007; Dimitrieva and Bucher, 2013).

Next, we subdivided TADs according to their overlap with GRBs in GRB-TADs (> 80% overlap) and non-GRB-TADs (< 20% overlap) as in the original study (Harmston et al., 2017). As expected, we observed a higher accumulation of breakpoints at boundaries and stronger depletion within TADs for GRB-TADs compared to non-GRB-TADs (Fig 3.3B). However, also the non-GRB-TADs, that have less than 20% overlap with GRBs, are enriched for rearrangements at TAD boundaries. This indicates that not only TADs overlapping GRBs are evolutionary conserved. In summary, we show that human TADs overlapping clusters of non-coding conserved elements are strongly depleted for rearrangements, likely due to strong selective pressure on the conserved regulatory environment around important developmental genes.

Rearranged TADs are associated with divergent gene expression between species

The enrichment of rearrangement breakpoints at TAD boundaries indicates that TADs are stable across large evolutionary time scales. However, the reason for this

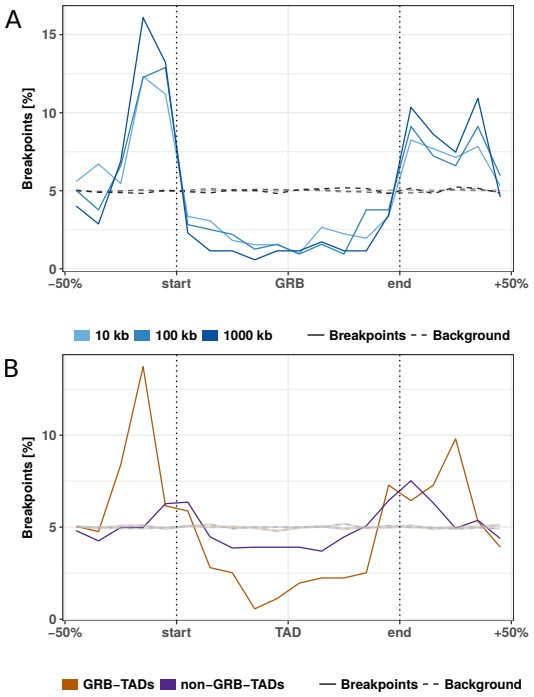

Figure 3.3.: **Rearrangement breakpoint distribution around GRBs and GRB-TADs. (A)** Rearrangement breakpoints between mouse and human around 816 GRBs. **(B)** Breakpoint distribution around GRB-TADs and non-GRB-TADs. GRB-TADs are defined as TADs overlapping more than 80% with GRBs and non-GRB-TADs have less than 20% overlap with GRBs. Breakpoints using a 10 kb fill size threshold are shown.

strong conservation of TAD regions is unclear. A mechanistic explanation could be that certain chromatin features at TAD boundaries promote or prevent DNA double strand breaks (DSBs) (Farré et al., 2015; Canela et al., 2017). Alternatively, selective pressure might act against the disruption of TADs due to their functional importance, for example in developmental gene regulation (Nora et al., 2013; Farré et al., 2015). TADs constitute a structural framework determining possible interactions between promoters and cis-regulatory sequences while prohibiting the influence of other sequences (Symmons et al., 2014; Lupiáñez et al., 2015). TAD disruption would prevent formerly established contacts. Rearrangements of TADs might also enable the recruitment of new cis-regulatory sequences which would alter the expression patterns of genes in rearranged TADs (Lupiáñez et al., 2015; Redin et al., 2017). Because of these detrimental effects, rearranged TADs should largely be eliminated by purifying selection. However, rearrangement of TADs could also enable the expression of genes in a new context and be selected if conferring an advantage. Therefore, we hypothesized that genes within conserved TADs might have a more stable gene expression pattern across tissues, whereas genes in rearranged TADs between two species might have a more divergent expression between species.

To test this, we analyzed the conservation of gene expression of ortholog genes between human and mouse across 19 matched tissues from the FANTOM5 project (Table S1) (Forrest et al., 2014). If a human gene and its mouse ortholog have high correlation across matching tissues, they are likely to have the same regulation and eventually similar functions. Conversely, low correlation of expression across tissues can indicate functional divergence during evolution, potentially due to altered gene regulation.

First, we separated human genes according to their location within TADs or outside of TADs. From 12,696 human genes with expression data and a unique one-to-one ortholog in mouse (Table S2), 1,525 have a transcription start site (TSS) located outside hESC TADs and 11,171 within. Next, we computed for each gene its expression correlation with mouse orthologs across 19 matching tissues. Genes within TADs have significantly higher expression correlation with their mouse ortholog (median R = 0,340) compared to genes outside TADs (mean R = 0,308, p = 0.0015, Fig 3.4A). This indicates higher conservation of gene regulation in TADs and is consistent with the observation of housekeeping genes at TAD boundaries (Dixon et al., 2012) and the role of TADs in providing conserved regulatory environments for gene regulation (Harmston et al., 2017; Ibn-Salem et al., 2017).

Next, we further subdivided TADs in two groups, rearranged and conserved, according to syntenic blocks and rearrangements between human and mouse genomes. In brief, a TAD is defined as conserved, if it is completely enclosed by a syntenic alignment block and does not overlap any rearrangement breakpoint. Conversely, a

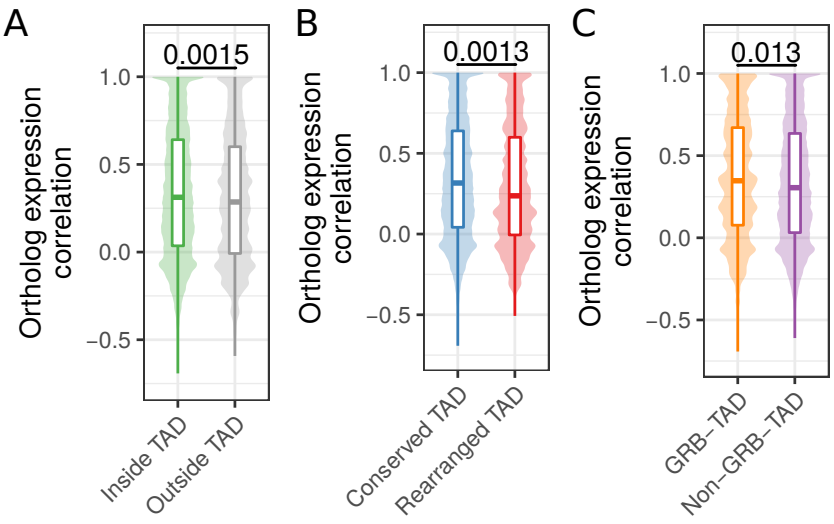

Figure 3.4.: **Ortholog gene expression correlation across tissues in conserved and rearranged TADs. (A)** Expression correlation of orthologs across 19 matching tissues in human and mouse for human genes within or outside of hESC TADs. **(B)** Expression correlation of orthologs across 19 matching tissues in human and mouse for genes in conserved or rearranged TADs. **(C)** Expression correlation of orthologs across 19 matching tissues in human and mouse for genes in GRB-TADs and non-GRB TADs. All P-values according to Wilcoxon rank-sum test.

rearranged TAD is not enclosed by a syntenic alignment block and overlaps at least one breakpoint that is farther than 80 kb from its boundary (see Methods). For the hESC TAD data set, this leads to 2,542 conserved and 137 rearranged TADs. The low number of rearranged TADs is consistent with the depletion of rearrangement breakpoints within TADs in general (Fig. 2). In total 8,740 genes in conserved and 645 genes in rearranged TADs could be assigned to a one-to-one ortholog in mouse and are contained in the expression data set. The expression correlation with mouse orthologs were significantly higher for genes in conserved TADs (median R = 0.316) compared to genes in rearranged TADs (median R = 0.237, p = 0.0013) (Fig 3.4B). This shows that disruptions of TADs by evolutionary rearrangements are associated with less conserved gene expression profiles across tissues. Although not significant, we also observed a slightly higher expression correlation for 1,003 genes in GRB-TADs compared to 8,038 genes in non-GRB TADs (Fig 3.4C, p = 0.13).

In summary, we observed higher expression correlation between orthologs for human genes inside TADs than outside. Moreover, we saw that genes in rearranged TADs show lower gene expression conservation than those in conserved TADs. These results not only support a functional role of TADs in gene regulation, but further support the hypothesis that TAD regions are subjected to purifying selection against their disruption by structural variations such as rearrangements.

Discussion

Our analysis of rearrangements between human and 12 diverse species shows that TADs are largely stable units of genomes, which are often reshuffled as a whole instead of disrupted by rearrangements. Furthermore, the decreased expression correlation with orthologs in mouse and human in rearranged TADs shows that disruptions of TADs are associated with changes in gene regulation over large evolutionary time scales.

TADs exert their influence on gene expression regulation by determining the set of possible interactions of cis-regulatory sequences with their target promoters (Nora et al., 2012; Symmons et al., 2014; Schoenfelder et al., 2015). This might facilitate the cooperation of several sequences that is often needed for the complex spatiotemporal regulation of transcription (Andrey and Mundlos, 2017). The disruption of these enclosed regulatory environments enables the recruitment of other cis-regulatory sequences and might prevent formerly established interactions (Montavon et al., 2012). The detrimental effects of such events have been shown in the study of diseases (Redin et al., 2017; Zepeda-Mendoza et al., 2017). There are also incidences where pathogenic phenotypes could be specifically attributed to enhancers establishing contacts to promoters that were formerly out of reach because of intervening TAD boundaries (Ibn-Salem et al., 2014; Lupiáñez et al., 2015; Spielmann et al., 2012). This would explain the selective pressure to maintain TAD integrity over large evolutionary distances and why we observe higher gene expression conservation for human genes within TADs compared to genes outside TADs.

Disruptions of TADs by large-scale rearrangements change expression patterns of orthologs across tissues and these changes might be explained by the altered regulatory environment which genes are exposed to after rearrangement (Farré et al., 2015).

Our results are largely consistent with the reported finding that many TADs correspond to clusters of conserved non-coding elements (GRBs) (Harmston et al., 2017). We observe a strong depletion of evolutionary rearrangements in GRBs and enrichment at GRB boundaries. This is consistent with comparative genome analysis revealing that GRBs largely overlap with micro-syntenic blocks in *Drosophila* (Engström et al., 2007) and fish genomes (Dimitrieva and Bucher, 2013). However, over 60% of human hESC TADs do not overlap GRBs (Harmston et al., 2017), raising the question of whether only a small subset of TADs are conserved. Interestingly, we find also depletion of rearrangements in non-GRB-TADs. This indicates that our rearrangement analysis identifies conservation also for TADs that are not enriched for CNEs. High expression correlation of orthologs in conserved TADs suggestss that the maintenance of expression regulation is important for most genes and probably even more crucial for developmental genes which are frequently found in GRBs.

Previous work using comparative Hi-C analysis in four mammals revealed that insulation of TAD boundaries is robustly conserved at syntenic regions, illustrating this with a few examples of rearrangements between mouse and dog genomes, which were located in both species at TAD boundaries (Vietri Rudan et al., 2015). The results of our analysis of thousands of rearrangements between human and 12 other species confirmed and expanded these earlier observations.

The reliable identification of evolutionary genomic rearrangements is difficult. Especially for non-coding genomic features like TAD boundaries, it is important to use approaches that are unbiased towards coding sequence. Previous studies identified rearrangements by interrupted adjacency of ortholog genes between two organisms (Vietri Rudan et al., 2015; Pevzner and Tesler, 2003). However, such an approach assumes equal inter-genic distances, which is violated at TAD boundaries, which have in general higher gene density (Dixon et al., 2012; Hou et al., 2012). To avoid this bias we used whole-genome-alignments. However, low quality of the genome assembly of some species might introduce alignment problems and potentially false positive rearrangement breakpoints.

Rearrangements are created by DNA double strand breaks (DSBs), which are not uniquely distributed in the genome. Certain genomic features, such as open chromatin, active transcription and certain histone marks are shown to be enriched at DSBs in somatic translocation sites (Roukos and Misteli, 2014) and evolutionary rearrangements (Murphy et al., 2005; Hinsch and Hannenhalli, 2006). Furthermore, induced DSBs and somatic translocation breakpoints are enriched at chromatin loop anchors (Canela et al., 2017). This opens the question of whether our finding of significantly enriched evolutionary rearrangement breakpoints at TAD boundaries could be explained by the molecular properties of the chromatin at TAD boundaries, rather than by the selective pressure to keep TAD function. Although, we cannot distinguish the two explanations entirely, our gene expression analysis indicates stronger conservation of gene expression in conserved TADs and more divergent expression patterns in rearranged TADs. This supports a model in which disruption of TADs are most often disadvantageous for an organism. Structural variations disrupting TADs can lead to miss regulation of neighboring genes as shown for genetic diseases (Ibn-Salem et al., 2014; Lupiáñez et al., 2015; Redin et al., 2017; Franke et al., 2016) and cancers (Hnisz et al., 2016; Northcott et al., 2014; Weischenfeldt et al., 2016).

Interestingly, we observed higher gene expression conservation for human genes within TADs compared to genes outside TADs. The larger syntenic structure of TADs might conserve the regulation likely by maintaining the proximity of promoters and cis-regulatory sequences while genes outside such frameworks are more exposed to changing genomic landscapes, presumably resulting in a greater susceptibility to the recruitment of regulatory sequences.

Apart from the described detrimental effects, our results suggest that TAD rearrangements occurred between genomes of human and mouse and led to changes in expression patterns of many orthologous genes. Since this is likely attributed to changing regulatory environments, it is also conceivable that some rearrangements led to a gain of function. Hence, TAD rearrangements might also provide a vehicle for evolutionary innovation. A single TAD reorganization has the potential to affect the regulation of a whole set of genes in contrast to the more confined consequences of other types of mutations (Acemel et al., 2017). Since it is also believed that changes in cis-regulatory sequences of developmental genes play a big part in evolutionary innovation (Carroll, 2008), the development of the enormous diversity of animal traits in evolution might have been promoted by the rearrangement of structural domains. This is consistent with a model in which new genes can arise by tandem-duplication and during evolution are then re-located to other environments (Ibn-Salem et al., 2017). These changes might have facilitated significant leaps in morphological evolution explaining the emergence of features that could not appear in small gradual steps. Following this hypothesis, TADs would not only constitute structural entities that perform the function of maintaining an enclosed regulatory landscape but could also be a driving force for change by exposing many genes at once to different genomic environments following single events of genomic rearrangement.

Conclusion

Our results indicate that TADs represent conserved functional building blocks of the genome. We have shown that the majority of evolutionary rearrangements do not affect the integrity of TADs and instead breakpoints are strongly clustered at TAD boundaries. This leads to the conclusion that TADs constitute conserved building blocks of the genome that are often reshuffled as a whole rather than disrupted during evolution. The conservation of TAD regions can be explained by detrimental effects of disrupting cis-regulatory environments that are essential for the spatio-temporal control of gene expression. Indeed we observe a significant association of conserved gene expression in intact TADs and divergent expression patterns in rearranged TADs explaining both why there could be selective pressure on the integrity of TADs over large evolutionary time scales, but also how TAD rearrangement can explain evolutionary leaps.

Methods

Rearrangement breakpoints from whole-genome alignments

Rearrangement breakpoints were identified between human and 12 selected vertebrate species from whole-genome-alignment data (Table 3.1). Alignment data were

downloaded as net files from UCSC Genome Browser for human genome hg38 and the genomes listed in Table 3.1. The whole-genome data consists of consecutive alignment blocks that are chained and hierarchically ordered in the so-called nets (Kent et al., 2003). Chains represent blocks of interrupted syntenic regions and may include larger gaps. When hierarchically arranged in a net file, child chains can complement their parents when they align nearby segments that fill the alignment gaps of their parents but may also break the synteny when incorporating distal segments. We implemented a computer program to extract rearrangement breakpoints from net files based on the length and type of fills. Start and end points of top-level or non-syntenic fills are reported as rearrangement breakpoint if the fill exceeds a given size threshold. We used different size thresholds to optimize both the number of identified breakpoints and to avoid biases of transposable elements that might be responsible for many small interruptions of alignment chains. In this way, we extracted rearrangement breakpoints between human and 12 genomes using size thresholds of 10 kb, 100 kb, and 1000 kb. To compare breakpoints to TADs we converted the breakpoint coordinates from hg38 to hg19 genome assembly using the liftOver tool from UCSC Genome Browser (Hinrichs et al., 2006).

Table 3.1.: Species used for breakpoint identification from whole-genome alignments with human.

Common name	Species	Genome Assembly	Divergence to human (mya)
Chimpanzee	*Pan troglodytes*	panTro5	6.65
Gorilla	*Gorilla gorilla gorilla*	gorGor5	9.06
Orangutan	*Pongo abelii*	ponAbe2	15.76
Rhesus	*Macaca mulatta*	rheMac8	29.44
Mouse lemur	*Microcebus murinus*	micMur2	74
Mouse	*Mus musculus*	mm10	90
Cattle	*Bos taurus*	bosTau8	96
Manatee	*Trichechus manatus latirostris*	triMan1	105
Opossum	*Monodelphis domestica*	monDom5	159
Chicken	*Gallus gallus*	galGal5	312
Clawed frog	*Xenopus tropicalis*	xenTro7	352
Zebrafish	*Danio rerio*	danRer10	435

Topologically associating domains and contact domains

We obtained topologically associating domain (TAD) calls from published Hi-C experiments in human embryonic stem cells (hESC) (Dixon et al., 2012) and contact domains from published *in situ* Hi-C experiments in human GM12878 cells (Rao et al., 2014). Genomic coordinates of hESC TADs were converted from hg18 to hg19 genome assembly using the UCSC liftOver tool (Hinrichs et al., 2006).

Breakpoint distributions at TADs

To quantify the number of breakpoints around TADs and TAD boundaries we enlarged TAD regions by 50% of their total length on each side. The range was then subdivided into 20 equal sized bins and the number of overlapping breakpoints computed. This results in a matrix in which rows represent individual TADs and columns represent bins along TAD regions. The sum of each column indicates the number of breakpoints for corresponding bins and therefore the same relative location around TADs. For comparable visualization between different data sets, the column-wise summed breakpoint counts were further normalized as percent values of the total breakpoint number in the matrix.

Quantification of breakpoint enrichment

To quantify the enrichment of breakpoints at domain boundaries, we generated random breakpoints as background control. For each chromosome, we placed the same number of actual breakpoints at a random position of the chromosome. For each breakpoint data set we simulated 100 times the same number of random breakpoints. We then computed the distribution of random breakpoints around TADs in the same way as described above for actual breakpoints. To compute enrichment of actual breakpoints compared to simulated controls, we classified each breakpoint located in a window of 400 kb around TAD borders in either close to a TAD boundary, if distance between breakpoint and TAD boundary was smaller or equal to 40 kb or as distant, when distance was larger than 40 kb. This results in a contingency table of actual and random breakpoints that are either close or distal to TAD boundaries. We computed log odds ratios as effect size of enrichment and p-values according to Fishers two-sided exact test. Additionally, we compared the distance of all actual and random breakpoints to their nearest TAD boundary using the Wilcoxon's rank-sum test.

Expression data for mouse and human orthologs

Promoter based expression data from CAGE analysis in human and mouse tissues from the FANTOM5 project (Forrest et al., 2014) were retrieved from the EBI Expression Atlas (Hinrichs et al., 2006) as baseline expression values per gene and tissue. The meta data of samples contains tissue annotations as term IDs from Uberon, an integrated cross-species ontology covering anatomical structures in animals (Herrero et al., 2016). Human and mouse samples were assigned to each other if they had the same developmental stage and matching Uberon term IDs. This resulted in 19 samples for each organism with corresponding tissues.

We used the R package biomaRt to retrieve all human genes in the Ensembl database (version grch37.ensembl.org) and could assign 13,065 to ortholog genes in mouse by allowing only the one-to-one orthology type (Herrero et al., 2016). Of

these ortholog pairs, 12,696 are contained in the expression data described above. For each pair of orthologs we computed the correlation of expression values across matching tissues as Pearson's correlation coefficient.

Classification of TADs and genes according to rearrangements and GRBs

We classified hESC TADs according to rearrangements between human and mouse genomes. We define a TAD as conserved if it is completely enclosed within a fill in the net file and no rearrangement breakpoint from any size threshold is located in the TAD region with a distance larger than 80 kb from the TAD boundary. A TAD is defined as rearranged, if the TAD is not enclosed completely by any fill in the net file, overlaps at least one breakpoint inferred using a 1000 kb fill size threshold, and this breakpoint is further than 80 kb away from each TAD boundary. TADs were also classified according to their overlap with GRBs as in (Harmston et al., 2017). A given TAD is a GRB-TAD if it overlaps with more than 80% of the TAD size with a GRB. A TAD is classified as non-GRB if it has less than 20% overlap with GRBs. The 12,696 human genes with mouse ortholog and expression data were grouped according to their location with respect to hESC TADs. We used the transcription start site (TSS) of the longest transcript per gene to group each gene as within TAD if the TSS overlaps a hESC TAD or as outside TADs, if not. Furthermore, we grouped genes in TADs according to conserved or rearranged TADs and separately according to GRB and non-GRB TADs.

Source code and implementation details

The source code of the entire analysis described here is available on GitHub: *https://github.com/Juppen/TAD-Evolution*. The identification of breakpoints and extraction of fills from whole-genome alignment data was implemented in Python scripts. Reading of BED files and overlap calculations with TADs and TAD bins were computed in R with Bioconductor (Huber et al., 2015) packages rtracklayer (Lawrence et al., 2009) and GenomicRanges (Lawrence et al., 2013). Gene coordinates and ortholog assignments were retrieved from Ensemble data base (version grch37.ensembl.org) using the package biomaRt (Durinck et al., 2009). For data integration and visualization we used R packages from tidyverse (Wickham and Grolemund).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The source code of all analysis is available on GitHub: *https://github.com/Juppen/TAD-Evolution*. All the genomic data used for analyses are freely available to be downloaded from the UCSC Genome Browser and EBI Expression Atlas with identifiers listed in Table 3.1 and Table S1.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors‘ contributions

JK and JI developed and implemented the methods and performed the analysis. JI conceived the study. JK wrote the first draft of the manuscript. JK, MA and JI wrote the manuscript. MA supervised the study.

Acknowledgments

The authors thank all members of the CBDM group for fruitful discussions.

4

Position effects of rearrangements in disease genomes

Preamble

This chapter was published as a co-first-author paper in the American Journal of Human Genetics:

Cinthya J. Zepeda-Mendoza[*], **Jonas Ibn-Salem**[*], Tammy Kammin, David J. Harris, Debra Rita, Karen W. Gripp, Jennifer J. MacKenzie, Andrea Gropman, Brett Graham, Ranad Shaheen, Fowzan S. Alkuraya, Campbell K. Brasington, Edward J. Spence, Diane Masser-Frye, Lynne M. Bird, Erica Spiegel, Rebecca L. Sparkes, Zehra Ordulu, Michael E. Talkowski, Miguel A. Andrade-Navarro, Peter N. Robinson, Cynthia C. Morton[#]. ***Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements***. Am J Hum Genet. 2017;101(2):206-217. doi:10.1016/j.ajhg.2017.06.011.

The publication is availible online: https://doi.org/10.1016/j.ajhg.2017.06.011. My contributions to this publication is indicated in Table E.1. The source code of the complete analysis is available at GitHub: https://github.com/ibn-salem/position_effect. Supplementary information is shown in Appendix C.

[*]These authors contributed equally to this work
[#]corresponding author

Abstract

Interpretation of variants of uncertain significance, especially chromosome rearrangements in non-coding regions of the human genome, remains one of the biggest challenges in modern molecular diagnosis. To improve our understanding and interpretation of such variants, we used high-resolution 3-dimensional chromosome structure data and transcriptional regulatory information to predict position effects and their association with pathogenic phenotypes in 17 subjects with apparently balanced chromosome abnormalities. We find that the rearrangements predict disruption of long-range chromatin interactions between several enhancers and genes whose annotated clinical features are strongly associated with the subjects' phe-

notypes. We confirm gene expression changes for a couple of candidate genes to exemplify the utility of our position effect analysis. These results highlight the important interplay between chromosome structure and disease, and demonstrate the need to utilize chromatin conformation data for the prediction of position effects in the clinical interpretation of cases of non-coding chromosome rearrangements.

Introduction

The importance of the integrity of chromosome structure and its association with human disease is one of the oldest and most studied topics in clinical genetics. As early as 1959, cytogenetic studies in humans linked specific genetic or genomic disorders and intellectual disability syndromes to changes in chromosomal ploidy, translocations, and DNA duplications and deletions (LeJeune et al., 1959; Ford et al., 1959; Jacobs and Strong, 1959; Stankiewicz and Lupski, 2002; Iafrate et al., 2004). The discovery of copy-number variants (CNVs) by microarray and sequencing technologies expanded the catalogue of genetic variation between individuals to test such associations at higher resolution (Iafrate et al., 2004; Sebat et al., 2004; Hinds et al., 2006; Conrad et al., 2006, 2010; Korbel et al., 2007; Stankiewicz and Lupski, 2010; International HapMap 3 Consortium et al., 2010; Carvalho and Lupski, 2016).

Over the years, analysis of disease-related structural rearrangements has illuminated genes that are mutated in various human developmental disorders (Zhang et al., 2009; Theisen and Shaffer, 2010; Nambiar and Raghavan, 2011; Higgins et al., 2008). Such chromosome aberrations can directly disrupt gene sequences, affect gene dosage, generate gene fusions, unmask recessive alleles, reveal imprinted genes, or result in alterations of gene expression through additional mechanisms such as position effects (Zhang et al., 2009). The latter is particularly important for the study of apparently balanced chromosome abnormalities (BCAs), such as translocations and inversions, often found outside of the hypothesized disease-causing genes (reviewed in (Kleinjan and Van Heyningen, 2005)).

Position effects were first identified in *Drosophila melanogaster*, where chromosomal inversions placing *white+* near centric heterochromatin caused mosaic red/white eye patterns.(Weiler and Wakimoto, 1995) In humans, BCAs can induce position effects through disruption of a gene's long-range transcriptional control (*i.e.*, enhancer-promoter interactions, insulator influence, etc.), or its placement in regions with different local chromatin environments as observed in the classical *Drosophila* position effect variegation (reviewed in (Kleinjan and Van Heyningen, 2005; Zhang and Wolynes, 2015; Spielmann and Mundlos, 2016)). Examples of position effect genes include paired box gene 6 (*PAX6* [MIM: 607108]), for which downstream chromosome translocations affect its *cis*-regulatory control and produce aniridia (AN [MIM: 106210]);(Fantes et al., 1995; Kleinjan et al., 2001) twist family bHLH transcription factor 1 (*TWIST1* [MIM: 601622]), where downstream

translocations and inversions are associated with Saethre-Chotzen syndrome (SCS [MIM: 101400]);(Cai et al., 2003) paired like homeodomain 2 (*PITX2* [MIM: 601542]) for which translocations are associated with Axenfeld-Rieger syndrome type 1 (RIEG1 [MIM: 180500]);(Flomen et al., 1998; Trembath et al., 2004) SRY-box 9 (*SOX9* [MIM: 608160]), where translocation breakpoints located up to 900 Kilobases (Kb) upstream and 1.3 Megabases (Mb) downstream are associated with campomelic dysplasia (CMPD [MIM: 114290]),(Velagaleti et al., 2005) in addition to several others.(Kleinjan and Van Heyningen, 2005; Kleinjan and van Heyningen, 1998; Lupski and Stankiewicz, 2005)

The availability of genome sequencing in the clinical setting has generated a need for rapid prediction and interpretation of structural variants, especially those pertaining to *de novo* non-coding rearrangements in individual subjects. With the development and subsequent branching of the chromosome conformation capture (3C) technique ((Dekker et al., 2002), reviewed in (de Wit and de Laat, 2012)), regulatory issues such as alteration of long-range transcriptional control and position effects can now be predicted in terms of chromosome organization. The high resolution view of chromosome architecture in diverse human cell lines and tissues(Lieberman-Aiden et al., 2009; Fullwood et al., 2009; Dixon et al., 2012; Sanyal et al., 2012; Phillips-Cremins et al., 2013; Rao et al., 2014; Mifsud et al., 2015; Dixon et al., 2015) has allowed molecular assessment of the disruption of regulatory chromatin contacts by pathogenic structural variants and single nucleotide changes; examples include the study of limb malformations,(Lupiáñez et al., 2015) leukemia,(Gröschel et al., 2014) and obesity,(Claussnitzer et al., 2015) among others.(Visser et al., 2012; Roussos et al., 2014; Giorgio et al., 2015; Oldridge et al., 2015; Ibn-Salem et al., 2014; Ordulu et al., 2016) These examples underscore the importance of chromatin interactions in quantitative and temporal control of gene expression, which can greatly enhance our power to predict pathologic consequences.

To test the feasibility of prediction and clinical interpretation of position effects of non-coding chromosome rearrangements, we analyzed 17 subjects from the Developmental Gene Anatomy Project (DGAP)(Higgins et al., 2008; Ligon et al., 2005; Kim and Marcotte, 2008; Lu et al., 2007; Redin et al., 2017) with *de novo* non-coding BCAs classified as variants of uncertain significance (VUS). Using publicly available chromatin contact information, annotated and predicted regulatory elements, and correlation between phenotypes observed in DGAP subjects and those associated with neighboring genes, we reliably predicted candidate genes exhibiting mis-regulated expression in DGAP-derived lymphoblastoid cell lines (LCLs). These results suggest that many VUS are likely to be further interpretable via long-range effects, and warrant their routine assessment and integration in clinical diagnosis.

Materials and Methods

Selection of subjects with apparently balanced chromosome abnormalities

BCA breakpoints and clinical data were obtained from DGAP cases for which whole-genome sequencing was performed using a previously described large-insert jumping library approach.(Higgins et al., 2008; Ligon et al., 2005; Kim and Marcotte, 2008; Lu et al., 2007; Redin et al., 2017; Talkowski et al., 2011) A total of 151 cases were filtered to select only subjects whose translocation or inversion breakpoints fall within intergenic regions (GRCh37) and did not overlap known long intergenic non-coding RNAs (lincRNAs) or pseudogenes, as these elements have been shown to exert functional roles (reviewed in (Quinn and Chang, 2016) (Pink et al., 2011; Muro and Andrade-Navarro, 2010)). Of 151 DGAP subjects, only 17 fulfilled our selection criteria, 12 of whom had available and reportedly normal clinical array results, suggesting lack of large duplications or deletions.

Clinical descriptions of DGAP cases

The clinical presentation of the 17 subjects varied, ranging from developmental delay to neurological conditions, offering the opportunity to assess long-range position effects in different phenotypes. Subjects' karyotypes are presented in the main text using the International System for Human Cytogenetic Nomenclature (ISCN2016) (Table 4.1). Detailed case descriptions are included in the Supplemental Note: Case Reports, as well as a nomenclature developed to describe chromosome rearrangements using next-generation sequencing.(Ordulu et al., 2014) Reported ages of DGAP subjects are from time of enrollment. All reported genomic coordinates use GRCh37.

Subject ID and Reported Karyotype	Disruption of Functional Elements	Breakpoints within TADs (hESC / IMR90 / GM12878)	Top-ranking Candidates +/-1 Mb

Table 4.1.: Description of the 17 analyzed DGAP cases with non-coding BCAs. Corresponding clinical karyotypes are reported, with overlap of breakpoints with regulatory elements (E = enhancer, DHS = DNaseI hypersensitive sites, CTCF = CTCF binding sites), and TADs from H1-hESC, IMR90, and GM12878 (1= one breakpoint within TAD, 2=both BCA breakpoints are located within TAD). Top-ranking position effect genes are provided for the ś1 Mb windows surrounding the BCA breakpoints; each gene is highlighted with different evidence supporting its inclusion (a = ClinGen known recessive genes, b= ClinGen genes with emerging and sufficient evidence suggesting haploinsufficiency is associated with clinical phenotype, c = HI scores less than 10, d = within H1-ESC TAD, e = DHS enhancer-promoter disrupted interactions).

Subject ID and Reported Karyotype	Disruption of Functional Elements	Breakpoints within TADs (hESC / IMR90 / GM12878)	Top-ranking Candidates +/-1 Mb
DGAP017	DHS	2/2/1	-
46,X,t(X;10)(p11.2;q24.3)			
DGAP111	CTCF	1/1/2	*ORC6*[a]
46,XY,t(16;20)(q11.2;q13.2)dn			
DGAP113	-	2/2/2	*ASPM*[a]
*46,XY,t(1;3)(q32.1;q13.2)dn			
DGAP126	-	2/1/2	-
46,XX,t(5;10)(p13.3;q21.1)dn			
DGAP138	-	2/2/2	*GRIK2*[ac]
46,XY,t(1;6)(q23;q13)dn			
DGAP153	-	1/1/1	-
46,X,t(X;17)(p11.23;p11.2)dn			
DGAP163	-	2/2/2	*SOS1*[cde],
46,XY,t(2;14)(p23;q13)dn			*COCH*[de]
DGAP176	DHS, CTCF	2/1/2	*ACSL4*[bd],
46,Y,inv(X)(q13q24)mat			*COL4A5*[bcde]
DGAP249	E, DHS	2/2/2	*SATB2*[bcde],
46,XX,t(2;11)(q33;q23)dn			*SORL1*[e]
DGAP252	-	2/2/2	*RBBP8*[a],*GATA6*[bcde]
46,XY,t(3;18)(q13.2;q11.2)dn			
DGAP275	DHS	1/1/2	*ANKLE2*[e],
46,XX,t(7;12)(p13;q24.33)dn			*POLE*[e]
DGAP287	CTCF	2/2/2	-
46,XY,t(10;14)(p13;q32.1)dn			
DGAP288	DHS	2/2/2	*SOX9*[bcd]
46,XX,t(6:17)(q13;q21)dn			

Subject ID and Reported Karyotype	Disruption of Functional Elements	Breakpoints within TADs (hESC / IMR90 / GM12878)	Top-ranking Candidates +/-1 Mb
DGAP315 *46,XX,inv(6)(p24q11)dn*	-	1/1/2	-
DGAP319 *46,XX,t(4;13)(q31.3;q14.3)dn*	-	2/1/2	-
DGAP322 *46,XY,t(1;18)(q32.1;q22.1)*	DHS	1/2/2	*IRF6*[bcd]
DGAP329 *46,XX,t(2;14)(q21;q24.3)dn*	-	1/2/2	*ZEB2*[bcde]

Analysis of genes bordering the rearrangement breakpoints

The presence of annotated genes or pseudogenes and lincRNAs was assessed in windows of ś3 and ś1 Mb neighboring each subject's translocation and inversion breakpoints, and within reported H1-hESC topologically associated domains (TADs)(Dixon et al., 2012) where the breakpoints were located. The gene annotation file was obtained from Ensembl GRCh37 archive,(Flicek et al., 2014) and we used the Human Body Map lincRNAs catalog.(Cabili et al., 2011) Haploinsufficiency (HI) and triplosensitivity scores were assigned using Huang *et al.*, 2010(Huang et al., 2010) and version hg19 of ClinGen(Rehm et al., 2015) data downloaded on 9/20/2016.

Assessment of disrupted functional elements and chromatin interactions bordering rearrangement breakpoints

The disruption of regulatory elements such as enhancers, promoters, locus control regions, and insulators can lead to disease-related gene expression changes; DNase I hypersensitive (DHS) sites have been used as markers for the identification of such elements.(Thurman et al., 2012) In addition, the alteration of TAD boundaries has been previously shown to cause a rewiring of enhancers with pathological consequences;(Lupiáñez et al., 2015; Giorgio et al., 2015; Narendra et al., 2015) CCCTC-Binding Factor (CTCF) binding sites have been found to be enriched in TAD boundaries,(Dixon et al., 2012) and several mutations of boundary-defining sites have been associated with cancer.(Flavahan et al., 2016; Hnisz et al., 2016) Based on these observations, we assessed the number of regulatory elements that were potentially disrupted by the analyzed DGAP breakpoints. We compared the breakpoint positions of the selected DGAP subjects against data corresponding to CTCF binding sites, DHS sites, and chromatin segmentation classifications (Broad ChromHMM) derived from a lymphoblastoid cell line (GM12878) and

human stem cells (H1-hESC), obtained from the Encyclopedia of DNA Elements (ENCODE) project(Dunham et al., 2012) and accessed through the University of California Santa Cruz Genome Browser.(Kent et al., 2002) Enhancer positions were additionally obtained from Andersson *et al.*, 2014(Andersson et al., 2014) for tissue and primary cells, and the VISTA Enhancer browser, human version hg19.(Visel et al., 2007) Finally, lists of transcription factor (TF) binding sites and gene promoters were obtained from the Ensembl database human version GRCh37.(Flicek et al., 2014) Hi-C interaction data and TAD positions for H1-hESC, GM06990, and IMR90 at 20 Kb, 40 Kb, 100 Kb, and 1 Mb resolution were obtained from Dixon *et al.*, 2012(Dixon et al., 2012) and the WashU EpiGenome Browser.(Zhou and Wang, 2012) A high-resolution dataset of chromatin loops and domains was obtained from Rao *et al.*, 2014 for IMR90 and GM12878 cells.(Rao et al., 2014) Lastly, distal DHS/enhancer–promoter connections(Thurman et al., 2012) were used to assess disrupted predicted cis-regulatory interactions by the BCAs. Genomic overlaps between the rearrangement breakpoints, functional elements and disrupted chromatin interactions were calculated using custom Perl scripts, the BEDtools suite(Quinlan and Hall, 2010) and the genomic association tester (GAT) tool.(Heger et al., 2013)

Ontological analysis of genes neighboring breakpoints

Phenotype similarity between potential position effect genes and DGAP cases was calculated by converting the phenotypes of the 17 subjects to Human Phenotype Ontology (HPO)(Köhler et al., 2014) terms and calculating their phenomatch score as described in Ibn-Salem *et al.*, 2014.(Ibn-Salem et al., 2014) The phenomatch score quanties the information content of the most specic HPO term that is part of or a common ancestor (more general term) of a set of phenotypes. Our set of phenotypes is constituted by the HPO terms associated to DGAP cases and the ones annotated to candidate position effect genes within windows of ś3 and ś1 Mb of sequence in proximity to the breakpoints. We used two background models to assess signicance of this similarity. The rst is based on randomly permuting the associations of phenotypes to genes; to this effect, the phenotype-gene associations are shuffled 100 times randomly and the similarity of these random phenotypes to the studied case clinical findings is calculated. The second background control is based on shifting the breakpoint location along the chromosome; each breakpoint is shifted by -9, -6, -3, +3, +6, and +9 Mb and the similarity of genes in proximity to the shifted breakpoints is computed.

Quantitative real-time PCR

LCLs derived from DGAP236-02m, DGAP244-02m and DGAP245-02m were used as karyotypically normal male controls. These are karyotypically normal fathers of enrolled DGAP cases with no history of disease. LCL 17402 (DGAP163) was

used to test differential gene expression for SOS Ras/Rac guanine nucleotide exchange factor 1 (*SOS1* [MIM: 182530]), and LCL 18060 (DGAP176) was used to test midline 2 (*MID2* [MIM: 300204]), p21 (RAC1) activated kinase 3 (*PAK3* [MIM: 300142]), and POU class 3 homeobox 4 (*POU3F4* [MIM: 300039]) expression using quantitative polymerase chain reaction (qPCR). Glucuronidase beta (*GUSB* [MIM: 611499]) was used as a housekeeping control. qPCR experiments were performed by the Harvard Biopolymers Facility using TaqMan probes Hs00264887_s1 (*POU3F4*), Hs00201978_m1 (*MID2*), Hs00176828_m1 (*PAK3*), Hs00893134_m1 (*SOS1*), and Hs00939627_m1 (*GUSB*). Data were analyzed using the CT method.

Assessment of DGAP breakpoints overlapping with non-coding structural variants in public databases

To find similar non-coding structural rearrangement subjects and compare their annotated clinical phenotypes to those observed in DGAP cases, we searched the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER)(Firth et al., 2009) version 2015-07-13, as well as the dbVar database from the National Center for Biotechnology Information (NCBI) Variation Viewer 1.5.(Lappalainen et al., 2012) Both databases are comprehensive community-supported repositories of clinical cases with novel and extremely rare genomic variants.

Results

Genomic characterization of non-coding breakpoints

To study the structural and evolutionary context of BCAs and their impact on nuclear architecture and gene expression, we used data generated by DGAP,(Higgins et al., 2008; Ligon et al., 2005; Kim and Marcotte, 2008; Lu et al., 2007; Redin et al., 2017) the largest collection of sequenced balanced chromosome rearrangements from individuals with abnormal developmental and cognitive phenotypes, many of which have yet to be investigated in detail. Each studied DGAP BCA has two breakpoint positions (as two distinct chromosome regions are involved in their generation), which we labeled with the DGAP#_A and DGAP#_B identifiers. We filtered DGAP data to select cases with both breakpoints in non-coding regions only, and excluding lincRNAs and pseudogenes; a total of 17 cases fulfilled our criteria, 15 translocations and 2 inversions (Figure 4.1 and Table S1). These subjects are phenotypically distinct, and most of them presented with congenital developmental and neurological conditions not recognized as a known syndrome or genomic disorder (see clinical descriptions in Supplemental Note: Case Reports).

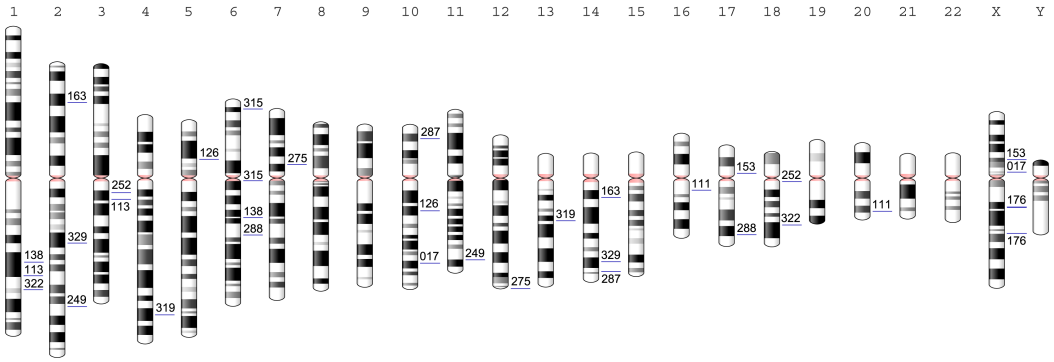

Figure 4.1.: **Chromosome locations of the 17 analyzed DGAP cases with non-coding BCAs.** Breakpoint positions are marked with a blue line and the corresponding DGAP number. All chromosomes are aligned by the centromere (marked in pink) and are indicated above by their corresponding chromosome number.

Further analysis revealed that BCA breakpoints were significantly depleted for overlapping annotated promoters or transcription factor (TF) binding sites (GAT TF p=0.0003, promoter p=0.0001, Table S2,3). Only one breakpoint (DGAP249_B) overlapped a ChromHMM enhancer in GM12878 cells (Table 4.1); the others had no overlap with annotated or predicted enhancers in the analyzed datasets, and this depletion was significant for VISTA (GAT p=0.0364) and Hi-ESC (GAT p=0.0036) but not for the annotated tissue and primary cell enhancers from Andersson *et al.*, 2014(Andersson et al., 2014) (Table S4). Eight breakpoints overlapped cell-type specific DHS sites (Table 4.1 and Table S5); these corresponded to DGAP cases 017, 176, 249, 275, 288 and 322; of these, DGAP176 and DGAP275 overlapped DHS sites at both BCA breakpoint sites. In addition, three DGAP cases overlapped CTCF binding sites in H1-hESC (DGAP cases 111, 176, and 287) and none in GM12878 cells (Table 4.1 and Table S6). Except for two cases in H1-hESC (DGAP17 and DGAP176), and four cases in GM12878 (DGAP 017, 126, 163 and 176), all rearrangements fall within ChromHMM repressed chromatin regions, but this association was not significant (GAT p= 0.40 for GM12878 and p= 0.15 for H1-hESC, Table S2F). Interestingly, 22 of the 34 breakpoints (~65%) overlap repeated elements at a significant level (GAT p=0.0002, Table S8), which may indicate a non-allelic homologous recombination process in their generation.(Gu et al., 2008; Cardoso et al., 2016)

Noticeably, either one or two breakpoints from all the non-coding DGAP BCAs fall within previously reported TADs in H1-hESC and IMR90 cell lines (Table 4.1 and Table S9).(Dixon et al., 2012) However, this overlap was not significant for both cell lines (GAT H1-ESC p=0.0537 and IMR90 p=0.28). We found that the breakpoints disrupt dozens, hundreds, or even thousands of chromatin contacts when assessed at the 20 and 40 Kb resolution in Hi-C data of H1-hESC and IMR90 cells, as well as chromatin contacts at 100 Kb and 1 Mb resolution in GM06990 cells

(Table S11). Breakpoint DGAP111_A had a consistent absence of disrupted chromatin contacts, which is expected as it overlaps a repetitive satellite region so no chromatin contacts could be mapped to the segment (Table S9 and Table S11). With the availability of higher resolution data, it is possible to detect whether BCA breakpoints disrupt smaller chromatin domains and loops not detected in previous studies. When analyzing high resolution IMR90 and GM12878 Hi-C data,(Rao et al., 2014) we discovered that 32 out of 34 breakpoints are contained within GM12878 sub-compartments (Table 4.1 and Table S10); interestingly, 28 of these are classified as members of the B compartment, which is less gene dense and less expressed compared to the A compartment. On the other hand, 18 and 24 breakpoints are contained within GM12878 and IMR90 arrowhead domains, respectively (Table S10), which are regions of enhanced contact frequency that tile the diagonal of each chromatin contact matrix. In addition, the breakpoints disrupt several significant short and long-range chromatin interactions in the GM12878 Hi-C data (Table S12).

Overall, the observation of breakpoint-associated DHS sites suggests the alteration of underlying regulatory elements with potential pathogenic outcomes, while the predicted extensive disruption of chromatin contacts and the alteration of TAD boundaries by the BCAs may affect long-range regulatory interactions of neighboring genes (see Discussion).

Identification of genes with potential position effects

To identify genes which could be generating the complex DGAP phenotypes via position effects from chromosome rearrangements, we analyzed all annotated genes within windows of ś3 and ś1 Mb proximal and distal to the breakpoints, and within the BCA-containing H1-hESC reported TAD positions. A total of 3081 genes were contained within the ś3 and ś1 Mb windows for all cases; 106 of these genes (~3.4%) have an HI score of <10%, which is a predictor of haploinsufficiency,(Huang et al., 2010) and 55 and two genes have ClinGen emerging evidence suggesting that dosage haplo/triplo-sensitivity, respectively, is associated with clinical phenotype (Table S15).

To further refine our search for genes which may exhibit position effects, we performed an unbiased correlation between DGAP case phenotypes and the clinical traits associated with genes bordering each breakpoint. To this end, we used the HPO dataset,(Köhler et al., 2014) which provides a standardized vocabulary of phenotypic abnormalities encountered in human disease, and currently contains ~11,000 terms and over 115,000 annotations to hereditary diseases. We translated DGAP clinical features to HPO terms (Table S16), and calculated phenotype similarity between DGAP cases and neighboring genes using the phenomatch score.(Ibn-Salem et al., 2014) The phenomatch score distinguishes between general and very specific phenotypic descriptions by quantifying the information content of the most

specic HPO terms that are common to, or a common ancestor of, the DGAP case and neighboring gene phenotypes. The similarity significance is then calculated based on randomly permuting the associations of phenotypes to genes, and in shifting the DGAP translocation and inversion breakpoint positions along the chromosome. We obtained phenomatch scores ranging from 0.003 to 91.48 for 179 genes within the ś3 and ś1 Mb windows, as well as within the TAD positions (Table S15).

In addition to dosage sensitivity and phenotypic similarity information, we complemented our analysis with assessment of enhancer-promoter interactions to make our candidate selection more specific. A typical mechanism by which chromosome rearrangements cause position effects is through disruptions in the association of genes with their regulatory regions.(Kleinjan and Van Heyningen, 2005; Kleinjan and van Heyningen, 1998) We therefore reasoned that genes and enhancers included in predicted enhancer-promoter interactions would be strong position effect candidates. We used the ENCODE distal DHS/enhancer–promoter connections(Thurman et al., 2012) to assess disrupted predicted *cis*-regulatory interactions by the DGAP breakpoints within a ś500 Kb window. The analysis revealed 193 genes that were separated from their predicted candidate enhancers, potentially altering gene expression (Table S13). A total of 133 candidate genes were separated from <10 of their predicted enhancers, while 60 genes were separated from their predicted interactions with 10 or up to 91 enhancers (Table S14).

For the 17 analyzed DGAP BCAs, there are a total of 645 genes with either evidence of dosage sensitivity, disrupted enhancer-promoter interactions, or significant phenotypic similarity. This represents ∼21% of the genes contained within the ś3 Mb windows, clearly an undesirable number for timely clinical interpretation and functional analyses. To filter the most promising candidates, we ranked them using their reported dosage sensitivity, disrupted regulatory interactions, and by selecting a phenomatch cut-off value capable of detecting pathogenic and likely pathogenic genes in 57 published DGAP cases from Redin *et al.*, 2017.(Redin et al., 2017) By taking into consideration the top quartile values of the reported phenomatch scores per case and adding up their dosage sensitivity and disrupted regulatory interaction data, we consistently ranked the reported pathogenic and likely pathogenic genes in the upper decile for 52 out of the 57 control DGAP cases (∼91%) when considering candidates within the TAD and ś1 Mb analysis windows (Table S17). 32 of these genes were the top-ranking candidates in their corresponding DGAP case, while 19 of them were positioned in the second-tier rank. Only five genes could not be found in the top decile ranking positions as they had one or no lines of evidence supporting their inclusion.

Applying this ranking strategy to the 17 non-coding BCAs, we predict 16 top-ranking candidates for 11 DGAP cases and 102 second-tier candidates for the 17 analyzed DGAP cases within ś1 Mb analysis windows (Table 4.1 and Table S15).

This is a significant reduction compared to the initial 645 possible candidates (~3.8% of the neighboring genes in the ś3 Mb windows considering top and second-tier candidates, and only 0.05% considering top candidates only). Of note, only nine of the 16 top-ranking candidates are included within the same TAD as the BCA breakpoint (H1-hESC TADs from (Dixon et al., 2012)), while the rest are located farther away. Nine top-ranking genes had an HI score <10%,(Huang et al., 2010) while ClinGen HI data revealed that four of these 16 genes are associated with autosomal recessive phenotypes, and an additional seven have sufficient or some evidence for haploinsufficiency. Only one candidate gene for DGAP138, glutamate ionotropic receptor kainate type subunit 2 (*GRIK2* [MIM: 138244]) was a confirmed triplosensitive annotated gene in ClinGen (Table S15).

Taken together, these cases represent more plausible candidates in the search for position effect genes with functional consequences in the subjects' phenotypes. Examples include *GRIK2* which could explain the intellectual disability observed in DGAP138; *SOS1*, forkhead box G1 (*FOXG1* [MIM: 164874]) and cochlin (*COCH* [MIM: 603196]) may be related to the neurological and developmental delay as well as hearing loss of DGAP163; acyl-CoA synthetase long-chain family member 4 (*ACSL4* [MIM: 300157]) and *POU3F4* could be involved in DGAP176's cognitive impairment and hearing loss; SATB homeobox 2 (*SATB2* [MIM: 608148]) may underlie the delayed speech and language development observed in DGAP249; RB binding protein 8 endonuclease (*RBBP8* [MIM: 604124]) may be involved in DGAP252's craniofacial dysmorphic features; *SOX9* most likely explains the cleft palate observed in DGAP288; DNA polymerase epsilon catalytic subunit (*POLE* [MIM: 174762]) may contribute to the extreme short stature observed in DGAP275, and zinc finger E-box binding homeobox 2 (*ZEB2* [MIM: 605802]) can potentially explain the hypotonia and neurological features observed in DGAP329. *SOX9* had been previously proposed to explain DGAP288's phenotype, and as predicted by our method, a decrease in its expression was observed in RNA derived from DGAP288's umbilical cord blood.(Ordulu et al., 2016) Additional quantitative real-time PCR analyses revealed *SOS1* as having reduced expression in DGAP163-derived LCLs compared to three normal sex-matched controls (Figure 4.2).Expression assessment for second-tier candidates *PAK3*, *MID2* and *POU3F4* in DGAP176 LCLs did not deviate substantially from their control expression values (Figure C.1); further searches into the Genotype-Tissue Expression (GTEx) project(Lonsdale et al., 2013) reveal that *PAK3*, *MID2* and *POU3F4* have low expression in LCLs, which would have made assessing changes in expression of these genes technically difficult. This points to the importance of the availability of tissues and cell lines relevant to the studied phenotypes, or the capacity to generate animal models that reproduce the observed BCAs for further analysis.

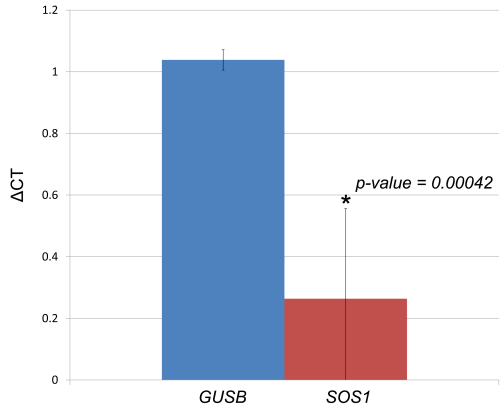

Figure 4.2.: Assessment of gene expression changes for DGAP163-derived LCLs. Each column represents the CT results of three culture replicates, with four technical replicates each, compared to three sex-matched control cell lines. Error bars indicate the standard deviation calculated from the biological replicates. The Mann-Whitney U test p-value is provided for the comparison between expression values of *SOS1* and the control *GUSB*.

Identification of subjects with shared non-coding chromosome alterations and phenotypes

The identification of subjects with shared non-coding chromosome alterations and phenotypes as described herein would further support our idea of these rearrangements exerting their pathogenic outcomes through long-range position effects. To identify such subjects, we searched the DECIPHER(Firth et al., 2009) and dbVar databases,(Lappalainen et al., 2012) both comprehensive community-supported repositories of clinical cases with novel or extremely rare genomic variants.

We found 494 DECIPHER cases overlapping our 34 non-coding BCA breakpoints (Table S19). Of these, 489 had rearrangements that overlapped one or more annotated genes (Table S20). Only five DECIPHER cases fulfilled our non-coding selection criteria (Table S21): cases 1985 and 1989, both of which overlap one of DGAP017's breakpoints in chromosome 10, but which have several other gene-altering genomic rearrangements; case 289720, a subject with a 161.44 Kb deletion in chromosome 10 described as likely benign and sharing a sequence breakpoint with DGAP126; case 289865 overlapping a breakpoint in DGAP126 in chromosome 10, very similar to case 289720, however with the presence of an additional pathogenic gene-altering rearrangement; and lastly case 293610, a pathogenic duplication of 364.43 Kb in chromosome 17 sharing a breakpoint with DGAP288. Only two of the five DECIPHER cases have reported clinical phenotypes. DECIPHER case 289720 presents with intellectual disability and psychosis, both pertaining to the superclasses of behavioral and neurodevelopmental abnormalities under the HPO classification. Interestingly, DGAP126 has abnormal aggressive, impulsive or violent behavior and

auto-aggression, as well as language and motor delays, which also fall under the classification of behavioral and neurodevelopmental abnormalities. DECIPHER case 293610 has reported gonadal tissue discordant for external genitalia or chromosomal sex as well as a non-obstructive azoospermia clinical phenotype;(Vetro et al., 2015) both features are not observed until puberty, and are associated with the female-to-male sex disorder observed for CNVs altering the *SOX9* genomic landscape. Although DGAP288 is still an infant, there is no report of sex reversal.

From the dbVar database, 675 non-coding structural rearrangements including CNVs, deletions, inversions, and translocations overlap DGAP breakpoints (Table S22). Of these, only five variants had associated clinical information, including variant nsv534336, a 530 Kb duplication overlapping the DGAP017 breakpoint in chromosome 10, classified as "uncertain significance"(Miller et al., 2010) and exhibiting a growth delay phenotype; nsv931775, a benign ~381.8 Kb deletion overlapping the DGAP113 breakpoint on chromosome 3, associated with developmental delay and/or other significant developmental or morphological phenotypes;(Miller et al., 2010) nsv534571, an ~639.7 Kb duplication of uncertain significance associated with muscular hypotonia and overlapping the DGAP287 breakpoint on chromosome 10; and variants nsv532026 and nsv917014, two duplications of ~613 Kb classified as "uncertain significance" and "likely benign," respectively, overlapping the DGAP315 breakpoint in chromosome 6, and associated with developmental delay and/or other significant developmental or morphological phenotypes as well as autism and global developmental delay. All the detected variants are associated with phenotypes observed in the DGAP cases, especially DGAP017's hypoplasia, the developmental delay observed in DGAP113, and DGAP315's significant developmental or morphological phenotypes.

Strictly speaking, these phenotypes are disparate, but fall under similar phenotypic categories, which could enable identification of long-range effect genes between different cases with similar clinical features and chromosome rearrangements. These comparisons highlight the importance of establishing detailed, specific, and unbiased guidelines for assigning phenotypes when performing computational phenotype comparisons.

Discussion

Structural variation of the human genome, either inherited or arising by *de novo* germline or somatic mutations, can give rise to different phenotypes through several mechanisms. Chromosome rearrangements can alter gene dosage, promote gene fusions, unmask recessive alleles, or disrupt associations between genes and their regulatory elements. The traditional clinical focus of studying genes disrupted by chromosome rearrangements has shifted to also assess regions neighboring these variants.(Ordulu et al., 2016) This search for positional effects has been particularly

important in the analysis of chromosome rearrangements associated with different clinical conditions and disrupting non-annotated genomic regions.(Zhang and Wolynes, 2015; Spielmann and Mundlos, 2016)

The study of chromatin conformation has been requisite in the analysis of such non-coding rearrangements. DNA is organized in the three-dimensional nucleus at varying hierarchical levels that are important for the regulation of gene expression,(de Wit and de Laat, 2012) with primary roles in embryonic development and disease.(Bonev and Cavalli, 2016) Several studies have analyzed the impact of structural variants in disruption of the regulatory chromatin environment leading to disease;(Lupiáñez et al., 2015; Gröschel et al., 2014; Visser et al., 2012; Roussos et al., 2014; Giorgio et al., 2015; Ibn-Salem et al., 2014) these studies have set the precedent for integrative analyses of disrupted chromatin conformation to expedite functional annotations of non-coding chromosome rearrangements.

We tested the possibility of utilizing chromatin contact information to dissect chromosome rearrangements which disrupt non-coding chromosome regions in clinical cases. We focused on 17 subjects from DGAP, 12 with available clinical microarray information, with different rare presentations and *de novo* non-coding BCAs classified as VUS. Of these, 15 corresponded to translocations and two were inversions. These cases represent ~11% of the total number of sequenced DGAP cases, which makes our predictions even more significant for future potential treatment or management of subjects who would not otherwise obtain a clinical diagnosis. Utilizing publicly available annotated genomic and regulatory elements, chromatin conformation capture information, predicted enhancer-promoter interactions, phenomatch scores, as well as haploinsufficiency and triplosensitivity information for all genes surrounding the BCA breakpoints at different window sizes (ś3 and ś1 Mb as well as BCA-containing TAD positions), we discovered 16 genes for 11 DGAP cases that are top-ranking position effect candidates for the subjects' clinical phenotypes (Table 4.1).

We observed that eight of the sequenced DGAP BCA breakpoints, corresponding to six DGAP cases (DGAP017, 176, 249, 275, 288 and 322), overlapped reported annotated and predicted enhancers and DHS sites. Disruption of these regulatory elements could potentially cause improper gene expression or repression through altered enhancer-promoter interactions or interactions with other DHS-associated elements such as insulators and locus control regions, among others. In fact, four of the breakpoints that disrupt annotated DHS sites and enhancers have been shown to establish chromatin contacts with our top position effect candidate genes in the region in Hi-C data of H1-hESC cells at 40 Kb resolution (Table S18). For example, the DGAP275_B breakpoint is involved in a chromatin interaction that puts it into physical proximity with *POLE* and *ANKLE2,* DGAP288_B contacts *SOX9*, and DGAP176_B interacts with *ACSL4*. Three additional breakpoints from DGAP111,

249 and 287 overlap CTCF binding sites. CTCF binding sites are enriched in TAD boundaries,(Dixon et al., 2012) and the elimination of these binding sites could potentially induce gene expression or other functional changes through alteration of the structural regulatory landscape of the region.(Lupiáñez et al., 2015)

There are nine DGAP cases (DGAP113, 126, 138, 153, 163, 252, 315, 319 and 329), six with normal arrays and two with benign CNVs, for which no overlap with genomic or other regulatory elements was detected. These cases thus represent events in which position effects are most likely caused by alteration of the underlying chromatin structure itself. This hypothesis is supported by detection of a vast number of disrupted chromatin contacts in four different cell lines (H1-hESC, IMR90, GM06990, GM12878) at different Hi-C window resolutions, 32 breakpoints included in H1-hESC TADs,(Dixon et al., 2012) and the separation of 193 genes from one and up to 91 of their predicted enhancers after the occurrence of the BCAs (Table S14). For example, *SOS1*, one of the most significant candidates in explaining DGAP163's global developmental delay, dysmorphic/distinctive facies and hearing loss, as observed in Noonan Syndrome 1 (NS1 [MIM: 163950]), is separated from its interaction with 88 predicted enhancers (Figure 4.3), and exhibited a decrease in expression in DGAP163-derived LCLs. However, NS1 is caused by autosomal dominant mutations in *SOS1*; we hypothesize that the reduced expression of *SOS1* might affect the RAS/MAPK signaling pathway and generate clinical features not completely overlapping those of NS1; however, this possibility remains to be functionally tested and complimented with analyses of genomic single nucleotide variants. A similar approach could be explored for DGAP275, where we hypothesize that *POLE*, associated with the facial dysmorphism, immunodeficiency, livedo, and short stature syndrome (FILS [MIM: 615139]) in an autosomal recessive manner,(Pachlopnik Schmid et al., 2012) may contribute to the extreme short stature observed in this DGAP subject; and *ZEB2*, etiologic for Mowat-Wilson syndrome (MOWS [MIM: 235730]) in an autosomal dominant manner (OMIM#235730), may potentially explain the hypotonia and neurological features observed in DGAP329 but not present all of the dysmorphic features or medical/non-neurologic phenotype of MOWS. Overall, more candidate genes will need to be analyzed rigorously to assess the validity of our position effect predictions and the disruption of important chromatin regulatory elements. Nonetheless, insight into the molecular pathway of disorders may be forthcoming from our approach and of value in the management of some individuals.

All predicted candidate genes have different lines of evidence supporting their selection, starting with a significant phenomatch score that correlates annotated gene phenotypes to those observed in the DGAP cases. HI and triplosensitivity evidence, inclusion in TAD regions, as well as HI scores build upon this selection, and can help laboratories and clinicians focus in subsequent analyses on candidates of their interest. As of now, the "top-ranking" candidates have the highest number of evidence

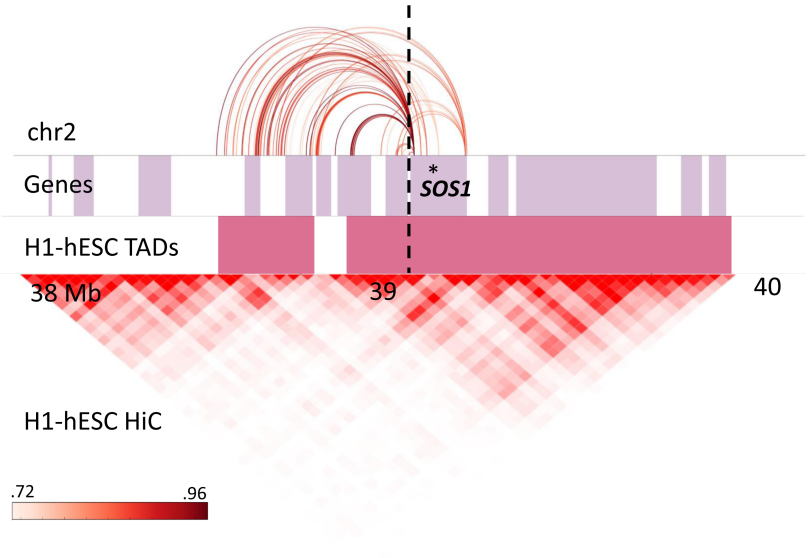

Figure 4.3.: **Disrupted enhancer-promoter DHS interactions predicted for *SOS1*** (gene position indicated by asterisk). The color graded rectangle represents the correlation values for the interactions as reported by ENCODE. The dashed line indicates the translocation breakpoint position in chromosome 2. Lilac colored rectangles represent genes, and pink rectangles show TAD positions annotated in H1-hESC.

supporting their selection; however, there are also 102 second-tier candidates for the 17 analyzed DGAP cases within ś1 Mb analysis windows which may well play a functional role. Presently, we are unable to give "weights" to any of these selection criteria (*i.e.*, a gene with a high phenomatch score and no evidence of HI is "more significant" than a gene with a medium phenomatch score and evidence of HI) mainly for two reasons: (i) we would need to collect more examples, which might not be easy to find and require a tremendous curation effort, and (ii) we need to understand the possibility, suggested by our results, that more than one gene may be contributory in the clinical presentation of the DGAP subjects, either acting simultaneously or throughout development. Moreover, many of the candidates have recessive inheritance modes, which make it necessary to assess the mutational status of both alleles as well as additional sequence variants not captured by our BCA breakpoint sequencing and the microarrays. Future in-depth exome, DNA and RNA sequencing as well as Hi-C experiments will provide a comprehensive view of the contribution of sequence variants, disruption of chromatin contacts, and changes in gene expression in the DGAP disease etiologies, such that guidelines might be developed as to which candidates should be followed up first and further studied with comprehensive functional validation using animal models and human cell lines that reproduce the BCA breakpoints.

Overall our results suggest that the integration of phenomatch scores, altered chromatin contacts, and other clinical gene annotations provide valuable interpretation

to many variants of uncertain significance through long-range position effects. The correct prediction of 52 out of 57 known pathogenic genes in DGAP cases used as positive controls supports such integration. Our computational analysis is rapid and can provide additional information to benefit the clinical assessment of both coding and non-coding genome variants. The latter is an important step towards prediction of pathogenic consequences of non-coding variantion observed in prenatal samples. For example, based on its position and chromatin contact alterations, we correctly predicted the involvement and decreased expression of *SOX9* in the cleft palate Pierre-Robin sequence (PRBNS [MIM: 261800]) association in DGAP288.(Ordulu et al., 2016)

Lastly, we would like to note that predicting the pathogenic outcome of disrupted chromatin contacts is not a straightforward endeavor: it has been shown that a single gene promoter can be targeted by several enhancers,(Thurman et al., 2012) therefore compensating for the perturbed interactions by the chromosome rearrangements. In addition, rearrangements can reposition gene promoters and enhancers outside of their preferred chromatin environments, leading to improper gene activation by enhancer adoption.(Lupiáñez et al., 2015) Our method currently identifies instances in which known and predicted enhancer/promoter interactions are disrupted by the rearrangement breakpoints and thus lead to decreased candidate gene expression. Enhancer adoption prediction will be incorporated once mathematical models of TAD formation upon changes in genomic sequence are refined and available to the greater scientific community. Presently, our predictions are as good as the availability of pathogenic gene annotations, chromatin conformation data, clinical phenotype information, and the presence of similar rearrangements in databases such as DECIPHER and dbVar. While the existence of other subjects with related phenotypes to the DGAP cases does not prove the involvement of neighboring genes in the etiology of these phenotypes, it is a step forward towards prediction of pathogenic effects starting from a simple computational analysis, pointing to a better phenotypic categorization when clinically examining affected individuals. By making our position effect prediction method available to the human genetics community, we hope to study additional cases with complete phenotypic information and be able to refine better the rules for the prediction of position effects on gene expression and discover new mechanisms of pathogenicity.

Acknowledgements

We offer heartfelt gratitude to all DGAP research participants and their families, and to countless genetic counselors, clinical geneticists, cytogeneticists, and physicians for their ongoing support of our study and for referrals to our project. This study was funded by the National Institutes of Health (GM061354 to CCM and MET). The authors declare no conflicts of interest.

Web Resources

The scripts used in this study to predict position effects can be downloaded from: *https://github.com/ibn-salem/position_effect*

OMIM, *http://www.omim.org*

Ensembl GRCh37 archive, *http://grch37.ensembl.org*

Human lincRNAs catalog, *http://portals.broadinstitute.org/genome_bio/human_lincrnas*

Haploinsufficiency scores, *https://decipher.sanger.ac.uk*

ClinGen GRCh37 data, *ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/clingen*

University of California Santa Cruz Genome Browser, *https://genome.ucsc.edu*

Human Phenotype Ontology, *http://human-phenotype-ontology.github.io*

Harvard Biopolymers Facility, *https://genome.med.harvard.edu*

dbVar Variation Viewer, *https://www.ncbi.nlm.nih.gov/variation/view*

3D Genome Browser, *http://promoter.bx.psu.edu/hi-c*

ENCODE, *https://www.encodeproject.org*

WashU EpiGenome Browser, *http://epigenomegateway.wustl.edu/*

GTEx portal, *https://www.gtexportal.org/home*

5 Prediction of chromatin looping interactions

Preamble

This chapter is submitted as a first-author paper to Genome Biology. A preprint is published at bioRxiv:

Ibn-Salem J, Andrade-Navarro MA. Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs. bioRxiv. 2018. doi:10.1101/257584.

The preprint is availible online: https://www.biorxiv.org/content/early/2018/02/01/257584. My contributions to this publication is indicated in Table E.1. The source code of the complete analysis is available at GitHub: https://github.com/Juppen/sevenC and https://github.com/Juppen/sevenC_analysis. Supplementary figures and links to supplementary tables are shown in Appendix D.

Abstract

Background: Knowledge of the three-dimensional structure of the genome is necessary to understand how gene expression is regulated. Recent experimental techniques such as Hi-C or ChIA-PET measure long-range interactions genome-wide but are experimentally elaborate and have limited resolution. Here, we present Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs (7C).

Results: While ChIP-seq was not designed to detect contacts, the formaldehyde treatment in the ChIP-seq protocol cross-links proteins with each other and with DNA. Consequently, also regions that are not directly bound by the targeted TF but interact with the binding site via chromatin looping are co-immunoprecipitated and sequenced. This produces minor ChIP-seq signals at loop anchor regions close to the directly bound site. We use the position and shape of ChIP-seq signals around CTCF motif pairs to predict whether they interact or not.

We applied 7C to all CTCF motif pairs within 1 MB in the human genome and validated predicted interactions with high-resolution Hi-C and ChIA-PET. A single ChIP-seq experiment from known architectural proteins (CTCF, Rad21, Znf143) but also from other TFs (like TRIM22 or RUNX3) predicts loops accurately. Importantly, 7C predicts loops in cell types and for TF ChIP-seq datasets not used in training.

Conclusion: 7C predicts chromatin loops with base-pair resolution and can be used to associate TF binding sites to regulated genes in a condition-specific manner. Furthermore, profiling of hundreds of ChIP-seq datasets results in novel candidate factors functionally involved in chromatin looping. Our method is available as an R package: https://ibn-salem.github.io/sevenC/

Keywords

Chromatin interactions, three-dimensional genome architecture, transcription factors, ChIP-seq, 3C, 4C, 5C, Hi-C, 6C, ChIA-PET, 7C, prediction, chromatin loops

Introduction

The three-dimensional folding structure of the genome and its dynamic changes play a very important role in the regulation of gene expression (Merkenschlager and Nora, 2016; Krijger and de Laat, 2016). For example, while it was well known that transcription factors (TFs) can regulate genes by binding to their adjacent promoters, many TF binding sites are in distal regulatory regions, such as enhancers, that are hundreds of kilo bases far from gene promoters (Spitz and Furlong, 2012). These distal regulatory regions can physically interact with promoters of regulated genes by chromatin looping interactions (Tolhuis et al., 2002; Sanyal et al., 2012), thus it is not trivial to associate TFs to regulated genes without information of the genome structure (Mora et al., 2015). Such looping interactions can be measured by chromosome conformation capture (3C) experiments (Dekker et al., 2002) and its variations to either study all interactions from single targeted regions (4C) (Simonis et al., 2006) or multiple target regions (5C) (Dostie et al., 2006), interactions between all regions genome-wide (Hi-C) (Lieberman-Aiden et al., 2009; Rao et al., 2014) or interactions mediated by specific proteins (6C (Tiwari et al., 2008) and ChIA-PET (Fullwood et al., 2009; Tang et al., 2015)).

While these experimental methods have brought many exciting insights into the three-dimensional organization of genomes (Merkenschlager and Nora, 2016; Krijger and de Laat, 2016; Bonev and Cavalli, 2016), these methods are not only elaborate and expensive but also require large amounts of sample material or have limited resolution (Sati and Cavalli, 2016; Schmitt et al., 2016). As a consequence, genome-wide chromatin interaction maps are only available for a limited number of cell types and conditions.

In contrast, the binding sites of TFs can be detected genome-wide by ChIP-seq experiments, and are available for hundreds of TFs in many cell types and conditions (Dunham et al., 2012; Davis et al., 2017). Here, we propose that it is possible to use these data to detect chromatin loops.

Recent studies provide functional insights about how chromatin loops are formed and highlight the role of architectural proteins such as CTCF and cohesin (Merkenschlager and Nora, 2016). CTCF recognizes a specific sequence motif, to which it binds with high affinity (Kim et al., 2007; Nagy et al., 2016). Interestingly, CTCF motifs are present in convergent orientation at chromatin loop anchors (Rao et al., 2014; Tang et al., 2015; Vietri Rudan et al., 2015). Furthermore, experimental inversion of the motif results in changes of loop formation and altered gene expression (Guo et al., 2015; de Wit et al., 2015). Polymer simulations and experimental perturbations led to a model of loop extrusion, in which loop-extruding factors, such as cohesin, form progressively larger loops but stall at CTCF binding sites in convergent orientation (Sanborn et al., 2015; Fudenberg et al., 2016). According to these models, CTCF binding sites can function as anchors of chromatin loops.

Our hypothesis is, that we can use convergently aligned CTCF motifs to search for similar ChIP-seq signals at both sites of chromatin loops to predict looping interactions from the largely available ChIP-seq data in many diverse cell-types and conditions (Fig. 5.1A). We then developed and tested a computational method to predict chromatin looping interactions from only genomic sequence features and TF binding data from ChIP-seq experiments. We show that our method has high prediction performance when compared to Hi-C and ChIA-PET loops and that prediction performance depends on the ChIP-seq target, which allows screening for TFs with potential novel functions in chromatin loop formation. The predicted looping interactions can be used to (i) associate TF binding sites or enhancers to regulated genes for conditions where Hi-C like data is not available, and (ii) to increase the resolution of interaction maps, where low resolution Hi-C data is available. We implemented our method as the R package *sevenC*.

Results

CTCF motif pairs as candidate chromatin loop anchors

In order to predict chromatin looping interactions from ChIP-seq data, we first analyzed which features at looping anchors correlate with interaction signals. As a starting point for all analyses we used 38,316 CTCF motif sites in the human genome as potential chromatin loop anchors. We built a dataset of all CTCF motif pairs located within a genomic distance of 1 Mb to each other. This resulted in 717,137 potential looping interactions; we expect that only a minority of these motif pairs will be in physical contact for a given cell type and condition. To label motif pairs as true loops, we used chromatin loops from published high-resolution *in-situ* Hi-C data and ChIA-PET data for CTCF and Pol2 in human GM12878 cells (Rao et al., 2014; Tang et al., 2015). If a motif pair was measured to interact in one

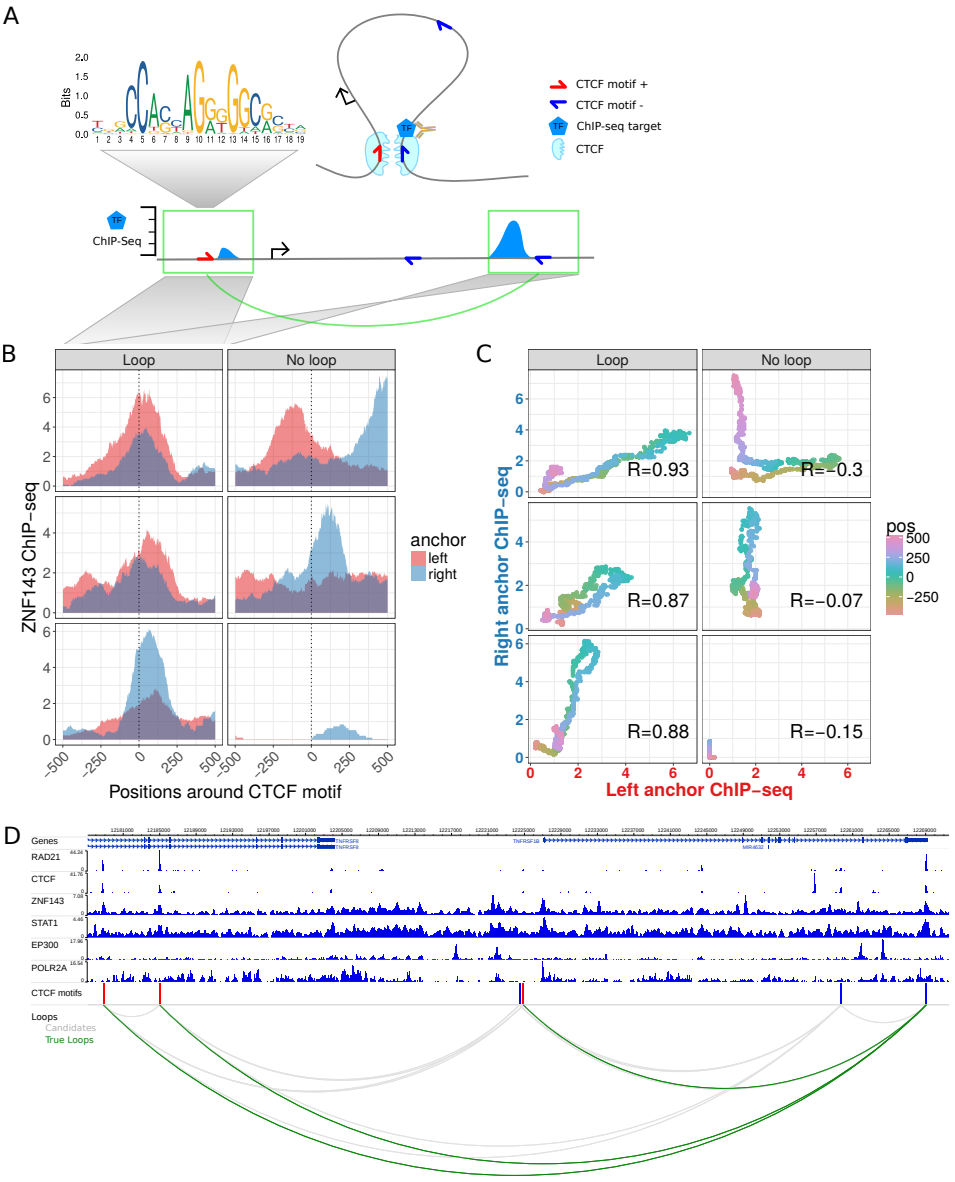

Figure 5.1.: **Chromatin looping interactions result in ChIP-seq coverage signals at direct and indirect bound loop anchors. (A)** Schematic illustration of a chromatin loop with CTCF motifs at the loop anchors (top right). A TF binds directly at the right loop anchor close to the CTCF motif. This results in a ChIP-seq coverage peak at the directly bound locus (bottom right) and in a minor signal at the other loop anchor (bottom left), both at the same distance to each CTCF motif. **(B)** Znf143 ChIP-seq coverage at six selected example CTCF motif pairs of which the ones in the left panel interact via loops according to Hi-C and ChIA-PET data and the ones in the right panel do not interact. The ChIP-seq coverage signal for each loci pair is shown in red for the left anchor region and in blue for the right anchor region, according to the distance to the CTCF motif (x-axis). Interacting CTCF motif pairs show more similar ChIP-seq coverage signals, which are often enriched at similar distances to the CTCF motif pairs, while the profiles of non-interacting pairs are less similar. **(C)** The similarity of ChIP-seq profiles by correlation of the ChIP-seq coverage signals of the selected motif pairs in (B). For each pair, the coverage at the right anchor is plotted versus the coverage at the left anchor at the same distance (color coded) from each CTCF motif. The Pearson correlation coefficient (R) of the dots is higher for interacting loci pairs. **(D)** Example loci on chromosome 1 shown in the genome-browser with six ChIP-seq tracks. Red and blue bars indicate CTCF recognition motifs on the forward and reverse strand, respectively. The bottom panel shows CTCF motif pairs in gray (candidates) and actually interacting pairs in green, according to ChIA-PET and Hi-C data.

of the data sets, we labeled it as true interaction. Overall 30,025 (4.19 %) of CTCF motif pairs were considered as true loops using these data sets.

Similarity of ChIP-seq signals at looping CTCF motifs

The ChIP-seq protocol involves a cross-linking step, in which formaldehyde treatment results in covalent bonds between DNA and proteins (Orlando et al., 1997). This allows the pull-down and detection of sites directly bound by the targeted protein. However, cross-linking occurs also between proteins, which results in detection of sites that are indirectly bound through protein-protein interactions or chromatin looping interactions (Hoffman et al., 2015; Starick et al., 2015).

We hypothesized that if a protein binds directly to a genomic region in chromatin contact with other genomic regions, DNA from both loci might be pulled out in the cross-linking and DNA-purification step of ChIP-seq protocols. As a result, we expect ChIP-seq signals (e.g. mapped reads) at both genomic regions: the directly bound one and the chromatin loop interaction partner locus (Fig. 5.1A).

To test this hypothesis, we used CTCF motif pairs as anchors and compared the ChIP-seq signal from one anchor to the (reversed) signal of the corresponding anchor. We found similar ChIP-seq coverage patterns around CTCF motifs more often when the two sites perform looping interactions than when they do not (Fig. 5.1B). To quantify the similarity of ChIP-seq coverage from any two CTCF sites, we correlated their ChIP-seq signals at +/- 500 bp around the CTCF motif (Fig. 5.1C) (see Methods for details). Measuring ChIP-seq profile similarity by correlation has the advantage that the correlation can be high even if the anchor that is not bound directly has a much lower ChIP-seq signal (which is often the case).

Next, we compared ChIP-seq similarity at looping and non-looping CTCF motif pairs for six selected TF ChIP-seq data sets (Fig. 5.1D). Compared to non-interacting CTCF sites the ChIP-seq correlation is significantly higher at looping interactions (Fig. 5.2A). However, the overall correlation as well as the difference between looping and non-looping CTCF sites varies between TF ChIP-seq datasets (Fig. 5.2A). As expected, we observed a large difference for the CTCF ChIP-seq dataset but, interestingly, also for other known architectural proteins, such as Rad21 and Znf143. Moreover, other TFs, such as STAT1 have significantly higher ChIP-seq signal similarity at CTCF motifs that interact via chromatin looping. Overall, this analysis shows that ChIP-seq signals are more similar at interacting CTCF sites, indicating that this similarity can be used to predict looping interactions.

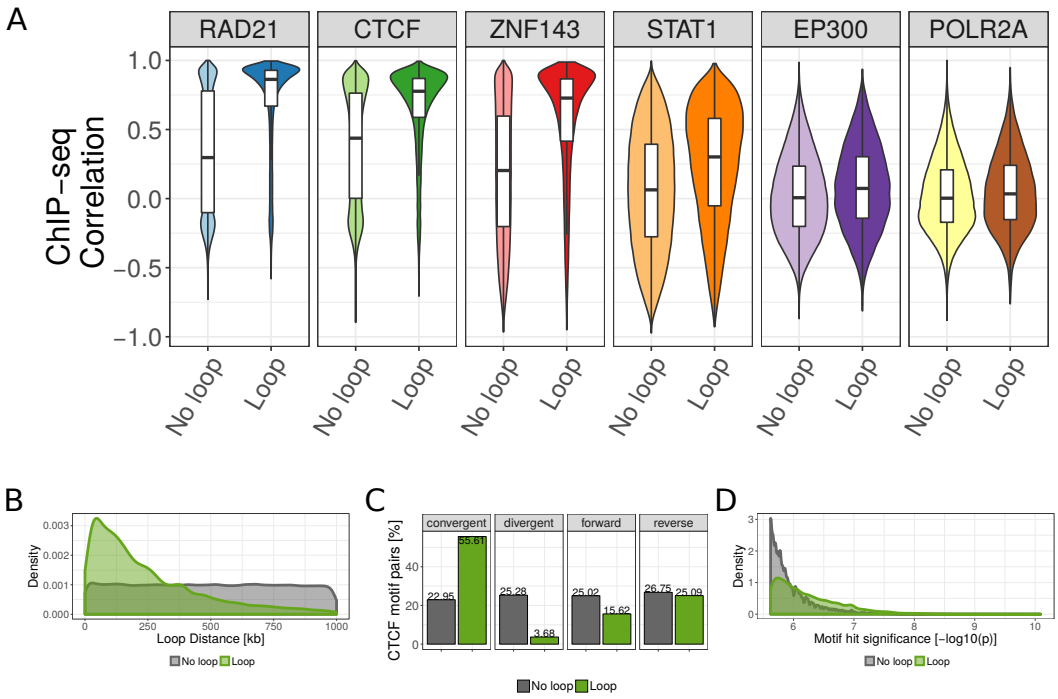

Figure 5.2.: ChIP-seq similarity and genomic features of looping and non-looping CTCF motif pairs. **(A)** Boxplot of Pearson correlation coefficient of ChIP-seq signals between CTCF motif pairs for all CTCF motif pairs within 1 Mb genome-wide. The correlation is shown separately for non-looping and looping motif pairs (according to HI-C and ChIA-PET data in GM12878 cells), and for six selected ChIP-seq data sets in GM12878 cells. **(B)** Distance distribution between looping (green) and non-looping CTCF motif pairs. **(C)** Number of looping and non-looping CTCF motif pairs in convergent, divergent, both forward, or both reverse orientation. **(D)** Distribution of CTCF motif hit significance as $-\log_{10}$ transformed p-value for looping and non-looping CTCF motif pairs. For each motif pair only the less significant motif is considered.

Genomic sequence features of CTCF motif pairs are associated with looping

The frequency of two genomic regions to physically interact depends on their genomic distance (Lieberman-Aiden et al., 2009). Consequently, we observed that CTCF motif pairs are more often in contact when they are close to each other in the genomic sequence (Fig. 5.2B). Recent studies on 3D chromatin structure led to an increased understanding of the molecular mechanism of chromatin loop formation and suggested a functional role of CTCF proteins, which bind specific DNA sequences (Merkenschlager and Nora, 2016). The canonical CTCF motif is non-palindromic and therefore occurs either in the positive or in the negative DNA strand. Importantly, it is known that CTCF motifs occur predominantly in convergent orientation to each other at chromatin loop anchors (Rao et al., 2014; Vietri Rudan et al., 2015). Experimental inversions of CTCF motifs lead to changes of the interactions and expression of the associated genes (Guo et al., 2015; de Wit et al., 2015). Accordingly, we observed that 55.6% of the looping CTCF pairs have convergent orientation versus only 25.3% of the non-looping pairs (Fig. 5.2C). We also observed that the motif match strength, as measured by the significance of a motif location to match the canonical CTCF motif (Khan et al., 2018), is higher for motifs involved in looping interactions (Fig. 5.2D). Together, the linear genome encodes several features, such as motif strength, orientation, and distance, that correlate with chromatin looping and can be used to predict such interactions.

Chromatin loop prediction using 7C

To make use of both the condition specific ChIP-seq signals and the genomic features of CTCF motifs to predict chromatin loops, we trained a prediction model that takes only ChIP-seq data as input. To this end, we built a logistic regression model that takes into account only four features: the correlation coefficient between the ChIP-seq signals of the paired CTCF motifs (in a window of 1000 bp around the motif), the genomic distance between motifs, the orientation, and the (minimum) motif hit significance score (see Methods for details). For each ChIP-seq data set, we trained and evaluated a separate model. The method is implemented as the R package 'sevenC', which predicts chromatin loops using as only input a bigWig file from a ChIP-seq experiment.

Prediction performance evaluation

We used 10-fold cross-validation to assess the performance of the predictions on independent data that was not seen in the training phase. For each cutoff on the predicted interaction probability score, we computed the sensitivity, specificity, precision and recall to plot receiver operator characteristic (ROC) and precision recall curves (PRC). Since only 4.2% of CTCF pairs are measured to interact, we mainly

used the area under the PRC (auPRC) to evaluate prediction performance since, compared to ROC, the PRC gives a more accurate classification performance in imbalanced datasets in which the number of negatives outweighs the number of positives significantly (Saito and Rehmsmeier, 2015). Furthermore, we defined an optimal cutoff for the prediction probability p based on optimizing the f1-score. The six selected TF ChIP-seq data sets have optimal f1-scores at about $p = 0.15$ (Figure S1B). For binary prediction, we provide a default prediction score threshold as the average of thresholds with optimal f1-score for the 10 best performing TF ChIP-seq datasets.

Prediction performance of 7C with sequence features and single TF ChIP-seq data sets

First, we evaluated how the sequence-encoded features can predict chromatin interactions. For this, we built regression models that use only these features. Each of these features alone, CTCF motif hit significance, motif orientation or distance, were very poor predictors, and resulted in auROC between 0.67 and 0.74 (Fig. 5.3A) and auPRC scores between 0.08 and 0.09 (Fig. 5.3C). Using the three sequence features together improved prediction performance (auROC = 0.85, auPRC = 0.22).

Next, we tested the addition of ChIP-seq data as feature in the prediction model using ChIP-seq data for each of six different TFs. Three of them, CTCF, RAD21, and ZNF143, have known function in chromatin loop formation (Merkenschlager and Nora, 2016; Ye et al., 2016; Bailey et al., 2015), while STAT1, P300, and POL2, are to our knowledge not directly involved in chromatin loop formation (Fig. 5.1D). Adding any of these TF ChIP-seq datasets to the model increased prediction performance. STAT1, EP300, and POL2 only moderately increased prediction performance with auROC values between 0.86 and 0.87 (Fig. 5.3A) and auPRC between 0.24 and 0.26 (Fig. 5.3B, C). However, ChIP-seq of the known architectural proteins CTCF, RAD21, and ZNF143 resulted in markedly increased prediction performance with auPRCs of 0.31, 0.37, and 0.38 for CTCF, RAD21, and ZNF143, respectively (Fig. 5.3B, C). For visual comparison, we show the actual looping interactions and 7C predictions on an example region at chromosome 11 (Fig. 5.3D).

Finally, we built a full model using the sequence based features and the ChIP-seq data of all six selected TFs. This only resulted in a slight increase of prediction performance to auPRC = 0.42 (Fig. 5.3B, C), indicating that a single ChIP-seq experiment might be sufficient for accurate prediction of chromatin loops.

We also tested if a single value of correlation of ChIP-seq signal at both loop anchors across the six different TFs is predictive. Indeed, we find high prediction performance of auPRC = 0.34 for this approach. However, this was lower than us-

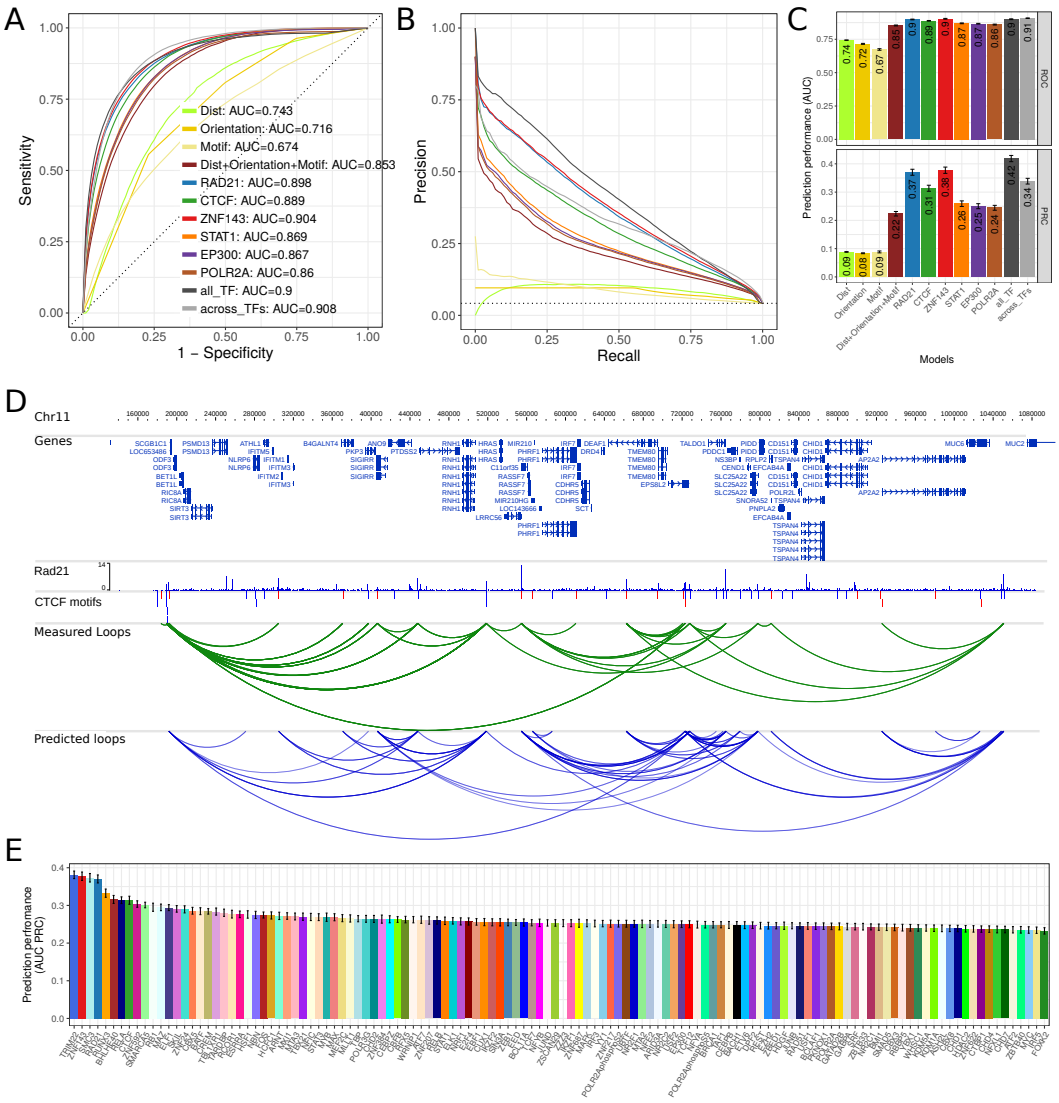

Figure 5.3.: Prediction performance using cross-validation. **(A)** ROC plot for different models to predict chromatin looping interactions. The sensitivity (y-axis) is shown against the false discovery rate (1 – specificity, x-axis) for thresholds of the prediction score. Curves show averages of 10-fold cross-validation experiments. The models "Dist", "Orientation", and "Motif" contain only a single feature as indicated and all three genomic features are combined in the model "Dist+Orientation+Motif". The models "RAD21", CTCF", "ZNF143", "STAT1", "EP300", and "POLR2A", contain the genomic features and the ChIP-seq correlation of the indicated factor. The model "all_TF" contains the genomic features and correlation of all indicated TFs. The model "across_TFs" contains the genomic features and a single correlation feature across the six ChIP-seq datasets as described in the main text. **(B)** PRC plot of precision against the recall for different prediction models. Color code as in (A). **(C)** Values of the area under the ROC (top) and PRC curves (bottom) as prediction performance. Error bars indicate standard deviation in 10-fold cross-validation experiments. **(D)** Example region on chromosome 11 in the genome browser showing: human genes, RAD21 ChIP-seq data in GM12878, CTCF motifs, CTCF motif pairs with that interact according to Hi-C or ChIA-PET data (green arcs) and predicted chromatin loops from RAD21 ChIPseq data using 7C (blue arcs). **(E)** Prediction performance of 7C as auPRC values for models with 124 TF ChIP-seq data sets from ENCODE. Error bars as in (C).

ing the correlation from single TF ChIP-seq experiments for RAD21 or ZNF143 and has the disadvantage of relying on ChIP-seq data from multiple experiments.

Together, these results show that sequenced based features alone have only a limited loop prediction performance, but integrating them with a single ChIP-seq experiment, 7C can predict chromatin loops with high accuracy.

Comparison of transcription factors by prediction performance

Our results can be used to better understand the molecular mechanisms of chromatin loop formation. We hypothesize that TFs whose ChIP-seq provides high prediction performance are likely to be functionally involved in chromatin looping. These TFs would be therefore interesting targets for further investigation of their potential function in chromatin looping.

To investigate this for as many TFs as possible, we used all available 124 TF ChIP-seq datasets from ENCODE for the human cell line GM12878 and compared transcription factors by their prediction performance. Notably, nearly all TF ChIP-seq data sets could increase the prediction performance of sequence-based features alone (Fig. 5.3E). However, there was a large variance in performance between TFs and a subset of TFs with high predictive power could be identified. These include for example the known architectural proteins mentioned above, CTCF, cohesin (RAD21 and SMC3), and ZNF143, but also factors, such as TRIM22, RUNX3, BHLHE40, or RELA, which might be interesting candidate factors with functional roles in chromatin loop formation.

Prediction performance in other cell types and for different TFs

Next, we wanted to test if 7C is general enough to predict looping interactions in a cell type different to the one used to train it. To test this, we used the models presented above (trained with data from human GM12878 cells) to predict loops using as input ChIP-seq data from human HeLa cells. The prediction performance was assessed using as positives 12,480 loops (1.74 % of all motif pairs) identified in HeLa cells (Rao et al., 2014; Tang et al., 2015). While prediction performance in HeLa cells is slightly lower as compared to the cross-validation in GM12878 cells, we see overall very good prediction performance also in HeLa cells by ROC curves (auPRC up to 0.91, Fig. 5.4A) and PRC curves (auPRC up to 0.27, Fig. 5.4B,C).

In this analysis, we compared the prediction performance of each specific TF model. However, in an application use case, one might not be able to train the model for a specific TF of interest and the model should predict loops for TFs that were not used in the training. Therefore, we built default 7C models by either averaging model

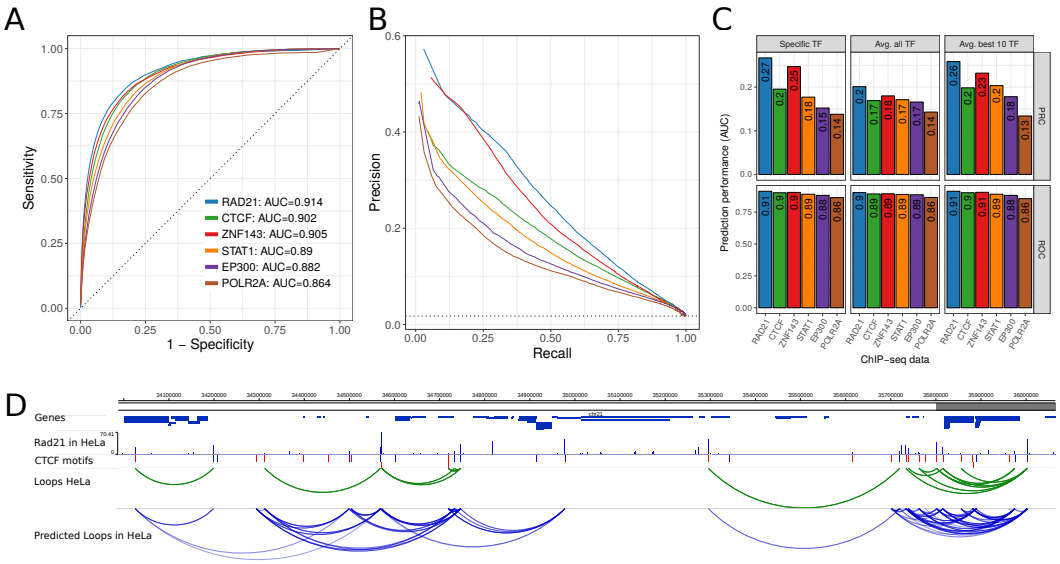

Figure 5.4.: Prediction performance in HeLa cells using 7C trained in GM12878 cells. **(A)** ROC curve of prediction performance of six selected TF ChIP-seq data sets. The 7C model was trained using ChIP-seq and true loop data in human GM1287 but loops were predicted using ChIP-seq data of the same TFs in HeLa cells and true loop data in HeLa cells. **(B)** Precision-Recall curves for the same analysis as in (A). **(C)** Prediction performance as auPRC (top) and auROC (bottom) in HeLa for the six TF ChIP-seq data sets (x-axis) and 7C models trained for the specific TF (left), 7C with parameters averaged across all 124 TF models (center), and 7C with parameters as average of the 10 best performing TF ChIP-seq data sets (right). **(D)** Example region on human chromosome 21 with genes, RAD21 ChIP-seq data in HeLa, CTCF motifs, true loops in HeLa cells according to Hi-C and ChIA-PET (green arcs) and predicted chromatin loops from RAD21 ChIP-seq data in HeLa (blue arcs).

parameters from all 124 TF models or by averaging across the model parameters of only the 10 best performing TFs.

While all three approaches result in good prediction performance for the six selected TFs (Fig. 5.4C), the model averaging parameters across all TFs performs poorer than the ones of only the best 10 models, which are actually nearly as good as the specific TF models. This is consistent with similar results from cross-validation analysis in GM12878 data (Fig. S1C). Furthermore, we visually inspected chromatin loop predictions from RAD21 ChIP-seq data in HeLa at an example loci on chromosome 21 (Fig. 5.4D). In summary, these results show that 7C can predict chromatin looping interactions in different cell types that were not used to train it. Similarly, the 7C default prediction model performs nearly as good as a TF specific model. This makes 7C applicable for ChIP-seq data from diverse TFs in many different cell types and conditions.

The high resolution of ChIP-nexus improves prediction performance

We wondered if the similarity of other genomics signals at loop anchors could potentially indicate looping interactions. Therefore, we used different genomic assays, such as DNase hypersensitivity (DNase-seq), ChIP-nexus and only ChIP-seq input material as input to our prediction methods (Fig. 5.5).

Furthermore, we compared different signal types of ChIP-seq. During computational processing of ChIP-seq raw data, reads are shifted in 5' direction by the estimated average fragment size (Zhang et al., 2008; Hansen et al., 2015). The coverage of these shifted reads is then compared to coverage of input control experiment (fold change over control). Furthermore, a recent study quantified read pairs (qfrags) in a specific distance to each other as estimate for the actual fragment numbers detected by ChIP-seq (Hansen et al., 2015). For most of the TFs tested here, we observed that the ChIP-seq signal types 'shifted reads' and 'qfrag' have better loop prediction performance than the 'fold change over control' (Fig. 5.5). Interestingly, also the ChIP-seq input signal alone results in better prediction performance than sequence features alone, indicating that cross-linking efficiency and density of chromatin itself is specifically distributed at chromatin loop anchors (Fig. 5.5). Also, DNase-seq, which measures chromatin accessibility, predicts looping interactions with similar accuracy than ChIP-seq input control (Fig. 5.5). This is consistent with specific open-chromatin profiles at TF binding sites (Pique-Regi et al., 2011; Yardmc et al., 2014).

However, using ChIP-nexus data for RAD21 and SMC3 (Tang et al., 2015), we could markedly improve chromatin loop predictions using 7C (Fig. 5.5). ChIP-nexus and ChIP-exo are variations of the ChIP-seq protocol, in which additionally, an exonuclease digestion step is applied to trim the DNA from the 5' end until the actual

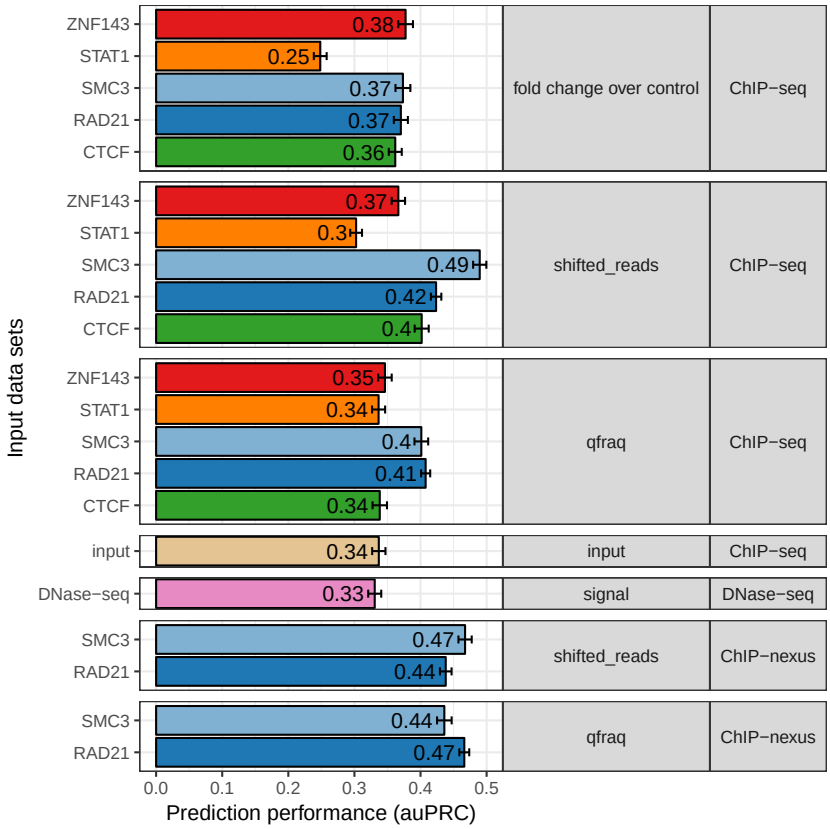

Figure 5.5.: Higher resolution of ChIP-exo and ChIP-nexus improves prediction performance. Prediction performance as area under the precision recall curve (auPRC, x-axis) for 7C models with different input data sets to predict chromatin looping (y-axis). Input data sets are grouped by signal-type (middle panel) and assay-type (right panel) and colored according to the TF (if any) used in the experiment.

bound protein (Rhee and Pugh, 2011; He et al., 2015). These signals result in high-resolution binding footprints that can be used to identify different TF binding modes and cooperation with co-factors (Starick et al., 2015). Therefore, we conclude, that the high-resolution binding profiles from ChIP-nexus allow to compute a more predictive binding signal similarity at chromatin loop anchors.

In summary, the comparison of different genomic signal types shows that cross-linking effect and chromatin density at chromatin anchors are predictive signals for long-range chromatin interactions and higher resolution TF binding assays, such as ChIP-nexus, result in improved prediction performance.

Discussion

We have developed 7C to reuse ChIP-seq data, profiling the interactions of proteins with genomes, for the prediction of chromatin looping interactions. We present this method as an alternative to dedicated techniques like Hi-C that directly measure genomic contacts. Since the results of ChIP-seq experiments are increasingly available for a large number of proteins, species, tissues, cell types, and conditions, our method offers a valid alternative when Hi-C data is not available, or cannot be produced due to cost or material limitations. Another major advantage of our method over Hi-C is that the predictions are at a base pair resolution, while Hi-C only reaches resolutions of at best kilo base pairs at a high cost.

Other computational approaches were developed to predict genomic contacts or assign regulatory regions to target genes. A commonly used approach is to compare activity signals at enhancers and promoters across many different conditions or tissues (Sheffield et al., 2013; Fishilevich et al., 2017; Andersson et al., 2014; O'Connor and Bailey, 2014): high correlation indicates association and potential physical interactions between enhancers and genes. However, these approaches lose the tissue specificity of the interactions. Other approaches integrate many diverse chromatin signals such as post-translational histone modifications, chromatin accessibility, or transcriptional activity (Roy et al., 2015; Whalen et al., 2015; Zhu et al., 2016; Schreiber et al., 2017; Dzida et al., 2017), and combine them with sequence features (Zhao et al., 2016), or evolutionary constrains (Naville et al., 2015). While these methods predict enhancer-gene association with good performance, they require for each specific condition of interest a multiplicity of input datasets, which are often not available.

Further computational approaches try to directly predict chromatin interactions by using diverse sequence features (Nikumbh and Pfeifer, 2017) or multiple chromatin features such as histone modifications (Brackley et al., 2016; Chen et al., 2016) or transcription (Rowley et al., 2017). One study makes use of the more recently discovered CTCF motif directionality to predict loop interactions from CTCF ChIP-seq

peak locations (Oti et al., 2016). Another study combines CTCF binding locations and motif orientation with polymer modeling to predict Hi-C interaction maps (Sanborn et al., 2015). However, none of these studies predicts chromatin loops from ChIP-seq signals of TFs different from CTCF by taking the CTCF motif orientation into account. Furthermore, CTCF binding sites are often only considered, when the signal is strong enough for peak calling algorithms to identify binding sites. In contrast, 7C takes the distribution of ChIP-seq signals from all TFs into account without a peak-calling step. Furthermore, the other studies do not provide a tool for the direct prediction of pairwise interactions from single ChIP-seq experiments. Interestingly, shadow peaks in ChIP-seq data of insulator proteins in *Drosophila* were previously associated to long-range interactions (Liang et al., 2014) and used to study the contribution of sequence motifs and co-factors in loop formation (Mourad et al., 2017), but not to directly predict chromatin loop interactions.

Compared to the predictive methods mentioned above, our approach has the clear advantage to directly predict chromatin looping interactions, and not enhancer-promoter associations, by making use of ChIP-seq signals from a single experiment with respect to CTCF motifs. This gives the prediction a base pair resolution since it relies in the alignment of a pair of CTCF motifs. In fact, given several CTCF motifs within a 1kb genomic bin, our looping prediction approach can be used to decide which of the CTCF sites is actually involved in the measured interactions and thus increase resolution even when Hi-C data is available. We showed that our approach, 7C, can work with just a single ChIP-seq experiment for many different TFs, making it usable for many diverse conditions of interest. Therefore, 7C can be used complementary to existing enhancer-promoter association tools or can be integrated in such predictive models to improve them.

Association of distal cis-regulatory elements, such as TF binding sites or enhancers, to their regulated target gene is a common problem in genomic studies (Mora et al., 2015), and this can be addressed by methods mapping contacts at a base pair resolution. While Hi-C measures pairwise contacts genome-wide in an unbiased manner and experimentally measuring genomic interactions is now becoming feasible due to the recent advances in 3C based technologies (Sati and Cavalli, 2016), a main drawback of the Hi-C method is the limitation of resolution. While the first Hi-C study analyzed chromatin interactions at bin sizes of 1Mb (Lieberman-Aiden et al., 2009), structural features such as topologically associating domains (TADs) were later called at 40kb bin resolution (Dixon et al., 2012), and the highest resolution for the human genome, reached only recently, is 1kb (Rao et al., 2014). This higher resolution requires largely increased sequencing depths (Rao et al., 2014; Bonev et al., 2017). Capture Hi-C identifies only the interactions of per-defined target regions such as promoters (Dryden et al., 2014; Mifsud et al., 2015). ChIA-PET restricts the interactions to those where a specific protein of interested is involved (Heidari et al., 2014; Fullwood et al., 2009; Tang et al., 2015). Therefore these

experiments are not always applicable and require a large amount of sample material. Ultimately, even in high resolution Hi-C it is not trivial to connect enhancers to interacting genes in a unique way, since enhancers are often found in clusters within short distances in the genome (Whyte et al., 2013; Parker et al., 2013).

Currently, our method, by using CTCF motifs, focuses on CTCF mediated chromatin loops. It is very likely that other DNA binding proteins mediate loops: for example, recent studies suggest that other TFs are involved in enhancer promoter interactions during differentiation (Bonev et al., 2017) and knockout of transcriptional repressor YY1 and other candidate factors result in loss of chromatin loops (Weintraub et al., 2018). Using motifs predicted for these different transcription factors, or combinations thereof, are open avenues for the future extension of our method.

Conclusion

We demonstrated that TF binding signals of ChIP-seq experiments at CTCF motifs are predictive for chromatin looping. We provided a method, 7C, that is simple to use and integrates these signals with genomic sequence features to predict long-range chromatin contacts from single ChIP-seq experiments. 7C is freely available as R package (https://ibn-salem.github.io/sevenC/). The analysis of ChIP-seq experiments for 124 different TFs highlighted the role of cohesin, ZNF143 and CTCF in chromatin loop formation, but also suggested many other TFs, such as TRIM22, RUNX3, and BHLHE40, to be functionally involved in chromatin looping, likely in cooperation and protein interaction (direct or indirect) with CTCF at loop anchors.

Since our method needs only a single ChIP-seq experiment as input, it enables the analysis of chromatin interactions in diverse cell types and conditions, where Hi-C like data is not available. Therefore, 7C can be used to enable condition specific associations of distal TF binding sites and enhancers to promoters of target genes. These might allow the interpretation of non-coding genetic variants by genes in physical contact with the variant loci in a specific cell type or condition of interest. Furthermore, 7C might improve the resolution of Hi-C interaction maps by facilitating base-pair specific pairing of CTCF motifs located in bins of several kb. With these applications, 7C increases the value of ChIP-seq datasets, which now can be used to improve the analysis of 3D genome folding and their dynamic changess between diverse cell types and conditions.

Methods

CTCF motifs in the human genome

The recognition motif of CTCF is well defined and available from the JASPAR database (MA0139.1) (Mathelier et al., 2015). We downloaded TF binding site predictions with the CTCF motif (MA0139.1) in the human genome hg19 from the JASPAR database (*http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/hg19/tsv/MA013*
Motif hits were filtered for p-value 2.5 x 10^{-6}, resulting in 38,316 highly significant CTCF motif hits genome-wide and 717,137 motif pairs within 1 Mb genomic distance that are considered as potential loop interaction anchors in this study.

Loop interaction data for training and validation

For training and validating the prediction model we used 9,448 published loops derived from high-resolution *in-situ* Hi-C experiments (Rao et al., 2014) and 206,399 CTCF and Pol2 ChIA-PET interactions (Tang et al., 2015) in human GM12878 cells. We considered each CTCF motif pair as positive (true looping interaction) if there was at least one measured looping interaction for which each loop anchor overlapped one of the CTCF motifs. Overlaps were calculated using the R package *InteractionSet* (Lun et al., 2016). This resulted in 30,025 (4.2%) of 717,137 candidate motif pairs that were labeled as true looping interactions in GM12878. For the prediction validation in HeLa cells we used the 3,094 Hi-C loops and 402,722 ChIA-PET interactions for CTCF and Pol2 in HeLa from the same studies (Rao et al., 2014; Tang et al., 2015) and labeled 12,480 (1.7 %) of motif pairs as true loops in HeLa cells.

ChIP-seq datasets in GM12878 cells

We downloaded publicly available ChIP-seq data from the ENCODE data portal (Dunham et al., 2012; Davis et al., 2017) by requiring the assay to be ChIP-seq, the target to be a transcription factor, the biosample term name to be GM12878, the genome assembly to be hg19, and the file-type to be bigWig. Furthermore, we filtered the data to have output type 'fold change over control' or 'signal' and to be built from two replicates. Then we selected for each TF only one unique experiment as bigWig file with either output type 'fold change over control' or, if unavailable, output type 'signal'. This resulted in 124 ChIP-seq experiments for different TFs (Table S1). ChIP-seq data for HeLa were retrieved analogously and filtered for the selected targets: RAD21, CTCF, ZNF143, STAT1, EP300, and ZNF143 (Table S2).

ChIP-seq data types

To analyze the effect of different ChIP-seq signal types and other genomic assays on loop prediction performance, we selected five TFs (ZNF143, STAT1, SMC3, RAD21, and CTCF) and downloaded the mapped reads of ChIP-seq experiments as BAM

files from the ENCODE data portal (Davis et al., 2017) and from the UCSC Genome Browser (Hinrichs et al., 2006). Furthermore, we downloaded signal tracks as bigWig files for ChIP-seq input control experiment and DNase-seq experiments in GM12878 cells. File accession identifiers and download links are provided in Table S3. We used the ChIP-seq peak caller *Q* (Hansen et al., 2015) with option '-w' for each human chromosome to generate signal tracks in BED format of shifted reads and qfrags. 'Shifted reads' are counts of mapped reads that are shifted in 5' direction by half of the estimated fragment size. 'qfrags' are pairs of forward and reverse mapped reads within a given distance (Hansen et al., 2015) and are shown to improve signal to noise ratio in ChIP-seq peak calling (Hansen et al., 2015). We then combined resulting BED files from all chromosomes and converted them to the bedGraph and bigWig formats using the *bedtools* (Quinlan and Hall, 2010) and *bedGrpahtoBigWig* tools from the UCSC Genome Browser (Kent et al., 2010).

ChIP-nexus data processing for RAD21 and SCM3

ChIP-nexus data for RAD21 and SMC3 in GM12878 cells were published recently (Tang et al., 2015). We downloaded the corresponding raw reads from the Sequence Read Archive (SRA) (Run IDs SRR2312570 and SRR2312571). Reads were processed using *felxcut* for barcode removal and adapter trimming as recommended in the user guide of the *Q-nexus* tool (Hansen et al., 2016). Reads were than mapped to human genome hg19 using *Bowtie* version 2.3.2 with default settings. Duplicate reads were removed using *nexcat* (Hansen et al., 2016). Finally, we created shifted-reads and qfraq profiles using *Q-nexus* (Hansen et al., 2016) with options '–nexus-mode' and '-w' for each chromosome and combined them to bigWig files as described above.

Similarity of ChIP-seq profiles as correlation of coverage around motifs

For each CTCF sequence motif in the human genome, we quantified the number of reads overlapping each base within +/- 500 bp around the motif center. This results in a vector $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,n})$ where $x_{i,k}$ is the coverage signal at position k around CTCF motif i. Coverage vectors for motif hits reported on the minus strand were reversed because CTCF motif sites are assumed to be symmetrically aligned to each other when cooperating at loop anchors (Fig. 5.1A) (Rao et al., 2014; Tang et al., 2015; Guo et al., 2015; de Wit et al., 2015; Sanborn et al., 2015). For all considered pairs of CTCF motifs i and j, we calculated the ChIP-seq profile similarity as Pearson correlation coefficient $r_{i,j}$ of the corresponding coverage vectors x_i and x_j.

Genomic sequence features of chromatin loops

Besides the correlation of ChIP-seq profiles, we used genomic features of motif pairs as features to predict interactions. The distance *d* is the number of bp between the two motif centers. The categorical variable orientation *o* is either, *convergent*, *forward*, *reverse*, or *divergent*, depending on the orientation of CTCF motifs in the pair (+-, ++, –, and -+, respectively). The motif hit similarity *s* is the minimum of the two motif hit scores in each pair; we derived these motif scores from the JASPAR motif hit tracks as $-\log_{10}$ transformed p-values (Khan et al., 2018).

Chromatin Loop prediction model

We used a logistic regression model to predict the log-likelihood probability of CTCF motif pairs to perform chromatin looping interactions. The probability p that two sites interact is modeled as:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1, + ... + \beta_k x_k$$

where β are the unknown model parameters and $x_1, ..., x_k$ the features.

More specifically, for the 7C model with a single ChIP-seq experiment as input, the logistic regression model for the interaction probability *p* is:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 d, + \beta_2 o + \beta_3 s + \beta_4 r$$

Parameters were estimated using the function 'glm()' with option 'family=binomial()' in R during model training as described below.

Training and validation of prediction model

We used the R package *rsample* for 10-fold cross-validation. Thereby, we randomly split the dataset of CTCF motif pairs into ten equal sized subsets. For each round of cross-validation one subset is held out (test dataset) and the model parameters are trained on the remaining 90% of the samples (training dataset). The model parameters are shown for six selected TFs and combined models in Figure S1A. For each split, the performance of the model is than evaluated on the test dataset. For prediction performance in HeLa cells, we trained on all motif pairs using ChIP-seq and true loops from GM12878 cells and evaluated performance on all motif pairs using the true loop data in HeLa.

Analysis of prediction performance

We quantified prediction performance using the receiver operating characteristic (ROC) and precision recall curves (PRC) as implemented in the R package *precrec* (Saito and Rehmsmeier, 2016).

Given the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the sensitivity is defined as TP/(TP+FN), specificity as TN/(TN+FP), precision as TP/(TP+FP), and recall as TP/(TP+FN). For each cross-validation split, the area under the curve is computed separately, and the mean across splits together with the standard deviation reported. To get binary prediction outputs, we computed the f1-score as harmonic mean of precision and recall for all prediction scores on all cross-validation folds using the R package *ROCR* (Sing et al., 2005). Then we computed the prediction score that maximizes the f1-score as default cutoff for binary prediction output (Fig. D.1B).

Implementation of 7C and compatibility to other tools

We implemented 7C as R package, termed *sevenC*, by using existing infrastructure for chromatin interaction data from the *interactionSet* package (Lun et al., 2016) and functionality for reading bigWig files from the *rtracklayer* package (Lawrence et al., 2009) from the Bioconductor project (Huber et al., 2015). Predicted loops can be written as interaction tracks for visualization in the WashU Epigenome Browser (Zhou et al., 2013) or as BEDPE format using the *GenomicInteractions* package (Harmston et al., 2015) for visualization in the Juicebox tool (Durand et al., 2016). The package is freely available and easy to install from GitHub under *https://ibn-salem.github.io/sevenC/* and has been submitted to the Bioconductor project (Huber et al., 2015). All analysis presented in this work were implemented in R and all scripts used have been made available in a separate GitHub repository: *https://github.com/ibn-salem/sevenC_analysis*.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The source code for the *sevenC* package is available on GitHub: *https://github.com/ibn-salem/sevenC*. The source code for all analyses in this manuscript is available on GitHub: *https://github.com/ibn-salem/sevenC_analysis*. All the genomic data used for analyses are freely available to be downloaded from the GEO repository or ENCODE as described in the methods section.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors‘ contributions

JI developed and implemented the method, conceived the study and performed all analyses. MA supervised the study. JI and MA wrote the manuscript.

Acknowledgments

The authors thank all members of the CBDM group for fruitful discussions, especially Katerina Takova for ideas on modeling and cross-validation. We further thank Idan Gabdank (Stanford University) for support in understanding ENCODE data set and updating it, Takaya Saito (University of Bergen) for updating and improving the *precrec* R package, Peter Hansen (Charité, Berlin) for support in ChIP-seq and ChIP-nexus analysis using *Q*, Morgane Thomas-Chollier (Ecole normale supérieure) for discussions on genome-wide motif analysis, and Sebastiaan Meijsing (Max Planck Institute for Molecular Genetics) for initial discussions on ChIP-exo profiles at chromatin loop anchors.

Discussion

6

- Limitations of 3C based methods
- Plasticity and dynamics of chromatin interactions
- Single cell resolution
- Functional mechanism
- Establishment of compartments / TADs / loops
- Regulation of compartments / TADs / loops
- Notes
  - Neo-TADs by tandem-duplication in evolution (Franke et al., 2016)
  - A key component and driver for new gene function in evolution or neo-functionalization can be the birth of new enhancers through acquisition of transcription factor binding and subsequent novel regulatory functions (Long et al., 2016).
  - This data can be computationally integrated with one-dimensional measurements along the genome and lead to exciting findings of higher order organisation.
  - TADs are not only structural units of chromosomes, but also functional building blocks of genomes.

Discussion paralog co-regulation in TADs

The association of gene expression with gene localizing in TADs is consistent with a recent computational study with the aim to separate the proportion of expression associated with genome organization from independent sources. A large fraction of expression can be attributed to positioning of genes in genome architecture and hihgly informative for TAD activity and organization (Rennie et al., 2018).

Discussion on TAD evolution

Follow up studes might experimentally support the functional importance of evolutionary conserved TADs. A very recent study investigated the requirenment of ultra-conserved non-coding region containing enhancers by knock-out experiments in mice (Dickel et al., 2018). Knowck-out mice that lack individual or combination of enhancers were viable but showed strong neurological phenotypes. Indicating

that the remarkably strong sequence conservation likely results from fitness deficit that appear subtle in a laboratory setting (Dickel et al., 2018). This could be similar for targeted deletions of conserved TAD boundaries.

The increased expression conservation of genes within TADs is somewhat consistent with a very recent study, in which promoter and enhancer activity was measured in liver samples from 15 species (Berthelot et al., 2017). In this study gene expression conservation could best be explained by the number and conservation of surrounding enhancers and promoters.

In this thesis, I analysed the functions of TADs for gene regulation and highlight their stability in evolution as well as their consideration when analyzing genomic variations in patient genomes. - Paralog genes as model for co-regulation in TADs - enhancer sharing - co-expression

As consequence of these association of TADs with co-regulation, enhancer sharing and co-expression, we hypothesized TADs provide regulatory environments for genes and therefore be conserved during evolution. More specifically, we asked whether genomic rearrangement between related species would more frequently occur at TAD boundaries. Furthermore, we hypothesized that disruption of TADs during evolution might be associated with changes of gene expression programs between the species.

The analysis of genomic rearrangements between human and other species during evolution lead to the conclusion that TADs are important regulatory building blocks of genomes. Indeed the changes of expression profiles are associated with the disruption of TADs during evolution. This might likely lead to severe disadvantages to the organism, as was observed for example in genetic diseases (Ibn-Salem et al., 2014; Lupiáñez et al., 2015) and cancers. Therefore, we interpret the depletion of evolutionary rearrangements in TADs and the expression change associated with TAD disruption to be a consequence of selective pressure on TAD structure. Therefore selective pressure is likely to act on TAD structures. While in neutral selection

- Relate to GRBs and CNE clusters (Harmston et al., 2017; Polychronopoulos et al., 2017)

TAD disruption as a new mechanism of disease pathogenesis.

The majority of disease-associated variants uncovered by genome-wide association studies (GWAS) reside in noncoding sequences. Many variants are located near cis-regulatory sequences and enhancers and could, therefore, contribute to pathogenesis by affecting transcription of specific genes (Hindorff et al., 2009). The ability to measure long-range chromatin interactions makes it possible to assign understand

the role of non-coding variants by predicting its interacting target gene (Smemo et al., 2014; Visser et al., 2012). This has been shown by measuring the interactions of promoters in 17 human primary hematopoietic cells types revealing more than 2,400 potential disease-associated genes liked to thousands of GWAS SNPs (Javierre et al., 2016). In another study, the noncoding variants associated with schizophrenia could be annotated using Hi-C contact maps from human cerebral cortex (Won et al., 2016).

Further directions

- Plasticity and dynamics of chromatin interactions
- Single cell resolution
  - Single cell Hi-C studies: (Nagano et al., 2017; Stevens et al., 2017)
  - Computational modelling of single cells (Sekelja et al., 2016).
- Functional mechanism
- Establishment of compartments / TADs / loops
- Regulation of compartment / TADs / loops

Conclusions

Recent methodological advances in chromatin conformation capture experiments resulted in genome-wide contact maps of genomes. These data lead to many interesting insights in the folding structures of genomes. One important discovery was that chromosomes fold locally into discreet genomic domains, called TADs.

In this thesis, I showed that TADs are not only structural units of genomes, but that they are also functionally important for the correct regulation of gene expression. TADs represent regulatory environment that restrict the interaction landscape of enhancers. Indeed, functionally related genes, such as paralogs, are co-regulated within TADs. During evolution, new genes can emerge by duplication and find established regulatory environments within TADs. Therefore, TADs represent productive nests for novel genes in evolution. The functional important of TADs was further stressed by their stability during hundred million years of evolution. Indeed stable TADs are associated with conserved expression profiles of genes.

Disruption of TADs by rearrangements is associated to changes of gene expression profiles during evolution as well as in genomes of subjects with developmental diseases. While these disruptions of TADs might be beneficial for an organism and lead to evolutionary leaps in some cases, I showed in disease genomes, that disruption of TADs can result in severe phenotypes like mental retardation. Therefore, the three-dimensonal folding structure of genomes, including TADs and enhancer-promoter interactions have to be considered for the interpretation of genomic variants of patient genomes.

While constantly decreasing costs of sequencing will further enable the analysis of individual genomes in many genetic syndromes or cancers, it will be increasingly important to correctly interpret these variants within their functional genomic context. To this end, we need a deeper understanding of the functional role of genome folding including its dynamics between single cells as well as its changes in specific cell types and conditions. To integrate diverse types of functional data that is measured along the genomes with the chromatin folding patterns and their interplay, we need carefully designed computational models. This will address not only fundamental questions such as evolution of genomes, mechanisms of gene regulation in differentiation and development, but also solve practical problems such as the interpretation of genetic variants in disease genomes for better developments of diagnosis and treatments.

Co-regulation of paralog genes: Supporting Information

A

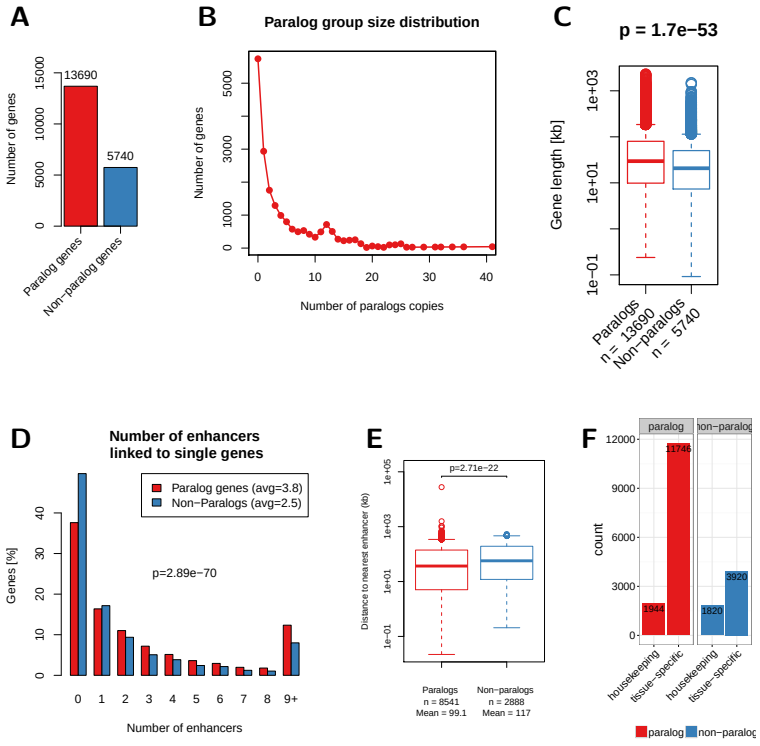

Figure A.1.: **(A)** Number of paralog and non-paralog genes in the human genome. **(B)** Paralog group size distribution in the human genome. **(C)** Gene length of paralog and non-paralog genes. **(D)** Distribution of the number of enhancers linked to single genes compared between paralog genes (red) and non-paralog genes (blue). **(E)** Genomic distance to nearest enhancer for paralogs and non-paralog genes. **(F)** Number of housekeeping genes among paralogs and non-paralog human genes. A recently published set of housekeeping genes was used here (Eisenberg and Levanon, 2013). The p-values shown in this figure were calculated using the Wilcoxon rank-sum test.

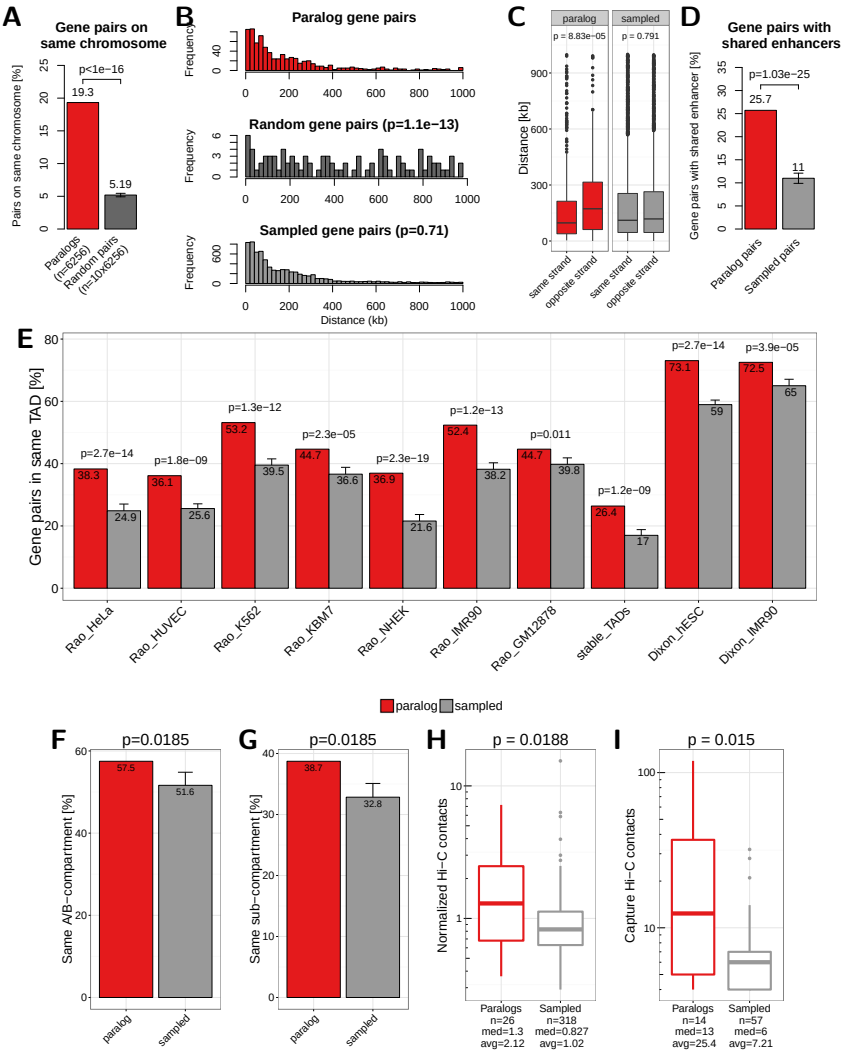

Figure A.2.: Main results of this study by changing the selection of paralog pairs from families with more than two paralogs. Here pairs are selected by maximizing the rate of synonymous mutations between them instead of minimizing, as in the main text. **(A)** Percent of paralogs pairs on the same chromosome compared to random pairs. **(B)** Distance distribution between pairs of paralogs (red), random pairs (dark grey), and sampled pairs according to the distances of paralogs (grey). **(C)** Genomic distance between close paralogs and sampled pairs separated by same strand or not same strand of gene pairs. **(D)** Percent of close paralogs and sampled pairs with at least one shared enhancer. **(E)** Percent of close gene pairs located within the same TAD for different TAD data sets. **(F)** Percent of paralog and sampled pairs that are in the same A/B compartment. **(G)** Percent of paralog and sampled pairs that are in the same subcompartment. **(H)** Normalized Hi-C contacts between distal paralogs and sampled genes. **(I)** Promoter capture-C contacts between distal paralogs and sampled genes.

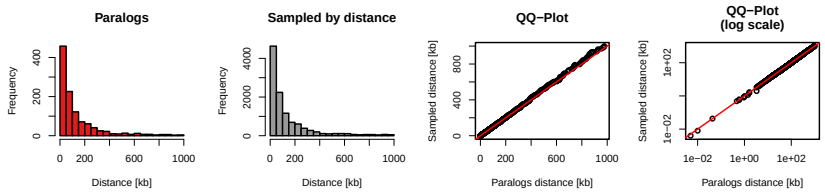

Figure A.3.: Sampling of gene pairs by distance. Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column).

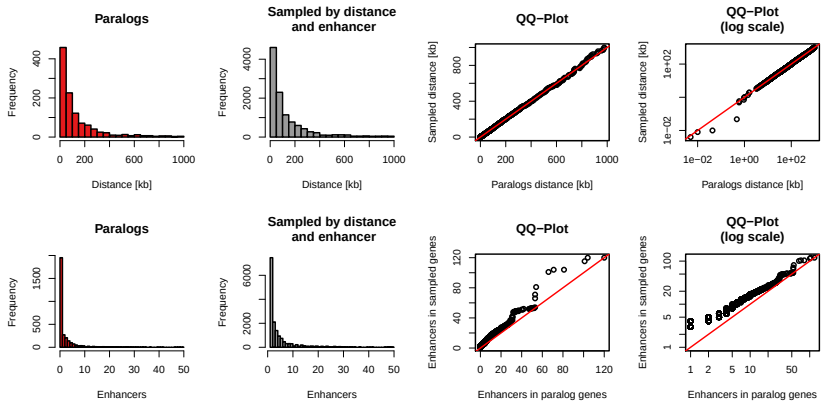

Figure A.4.: Sampling of gene pairs by distance and number of enhancers. Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Bottom row: Distance of the number of enhancers linked to each single gene in the pairs of paralogs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column).

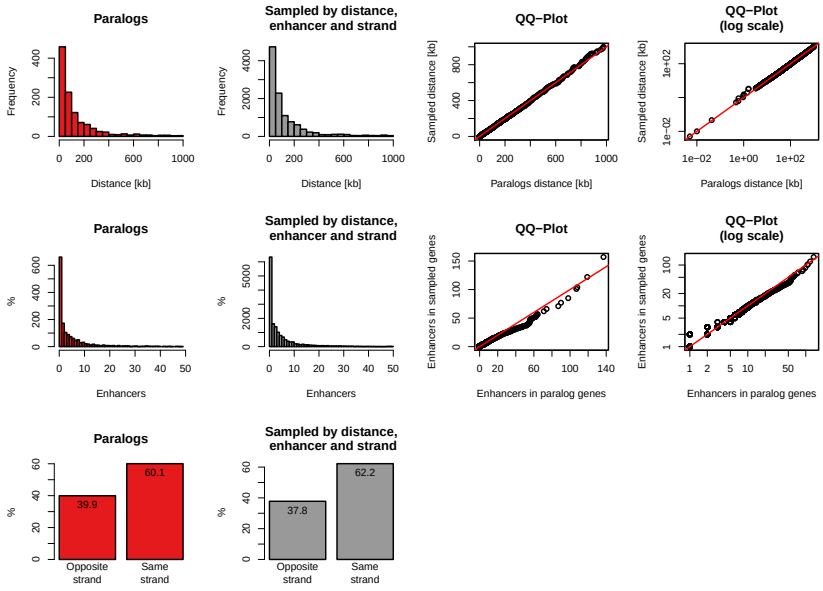

Figure A.5.: Sampling of gene pairs by distance, number of enhancers, and same strand frequency. Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Middle row: Distance of the number of enhancers linked to each single gene in the pairs of paralogs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Bottom row: Percentages of pairs of genes with opposite or same strand of transcription for paralog pairs (red) and sampled pairs (grey).

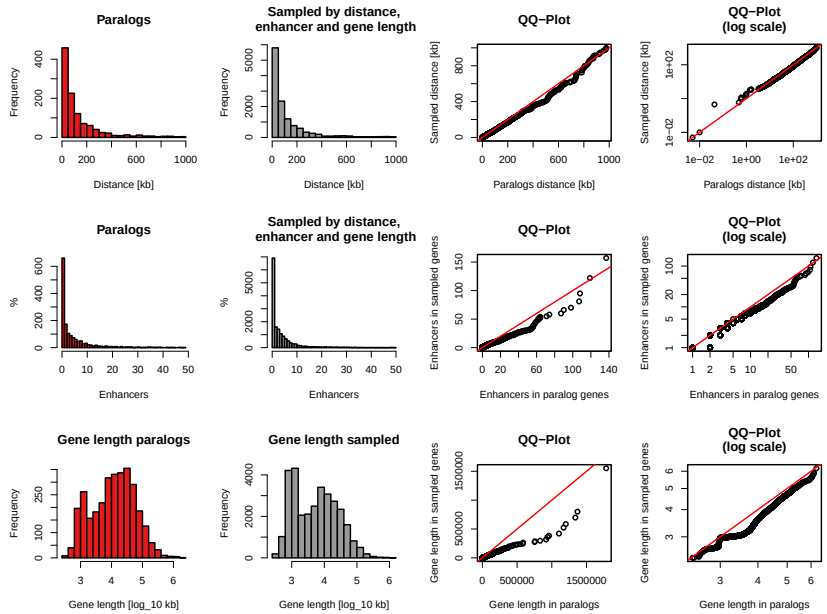

Figure A.6.: **Sampling of gene pairs by distance, number of enhancers, and same strand frequency.** Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Middle row: Distance of the number of enhancers linked to each single gene in the pairs of paralogs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Bottom row: Distribution of gene lengths of each single gene in the pairs of paralogs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and $\log_{10}$ of gene lengths (fourth column).

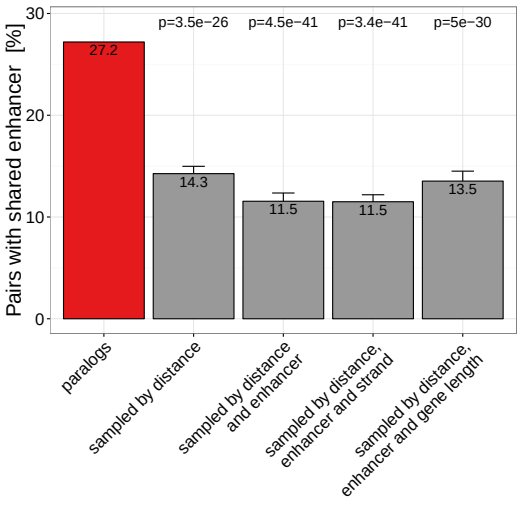

Figure A.7.: Percent of gene pairs with at least one shared enhancer in paralog pairs and four different types of sampled gene pairs. Only pairs with TSS distance $\leq$ 1Mb are considered. Error bars indicate standard variation of ten times replicated sampling.

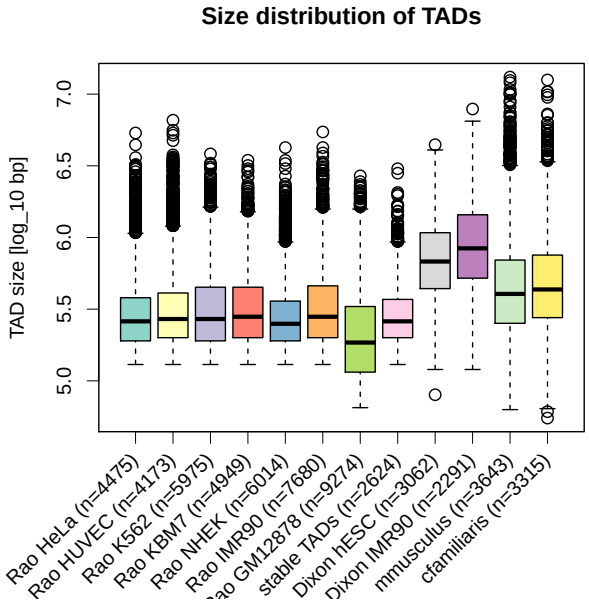

Figure A.8.: **Size distribution of TADs in different cell-types, studies, and species.** Each box shows the size-distribution of one data set of TADs. The labels indicate the study (Rao (Rao et al., 2014), or Dixon (Dixon et al., 2012)), cell type and number of TADs in each data set. The last two boxes are for TADs from Hi-C experiments in mouse and dog Hi-C liver cells (Vietri Rudan et al., 2015).

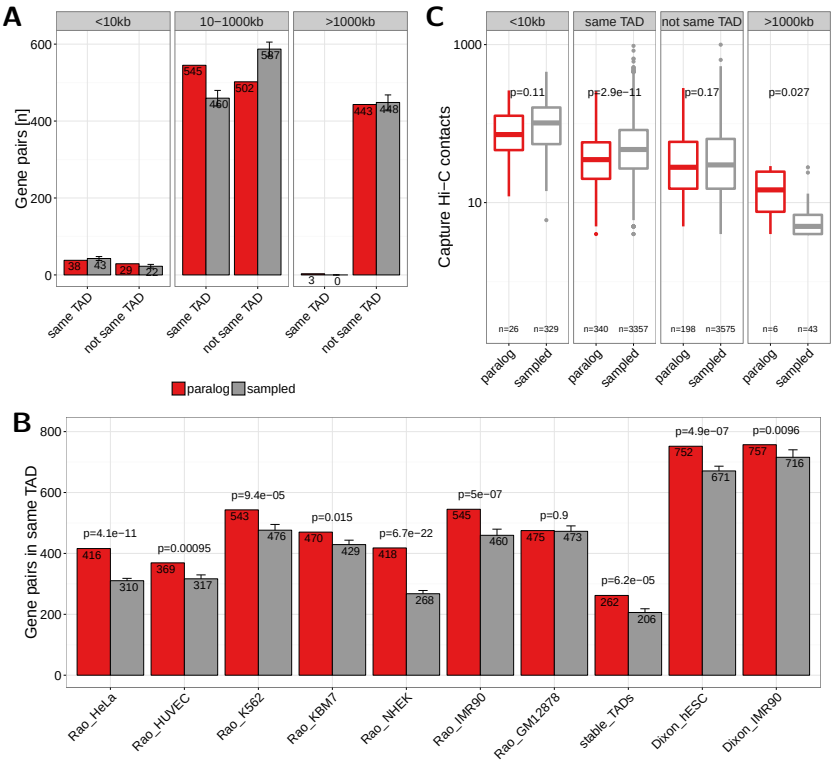

Figure A.9.: **(A)** Number of paralog (red) and sampled (grey) gene pairs that are in the same TAD or not separated in three groups of genomic distances (0-10kb, 10-1000kb and $>$ 1000kb). TADs called from IMR90 cells by (Rao et al., 2014) were used here. **(B)** Co-localization of gene pairs with genomic distances between 10kb and 1000kb within the same TAD for paralogs and sampled gene pairs and separated by TAD data sets from different cell types and studies. The first seven bars show values for TADs called in HeLa, HUVEC, K562, KBM7, NHEK, IMR90, and GM12878 cells by (Rao et al., 2014). The eighth bar shows the value for stable TADs across cell types form this study (at least 90% reciprocal overlap in 50% of cells). The last two bars show data for TADs called in hESC and IMR90 cells by (Dixon et al., 2012). Error bars indicate standard deviation in 10 times replicated sampling of gene pairs. P-values are computed using Fisher's exact test. **(C)** Promoter capture-C contacts between pairs of paralogs (red) and sampled gene pairs (grey) for the groups: $<$10kb genomic distance, located in the same TAD, not in the same TAD, and with genomic distance $>$1000kb.

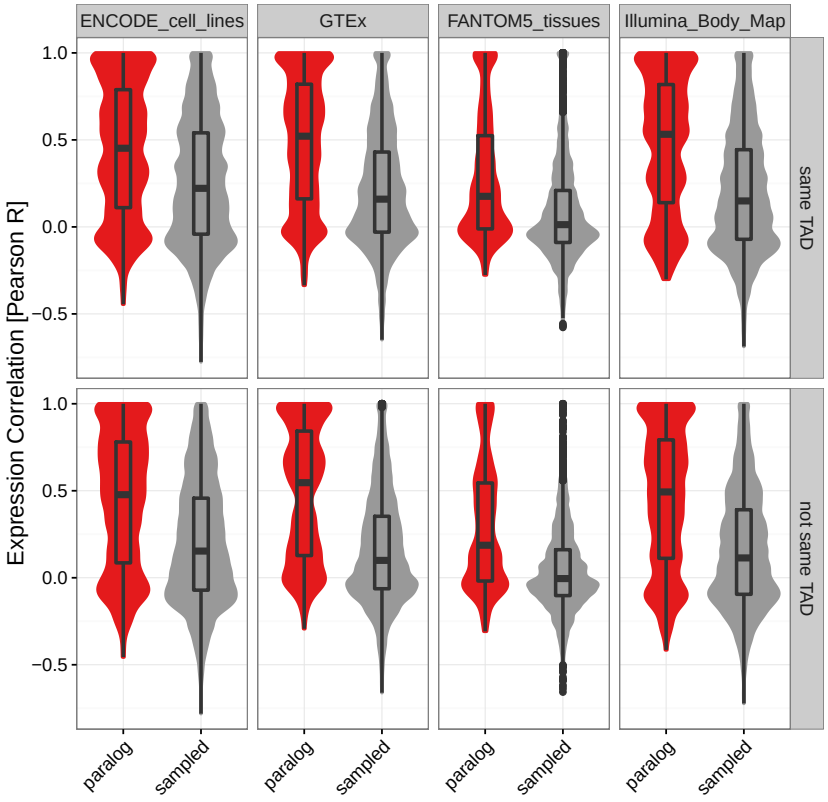

Figure A.10.: Distribution of Pearson correlation coefficients of gene expression values in four independent data sets between close paralog gene pairs (red) and sampled control gene pairs (grey) separated for gene pairs within the same IMR90 TAD (top) or not in the same TAD (bottom). Boxes show 25th, 50th and 75th percent quantile of the data and the filled areas indicate the density distribution.

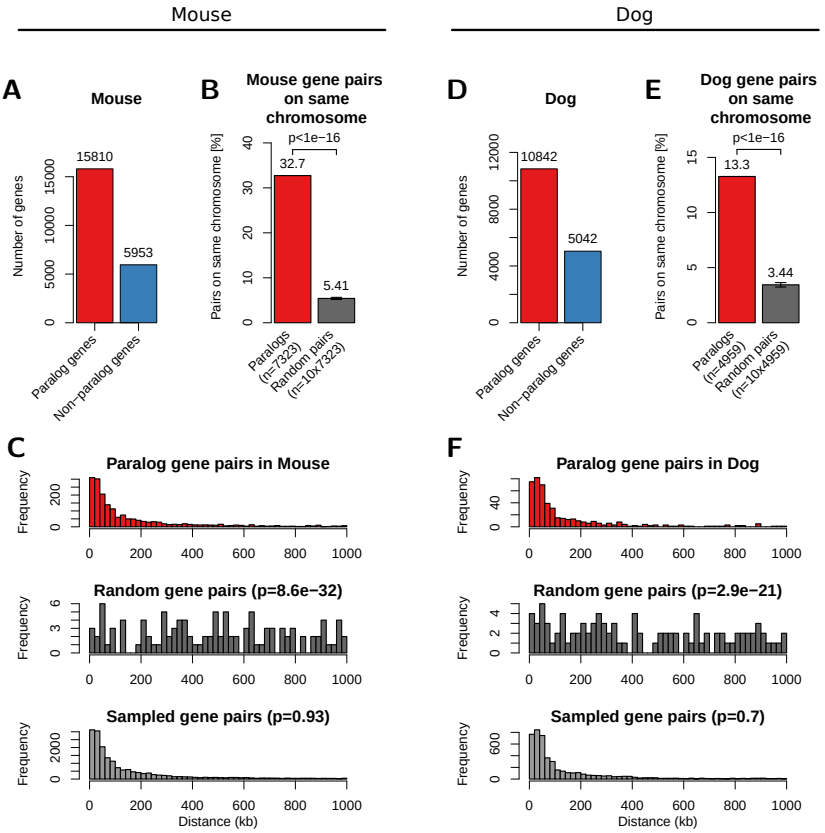

Figure A.11.: Paralog gene pairs in mouse (left) and dog (right) genome cluster on chromosome within short genomic distances. **(A)** Number of genes with paralogs (red) and without (blue) in mouse genomes. **(B)** Percent of filtered mouse paralog pairs on the same chromosome (red) and random gene pairs on the same chromosome (dark grey). Error-bars indicate standard deviation of 10 times replicated randomizations. **(C)** Distribution of linear genomic distances between mouse gene pairs for filtered paralog genes (top, red), random genes (center, dark grey) and sampled gene pairs (bottom, grey). **(D, E, F)** show the same data for the dog genome as figures A, B, C, respectively.

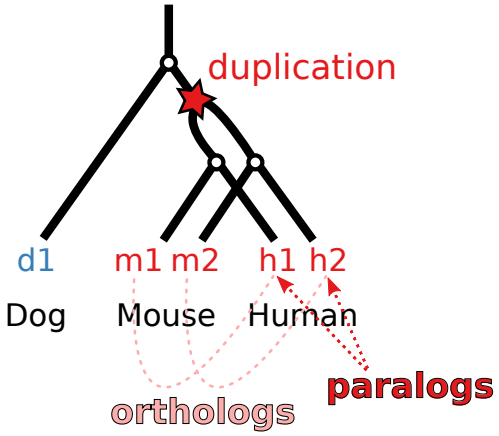

Figure A.12.: Phylogenetic gene tree model of a gene that is duplicated before the separation of mouse and human and consequently leads to two paralogs in mouse and human that are one-to-one orthologs to each other and a single ortholog in the dog genome that cannot be assigned uniquely to a human gene.

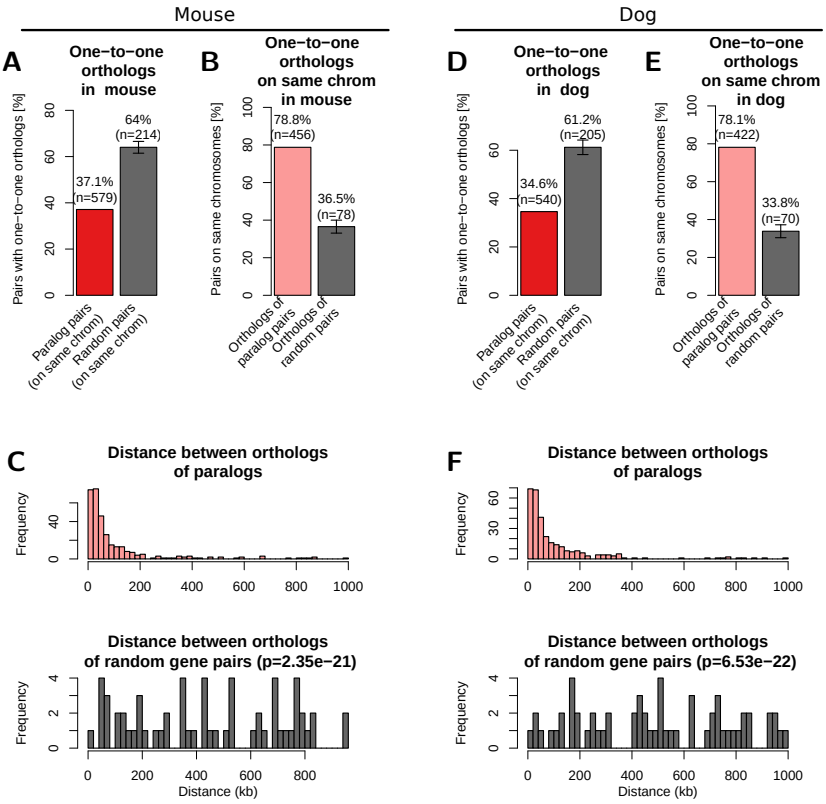

Figure A.13.: One-to-one orthologs of human paralogs in mouse (left) and dog (right) genome. **(A)** Percent of filtered human paralog pairs with one-to-one orthologs for both genes in mouse genome compared to random genes. **(B)** Percent of one-to-one orthologs on the same chromosome in the mouse genome (light red) and one-to-one orthologs of random human gene pairs on the same chromosome (dark grey). Error-bars indicate standard deviation of 10 times replicated randomizations. **(C)** Distribution of linear genomic distances between gene pairs for mouse one-to-one orthologs of human paralog gene pairs (top, light red) and one-to-one orthologs of random human gene pairs (bottom, dark grey). **(D, E, F)** show the same data for the dog genome as figures A, B, C, respectively.

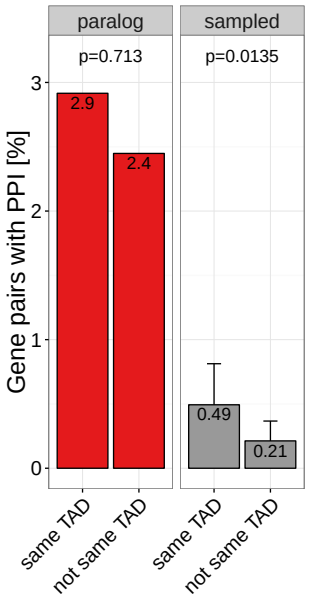

Figure A.14.: Percent of close paralogs (red) and sampled (grey) gene pairs in the same IMR90 TAD (left bar) or not same TAD (right bar) that have a direct protein protein interaction (PPI) with each other in the HIPPIE database (Schaefer et al., 2012).

TAD evolution: Supporting Information

B

Supplementary Tables

Table S1 Matching tissues and samples with CAGE expression data in human and mouse. https://www.biorxiv.org/highwire/filestream/70793/field_highwire_adjunct_files/2/231431-3.tsv

Table S2 Ortholog genes in human and mouse with gene expression correlation across tissues. https://www.biorxiv.org/highwire/filestream/70793/field_highwire_adjunct_files/3/231431-4.tsv

Supplementary Figures

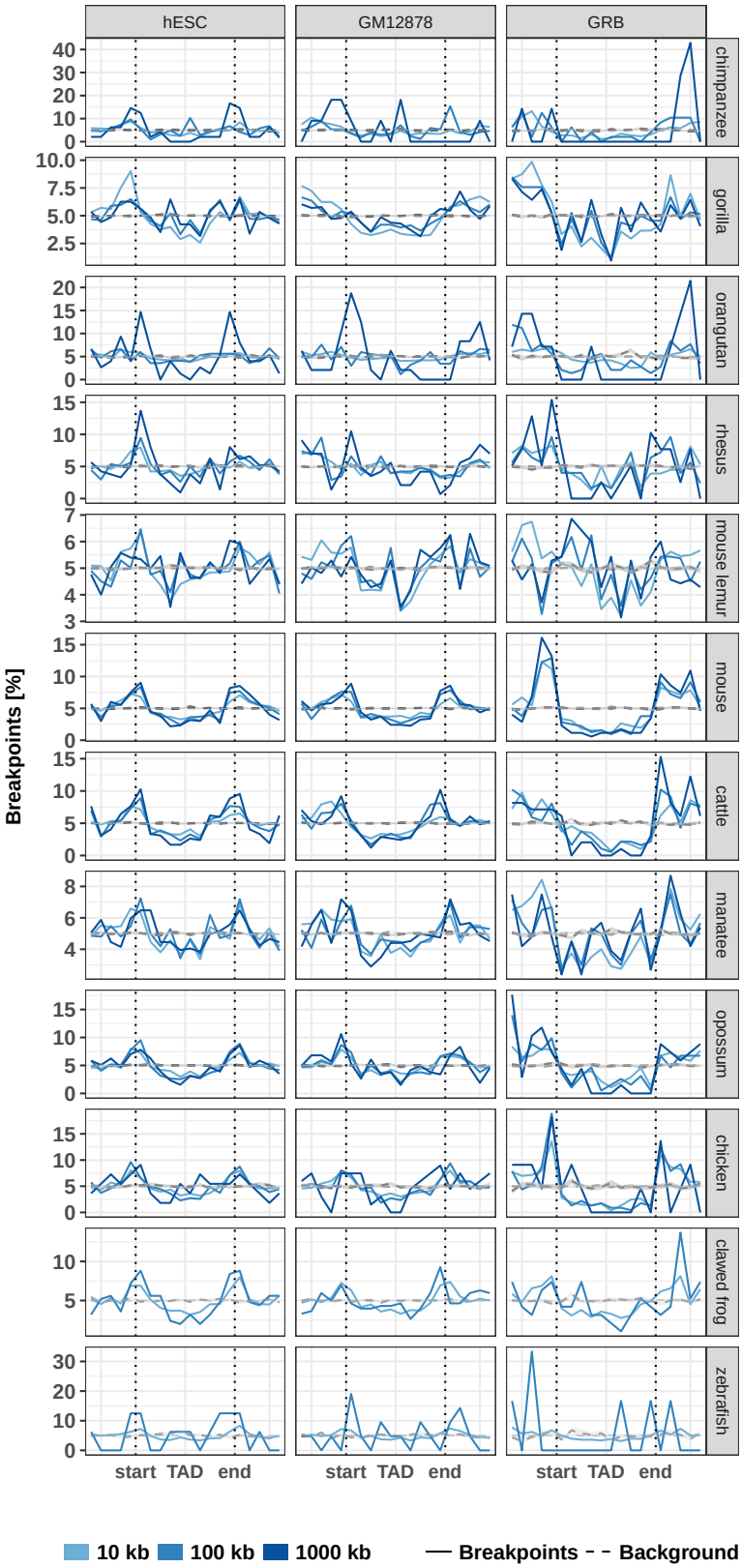

Figure B.1.: **Distribution of evolutionary rearrangement breakpoints between human and 12 vertebrate genomes around domains.** Relative breakpoint numbers from human and different species (horizontal panels) around hESC TADs (left), GM12878 contact domains (center), and GRBs (left). Blue color scale represents breakpoints from different fill-size thresholds. Dotted lines in gray show simulated background controls of randomly placed breakpoints.

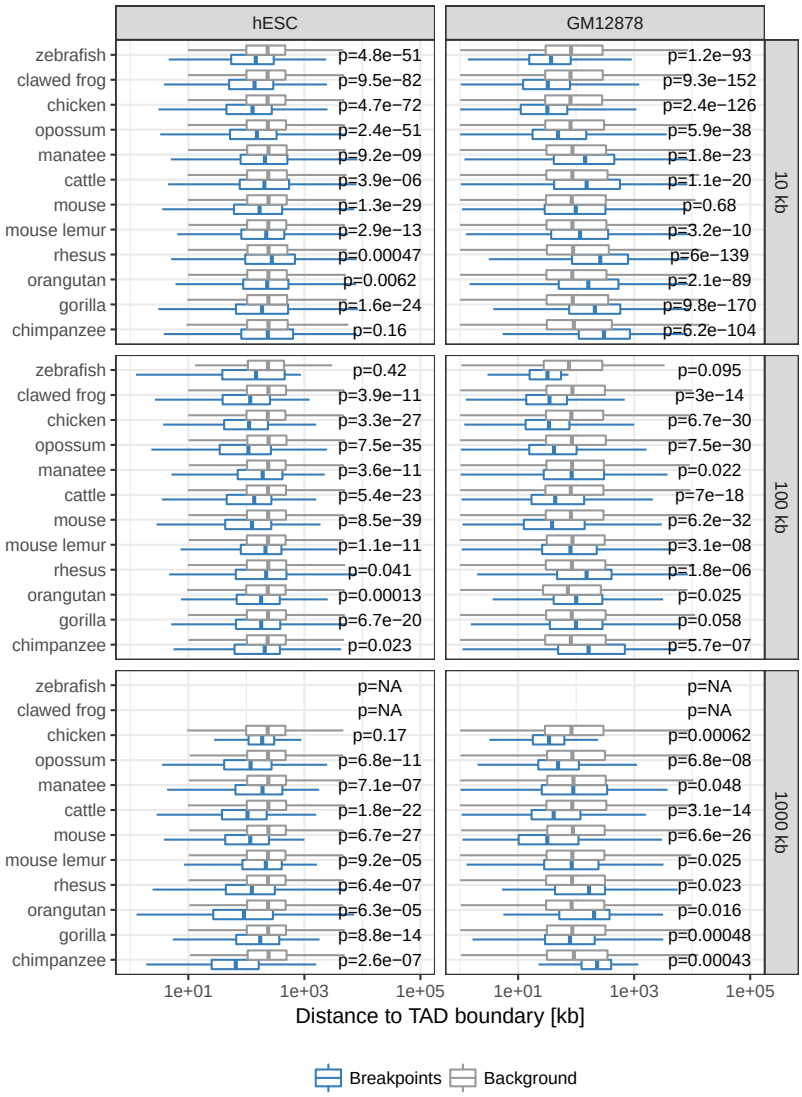

Figure B.2.: **Distance between rearrangement breakpoints and random controls to closest TAD boundary.** For each species (y-axis) and fill size threshold (vertical panels) the distances from all identified rearrangement breakpoints to its closest TAD boundary (x-axis) are compared between actual rearrangements (blue) and 100 times randomized background controls (gray). The left panel shows distances to next hESC TAD boundary and the right panel distances to closest GM12878 contact domain boundary. P-values according to Wilcoxon's rank-sum test.

Position Effect: Supplemental Data

C

Supplemental Note: Case Reports

DGAP017

46,X,t(X;10)(p11.2;q24.3)dn.arr(1-22,X)x2

Newborn female with a bicornuate uterus, diaphragmatic hernia, thenar hypoplasia, pulmonary hypoplasia, absent right olfactory lobe, loose skin, scoliosis, small thorax, hypoplastic labia, right clinodactyly and camptodactyly, as well as a scaphoid abdomen. This collection of features was reminiscent of Fryns syndrome (FRNS [MIM: 229850]). This case was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (GM00972) (Tang et al., 2013). An Affymetrix Genome-Wide Human SNP Array 6.0 performed at Coriell is reportedly normal.

DGAP111

46,XY,t(16;20)(q11.2;q13.2)dn.arr[hg18] 1q23.3(159763523_159905125)x3

Six-year-old male with congenital heart disease (one atrial septal defect, seven small ventricular septal defects), eye anomaly (Duane syndrome), poor growth, developmental delay, chronic constipation, left undescended testis, history of scoliosis (resolved), history of weak ankles and feet requiring braces (resolved), and asthma. Microarray analysis of DNA extracted from the DGAP111 EBV-transformed cell line contributed from DGAP to the NIGMS Human Genetic Cell Repository (GM22709, Coriell) was performed on the Affymetrix Genome-Wide Human SNP Array 6.0 and revealed a duplication of ~141.6 Kb in 1q23.3 (159763523-159905125) that was interpreted as likely benign.

DGAP113

46,XY,t(1;3)(q32.1;q13.2)dn

One-year-old male with bilateral congenital cataracts (TORCH screen, positive IgG and negative IgM for rubella, cytomegalovirus, herpes simplex virus; rubella virus isolation from urine and lens was negative), and mild developmental delay. Cranial magnetic resonance imaging (MRI) revealed prominent extra-axial cerebrospinal fluid spaces of uncertain significance, and the subject has marked macrocephaly

(head circumference >95th percentile) (Lachke et al., 2012). No microarray was performed.

DGAP126

46,XX,t(5;10)(p13.3;q21.1)dn.arr[hg18] 7q34(142030226_142154515)x1

Ten-year-old female with significant developmental delay with regression, autistic tendencies, and receptive and expressive language delay, disruptive behavior disorder, enuresis, dysthymia, sleep disturbance, self-injurious behaviors, and agitation. She had delays in gross and fine motor skills. No dysmorphic features were observed. Microarray analysis of DNA extracted from the DGAP126 EBV-transformed cell line contributed from DGAP to the NIGMS Human Genetic Cell Repository (GM18825, Coriell) was performed on the Affymetrix Genome-Wide Human SNP Array 6.0 and revealed a deletion of ~124.3 Kb in region 7q34 (142030226-142154515) that was interpreted to be benign.

DGAP138

46,XY,t(1;6)(q23;q13)dn.arr(1-22)x2,(X,Y)x1

Seven-year-old male with intellectual disability, fat distribution around trunk, gastroesophageal reflux, feeding problems (gastrostomy), seizure disorder, movement disorder (random, writhing type movements), wheelchair-dependence, Pierre-Robin sequence (mild micrognathia and cleft of the soft palate) (PRBNS [MIM: 261800]), microcephaly, pseudogynecomastia, and low growth hormone and high cortisone levels. Normal microarray results were reported from of DNA extracted from the DGAP138 EBV-transformed cell line contributed from DGAP to the NIGMS Human Genetic Cell Repository (GM20568, Coriell) on the Affymetrix Genome-Wide Human SNP Array 6.0.

DGAP153

46,X,t(X;17)(p11.23;p11.2)dn.arr(1-22,X)x2

Eight-year-old female with dysmorphic features (including mild synophrys, a flat philtrum and thin upper lip vermilion), mild developmental delay, sleep disturbance, and behavior problems (including temper tantrums, self-biting, and agitation). Deletion testing was negative for Smith-Magenis syndrome (SMS [MIM: 182290]). No cryptic aneusomies were reported to be detected by clinical aCGH. The DGAP153 EBV-transformed cell line was contributed to the NIGMS Human Genetic Cell Repository (GM20572, Coriell).

DGAP163

46,XY,t(2;14)(p23;q13)dn.arr(1-22)x2,(X,Y)x1

Four-year-old male with severe global developmental delay, absent speech, dysmorphic/distinctive facies, hypospadias (repaired), seizures as an infant (now seizure

free), myopia, nystagmus, small left retinal coloboma, and conductive hearing loss (history of otitis media). MRI showed periventricular white matter changes of unknown origin (no record of anoxic event), and recent electroencephalograms (EEGs) were normal. Fluorescence *in situ* hybridization (FISH) for SMS, DiGeorge syndrome (DGS [MIM: 188400]) and Velocardiofacial syndrome (VCFS [MIM: 192430]) was reportedly normal, as was aCGH using a 1M Agilent array with a resolution of 6.3 Kb.

DGAP176

46,Y,inv(X)(q13q24)mat

Four-year-old male with congenital, severe, bilateral sensorineural hearing loss, cognitive impairment, plagiocephaly, lax joints, and coordination difficulties. Dysmorphic features include macrocephaly, broad forehead, hypertelorism, downslanting palpebral fissures, epicanthic folds, flat midface, rounded nasal tip, flat nasal root, downturned corners of the mouth, simple helix of left ear, and full lips. He also had fifth finger clinodactyly and bridged palmar creases. No mutations were detected in the coding regions of gap junction protein beta 2 (*GJB2* [MIM: 121011]) or gap junction protein beta 6 (*GJB6* [MIM: 604418]). The mother is mosaic for inv(X)(q13q24) and 45,X but is reportedly healthy (Anger et al., 2014). No microarray was performed.

DGAP249

46,XX,t(2;11)(q33;q23)dn.arr(1-22,X)x2

Seven-year-old female with a history of global developmental delay. She has gross and fine motor delays, atypical oral motor skills and limited exploration of sensory materials. At four years she had an abnormal sleep-deprived EEG and increased bilateral electrocortical excitability; at six years EEG results were significantly abnormal with bifrontal symptoms consistent with epileptiform disturbance recorded in the interictal state. She has decreased visual motor integration, and a composite intellectual coefficient (IQ) of 71. Normal clinical microarray results were reported.

DGAP252

46,XY,t(3;18)(q13.2;q11.2)dn.arr(1-22)x2,(X,Y)x1

Four-month-old male whose prenatal course was complicated by polyhydramnios with an accompanying abnormal prenatal ultrasound and MRI, revealing an abnormal cerebellum, dilated cisterna magna, right lung apex cyst, intra-abdominal cysts and bilateral abnormal feet. Delivery was at term with two right posterior mediastinal cysts identified as a foregut duplication cyst and a bronchogenic cyst by pathology after surgical excision. Three ileal cysts were identified as duplication cysts with complete muscularis propria, small bowel/colon, and gastric oxyntic type

mucosa by pathologic examination after excision. Cerebellar hypoplasia was noted by MRI of his brain at one day of age. A wide anterior fontanelle (three finger widths) was observed, and his head was reportedly mildly turricephalic with a high forehead and a round bony protrusion of his skull at the occipital base. Normal clinical microarray results (CMA-HR + SNP (v.8.3)) were reported.

DGAP275

46,XX,t(7;12)(p13;q24.33)dn.arr(1-22,X)x2

Nine-year-old female with severe unexplained short stature (<4 SDs) and normal radiographs. An extensive endocrine workup revealed a normal growth hormone axis and no evidence of precocious puberty. She was non-dysmorphic and had normal cognitive development. A normal clinical Affymetrix Cytoscan SNP microarray was reported.

DGAP287

46,XY,t(10;14)(p13;q32.1)dn.arr(1-22)x2,(X,Y)x1

Four-year-old male with a history of global developmental delay and asymmetric spastic diplegia. He is ataxic, non-verbal, and drools frequently. He is non-dysmorphic, and a brain MRI was normal. Normal clinical Affymetrix Cytoscan HD SNP microarray results were reported.

DGAP288

46,XX,t(6;17)(q13;q21)dn.arr(1-22,X)x2

Prenatal case enrolled in study at 15 weeks, following ultrasound at 11 weeks revealing a cystic hygroma and chorionic villus sampling (CVS) at 12 weeks revealing the t(6;17) apparently balanced chromosome translocation. Normal clinical Affymetrix Cytoscan HD SNP microarray results were reported at 13 weeks. Micrognathia was seen on ultrasound at 18 weeks. At 19 weeks, DGAP sequencing results revealed no genes disrupted by the translocation, and the pregnancy was continued. Polyhydramnios and micrognathia were noted at 28 weeks. Fetal MRI at 34 weeks revealed a small jaw index consistent with micrognathia and retrognathia, glossoptosis, and cleft palate without cleft lip; findings were suspicious for PRBNS. Following delivery at 39 weeks, initial exams revealed a cleft palate. She was placed on continuous positive airway pressure, but otherwise was considered well.

DGAP315

46,XX,inv(6)(p24q11)dn.arr(1-22,X)x2

Fifteen-year-old female with severe static encephalopathy of unknown etiology. She uses a wheelchair, is microcephalic, nonverbal, and has severe generalized spasticity

with poorly controlled epilepsy. She had a normal echo and eye examination and reportedly normal aCGH results.

DGAP319

46,XX,t(4;13)(q31.3;q14.3)dn.arr(1-22,X)x2

Thirteen-year-old female with intellectual disability, and height, weight, and head circumference below the 3rd percentile. She has a grade II-IV systolic murmur, abnormal facies, finger and toe abnormalities. This case was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (GM00972) (Tang et al., 2013). This case was previously reported (Jenkins et al., 1975). The Affymetrix Genome-Wide Human SNP Array 6.0 performed at Coriell is reportedly normal.

DGAP322

46,XY,t(1;18)(q32.1;q22.1).arr(1-22)x2,(X,Y)x1

Male subject of unknown age with genitourinary malformations, third degree hypospadias, labialized scrotum with palpable descended testes, mild developmental delay, growth delay, and apparently intact hormonal axis. This case was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (GM16438) (Tang et al., 2013; Frizell et al., 1998). The Affymetrix Genome-Wide Human SNP Array 6.0 performed at Coriell is reportedly normal.

DGAP329

46,XX,t(2;14)(q21;q24.3)dn.arr[GRCh37/hg19] 18q22.3(72545050_72692202)x1 pat

Five-year-old female with a progressive neurologic disorder. She has nearly constant choreoathetosis, dystonia (including painful neck dystonia), and myoclonic movements, which are exacerbated by fatigue and emotional stress and are worsening with time. She is profoundly hypotonic and non-ambulatory. She is nonverbal but able to follow simple commands. She had a reported normal clinical CytoSure ISCA 8x60K v2.0 microarray, although a paternally inherited 150 Kb deletion at 18q22.3 from her phenotypically normal father was detected.

Supplemental Note: Nucleotide Level Nomenclature for DGAP karyotypes

Karyotypes of DGAP cases are described using a revised nomenclature that incorporates next-generation sequencing positions from Ordulu *et al.*, 2014.

DGAP017

46,X,t(X;10)(p11.2;q24.3)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(X;10)(10pter->10q25.1(107,711,256)::TATCCTTTG::Xp11.22(51,702,992)->Xpter;10qter->10q25.1(107,714,387)::GAGAAAAC::Xp11.22(51,707,815)->Xqter)dn

DGAP111

46,XY,t(16;20)(q11.2;q13.2)dn.arr[hg18] 1q23.3(159763523_159905125)x3.seq[GRCh37/hg19] (16,20)cx,der(16)(16pter->16q11.2(46,396,774)::16q11.2(46,397,625-46,397,900)::16q11.2(46,408,942-464093{69-70})::20q13.2(53,969,64{0-1}-53,970,162)::20q13.2(53,970,203)->20qter),der(20)(20pter->20q13.2(53,969,63{5-6})::16q11.2(46,403,29{1-2})->16qter)dn

DGAP113

46,XY,t(1;3)(q32.1;q13.2)dn.seq[GRCh37/hg19] t(1;3)(1pter->1q31.3(198,076,14{1})::3q13.13(110,275,>3qter;3pter->3q13.13(110,275,769)::AGAA::1q31.3(198,076,137)->1qter)dn

DGAP126

46,XX,t(5;10)(p13.3;q21.1)dn.arr[hg18] 7q34(142030226_142154515)x1.seq[GRCh37/hg19] t(5;10)(10qter->10q21.3(67,539,99{7-5})::5p13.3(29,658,44{0-2})->5qter;10pter->10q21.3(67,539,99{0})::5p13.3(29,658,42{6})->5pter)dn

DGAP138

46,XY,t(1;6)(q23;q13)dn.arr(1-22)x2,(X,Y)x1.seq[GRCh37/hg19] t(1;6)(1pter->1q31.2(193,491,602)::6q16.2(100,159,181)->6qter;6pter->6q16.2(100,159,182)::A::1q31.2(193,491,60>1qter)dn

DGAP153

46,X,t(X;17)(p11.23;p11.2)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(X;17)(17pter->17p11.2(20,682,69{0-1})::Xp11.3(44,372,16{4-5})->Xqter;17qter->17p11.2(20,682,68{7-4})::Xp11.3(44,372,1{72-69})->Xpter)dn

DGAP163

46,XY,t(2;14)(p23;q13)dn.arr(1-22)x2,(X,Y)x1.seq[GRCh37/hg19] t(2;14)(14qter->14q13(31,717,834)::G::2p23(39,206,240-39,206,384)::2p23(39,206,414)->2qter;14pter->14q13(31,717,73{3})::2p23(39,206,24{2})->2pter)dn

DGAP176

46,Y,inv(X)(q13q24)mat.seq[GRCh37/hg19] inv(X)(pter->q13(82,275,014)::ATCAATTTA::q24q13(108,129-82,320,86{7-5})::q24(108,149,24{9-7})->qter)mat

DGAP249

46,XX,t(2;11)(q33;q23)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(2;11)(2pter->2q33.1(199,943,78{1-9})::11q24.1(121,642,3{46-54})->11qter;11pter->11q24.1(121,638,616)::AGATCT::2q33.1(199,943,805)->2qter)dn

DGAP252

46,XY,t(3;18)(q13.2;q11.2)dn.arr(1-22)x2,(X,Y)x1.seq[GRCh37/hg19] t(3;18)(3pter->3q13.11(104,627,622)::TCAATACCTTTTA::18q11.2(19,498,398)->18qter;18pter->18q11.2(19,498,400)::AAAAATGGC::3q13.11(104,627,629)->3qter)dn

DGAP275

46,XX,t(7;12)(p13;q24.33)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(7;12)(12qter->12q24.33(132,983,131)::TC::7p12.3(46,111,841)->7qter;12pter->12q24.33(132,983,129)::7p12->7pter)dn

DGAP287

46,XY,t(10;14)(p13;q32.1)dn.arr(1-22)x2,(X,Y)x1.seq[GRCh37/hg19] t(10;14)(14qter->14q32.13(95,212,573)::AGTAAAGGGTTGGGTTAC::10p14(10,161,500-10,161,740)::TCG::10p14(->10qter;14pter->14q32.13(95,212,572)::TATCAG::10p14(10,161,498)->10pter)dn

DGAP288

46,XX,t(6;17)(q13;q21)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(6;17)(6pter->6q21(112,976,04{2-4})::17q24.3(69,728,01{7-9}->17qter;17pter->17q24.3(69,728,006)::CCCT->6qter)dn

DGAP315

46,XX,inv(6)(p24q11)dn.arr(1-22,X)x2.seq[GRCh37/hg19] inv(6)(qter->q11.1(63,115,715)::p24.3(63,115,685)::T::p24.3(9,394,994)->pter)dn

DGAP319

46,XX,t(4;13)(q31.3;q14.3)dn.arr(1-22,X)x2.seq[GRCh37/hg19] t(4;13)(4pter->4q32.2(161,913,247)::13q21.1(59,345,837)->13qter;13pter->13q21.1(59,345,83{5-6})::4q32.2(161,913,24{7-8})->4qter)dn

DGAP322

46,XY,t(1;18)(q32.1;q22.1)dn.arr(1-22)x2,(X,Y)x1.seq[GRCh37/hg19] t(1;18)(1pter->1q32.2(208,544,055)::ACTCCTCCAACTCCTATGTAGTTG::18q22.1(63,566,045)->18qter;18pter->18q22.1(63,566,053)::TACA::1q32.2(208,544,091)->1qter)dn

DGAP329

46,XX,t(2;14)(q21;q24.3)dn. arr[GRCh37/hg19] 18q22.3(72545050_72692202)x1 pat.seq[GRCh37/hg19] t(2;14)(2pter->2pter->2q22.3(145,110,93{6})::14q31.1(83,574,72{4})->14qter;14pter->14q31.1(83,574,71{5-9}::2q22.3(14,511,09{37-41}->2qter)dn

Supplemental Figure

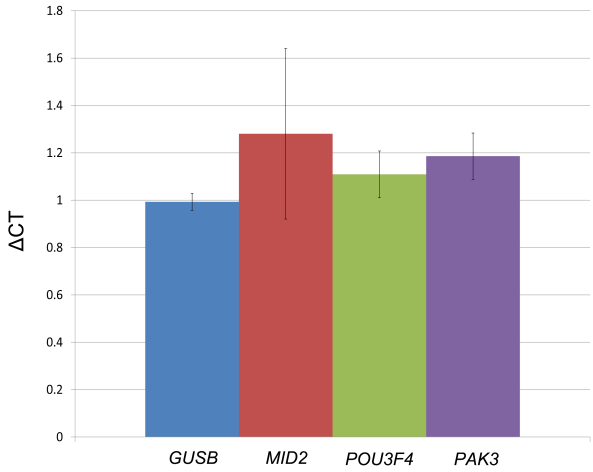

Figure C.1.: **Assessment of gene expression changes for DGAP176-derived LCLs.** Control gene expression is shown in blue and surveyed genes are marked in different colors. Each column represents the CT results of three culture replicates, with four technical replicates each, compared to three sex-matched control cell lines. Error bars indicate the standard deviation calculated from the biological replicates per gene.

Supplemental Table Legends

Supplemental Tables can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2017.06.011.

Table S1. Table describing the 17 cases with both breakpoints in non-coding regions. Case identifiers are provided per studied subject (Subject ID), in addition to their karyotypes using the International System for Human Cytogenetic Nomenclature (ISCN2016) and array information reported in hg19 unless otherwise stated in hg18. Each case has two reported breakpoints (A and B), and for each we provide cytogenetic band and nucleotide locations in hg19 coordinates for the derivative chromosomes involved in their generation (der(A) and der(B)). We also report the sequencing reads by which the breakpoints were identified, and the overlap with known annotated genes (Disrupted gene 1 and Disrupted Gene 2), as well as the two nearest genes (Closest Gene 1 and Closest Gene 2) and their distance in base pairs (bp) to the breakpoint locations (Distance to gene 1 and Distance to Gene 2) in the derivative chromosomes. Negative distance numbers indicate genes upstream of the breakpoint position, while positive numbers indicate genes located downstream of the breakpoint.

Table S2. Overlap of non-coding DGAP breakpoint positions with gene promoters. Table reporting the number of annotated Ensembl GRCh37 gene promoters (Ensembl_GRCh37_promoters) that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end).

Table S3. Overlap of non-coding DGAP breakpoint positions with transcription factor binding sites. Table reporting the number of annotated Ensembl GRCh37 transcription factor binding sites (Ensembl_GRCh37_tfbindingsites) that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end).

Table S4. Overlap of non-coding DGAP breakpoint positions with enhancers. Table reporting the number of primary cell (Primary_cell_enhancers), tissue (Tissue_enhancers), H1-ESC (ChromHMM_H1_ESC_enhancers), GM12878 (ChromHMM_GM12878_enhancers), and VISTA (VISTA_db_hg19) enhancers that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). Enhancer positions were obtained from Andersson *et al*., 2014, ENCODE, and the VISTA enhancer database human version hg19. Highlighted green rows indicate breakpoints which overlapped one or more of the enhancer categories analyzed.

Table S5. Overlap of non-coding DGAP breakpoint positions with DNaseI hypersensitive sites. Table reporting the number of DNaseI hypersensitive sites from H1-hESC, GM06990, GM12878, and the master table (a compilation of 125 cell lines DNaseI clusters) from ENCODE that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). Highlighted green rows indicate breakpoints which overlapped one or more of the DNaseI hypersensitive sites in the different cell lines analyzed.

Table S6. Overlap of non-coding DGAP breakpoint positions with CTCF binding sites. Table reporting the number of ENCODE CTCF binding sites from H1-hESC and GM12878 that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). Highlighted green rows indicate breakpoints which overlapped one or more of the CTCF binding sites in the two cell lines analyzed.

Table S7. Overlap of non-coding DGAP breakpoint positions with ENCODE chromatin state segments. Table reporting the ENCODE chromatin state segment classifications per non-coding DGAP breakpoint (DGAP id, chr, start, end) for H1-hESC and GM12878 cell lines. Chromatin state segment coordinates and other bed file information is displayed starting from column #bin until column itemRGB. Please refer to ENCODE's bed items description from here: http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgTables. Chromatin state names CTCF = CTCF binding site, E = enhancer, WE = weak enhancer, T = transcriptionally active, R = transcriptionally repressed.

Table S8. Overlap of non-coding DGAP breakpoint positions with repetitive elements. Table reporting the number of repetitive elements as assessed by Repeat Masker that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). Repetitive elements information such as coordinates (Rep_chr, Rep_start, Rep_end), name, class and family are provided for each overlap.

Table S9. Overlap of non-coding DGAP breakpoint positions with topologically associating domains (TADs). Table reporting the number of TADs in H1-hESC and IMR90 (Dixon *et al*., 2012) that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). TAD information such as coordinates (TAD_chr, TAD_start, TAD_end) are provided for each overlap.

Table S10. Overlap of non-coding DGAP breakpoint positions with high-resolution chromatin subcompartments and arrowhead domains. Table reporting the number of high-resolution chromatin subcompartments and arrowhead domains in IMR90 and GM12878 (Rao *et al*., 2014) that overlap non-coding DGAP breakpoints (DGAP id, chr, start, end). Chromatin subcompartments and arrowhead domains information such as coordinates and class are provided for each overlap.

Table S11. Disruption of chromatin contacts by non-coding DGAP breakpoint positions. Table reporting the number of chromatin contacts disrupted by non-coding DGAP breakpoint positions (DGAP id, chr, start, end) within their analysis windows (window_start, window_end) in Hi-C datasets of 20 and 40 Kb resolution of H1-hESC (Dixon *et al*., 2012) (Esc_20kb_HindIII_rep1, Esc_20kb_HindIII_rep2, Esc_40kb_hindIII_combined, Esc_40kb_hindIII_rep1, Esc_40kb_hindIII_rep2), 20 and 40 Kb resolution of IMR90 (Dixon *et al*., 2012) (IMR90_20kb_hindIII_rep1, IMR90_20kb_hindIII_rep2, IMR90_40kb_hindIII_combined, IMR90_40kb_hindIII_rep1, IMR90_40kb_hindIII_rep2), 100Kb and 1Mb resolution of GM06990 (http://epigenomegateway.wustl.edu/) (GM06990_obsexp_100kb, GM06990_obsexp_1mb) and looplists from Rao *et al*., 2014 for GM12878 and IMR90 (GSE63525_GM12878_primary+replicate_HiCCUPS_looplist, GSE63525_IMR90_HiCCUPS_looplist).

Table S12. Disruption of GM12878 chromatin contacts at various resolution levels by non-coding DGAP breakpoint positions. Table reporting the number of chromatin contacts disrupted by non-coding DGAP breakpoint positions (DGAP id, chr, start, end) within their analysis windows (window_start, window_end) in the 50Kb, 100Kb, 250Kb, 500Kb and 1Mb resolution Hi-C datasets from Rao *et al*., 2014 for GM12878.

Table S13. Disruption of predicted disrupted ENCODE distal DHS/enhancer–promoter connections by non-coding DGAP breakpoint positions. Table reporting the number of predicted ENCODE distal DHS/enhancer–promoter connections (Thurman *et al*., 2012) (promoter_DHS_chr, promoter_DHS_start, promoter_DHS_end, promoter_DHS_gene, distal_DHS_chr, distal_DHS_start, distal_DHS_end, promoter_distal_DHS_correlation) by non-coding DGAP breakpoint positions (DGAP id, chr, start, end) within their ś500 Kb analysis windows (window_start, window_end).

Table S14. Genes with predicted disrupted ENCODE distal DHS/enhancer–promoter connections by the non-coding DGAP breakpoint positions. Table reporting the names of genes (Genes) separated from their predicted enhancers (Disrupted_enh_prom_interactions) (Thurman *et al.*, 2012).

Table S15. Identification of genes with potential position effects. Table reporting the candidate genes (ensembl_gene_ID, Gene_chr, Gene_start, Gene_end, Gene_name) and their various lines of selection evidence for each non-coding DGAP breakpoint position (DGAP id, chr, start, end) within their analysis windows (window_start, window_end). Evidence lines include Hi-C domain inclusion (Hi_domain, HiC_chr, HiC_start, HiC_end), haploinsufficiency (HI_chr, Gene-start, gene_end, HI_prob, Haploinsufficiency_score,), triplosensitivity (Triplosensitivity_score), phenomatch score (PhenoScore, MaxPhenoScore, Phone_percentile, count_Pheno_percentile, MaxPheno_percentile, count_MaxPheno_percentile, Percentile_final_count). All of the evidence information is summarized (6Mb, 2Mb, TAD, DHS, Count_haplo, count_triplo) and the gene rankings are presented in the PERC+DHS+TAD+HAPLO+TRIPLO and PERC+DHS+2Mb+HAPLO+TRIPLO columns which take different evidence lines into considerarion. Green row highlight indicates highest ranking gene, and yellow row highlight indicates second best ranking genes.

Table S16. Translation of DGAP clinical features to HPO terms. Table reporting the HPO identifiers per DGAP case.

Table S17. Identification of genes with potential position effects for known pathogenic positive controls. Table reporting the candidate genes (ensembl_gene_ID, Gene_chr, Gene_start, Gene_end, Gene_name) and their various lines of selection evidence for the set of known pathogenic rearrangement positive controls (DGAP id, chr, start, end) within their analysis windows (window_start, window_end) from Redin *et al.*, 2017. Evidence lines include Hi-C domain inclusion (Hi_domain, HiC_chr, HiC_start, HiC_end), haploinsufficiency (HI_chr, Gene-start, gene_end, HI_prob, Haploinsufficiency_score,), triplosensitivity (Triplosensitivity_score), phenomatch score (PhenoScore, MaxPhenoScore, Phone_percentile, count_Pheno_percentile, MaxPheno_percentile, count_MaxPheno_percentile, Percentile_final_count). All of the evidence information is summarized (6Mb, 2Mb, TAD, DHS, Count_haplo, count_triplo) and the gene rankings are presented in the PERC+DHS+TAD+HAPLO+TRIPLO and PERC+DHS+2Mb+HAPLO+TRIPLO columns which take different evidence lines into considerarion. Yellow row highlight indicates pathogenic genes reported by Redin *et al.*, 2017.

Table S18. Identification of disrupted chromatin contacts between disrupted DHS and enhancers by the non-coding DGAP breakpoint positions. An agnostic search revealed the existence of chromatin contacts between breakpoint-disrupted sequences of DHS sites and gene enhancers in Hi-C data of H1-hESC cells at 40 Kb

resolution (Dixon *et al.*, 2012). The reported genes are our top position effect candidate genes in the region. Table columns report the candidate gene information (Gene_chr, Gene_start, Gene_end, Gene_name), the associated DGAP case information (DGAP_ID, DGAP_chr, DGAP_start, DGAP_end) and the disrupted Hi-C chromatin interaction (HiC_1_chr, HiC_1_start, HiC_1_end, HiC_2_chr, HiC_2_start, HiC_2_end, HiC_1_interaction).

Table S19. Overlap of non-coding DGAP breakpoint positions with DECIPHER cases. Table reporting the number of DECIPHER cases that overlap non-coding DGAP breakpoints (DGAP_ID, DGAP_chr, DGAP_start, DGAP_end). DECIPHER case information such as ID_patient, chr_start, chr_end, chr, mean_ratio, classification_type and phenotype are provided for each overlap.

Table S20. Genes contained within overlapped DECIPHER cases by non-coding DGAP breakpoint positions. Table reporting the number of genes contained within overlapped DECIPHER cases by the non-coding DGAP breakpoints (DGAP_ID, DGAP_chr, DGAP_start, DGAP_end). DECIPHER case and gene information such as gene_count, DECIPHER_ID, DECIPHER_chr, DECIPHER_start, DECIPHER_end, DECIPHER_value, DECIPHER_type_rearr, DECIPHER_phenotype and HG_symbol are provided for each overlapped DECIPHER case.

Table S21. DECIPHER cases overlapped by non-coding DGAP breakpoint positions that fulfilled non-coding selection criteria. Table reporting the number of DECIPHER cases that have non-coding breakpoints. DGAP comparison case information (DGAP_ID, DGAP_chr, DGAP_start, DGAP_end) is provided, as well as overlapped DECIPHER case information containing id_patient, chr_start, chr_end, chr, mean_ratio, classification_type and phenotype.

Table S22. Overlap of non-coding DGAP breakpoint positions with dbVar cases. Table reporting the number of dbVar cases that overlap non-coding DGAP breakpoints (DGAP_chr, DGAP_start, DGAP_end, DGAP_ID). dbVar case information such as dbVar ID, Start, End, Variant type, Gene, Molecular consequences, Most severe clinical significance, 1000G minor allele, 1000G MAF, GO-ESP minor allele, GO-ESP MAF, ExAC minor allele, ExAC MAF, Publications (PMIDs), Variant allele, Transcript change, RefSeq, Protein change, Molecular consequence, HGVS_c, HGVS_g, HGVS_ng, HGVS_p, Condition, Most severe clinical significance, Submitters, Highest review status and Last evaluated are provided for each overlap.

Loop prediction: Supplemental Information

D

Supplementary Tables

Table S1 Metadata of ChIP-seq experiments from ENCODE in human GM12878 cells with accession ID and download link. https://www.biorxiv.org/highwire/filestream/79233/field_highwire_adjunct_files/0/257584-1.tsv

Table S2 Metadata of ChIP-seq experiments from ENCODE human HeLa cells with accession ID and download link. https://www.biorxiv.org/highwire/filestream/79233/field_highwire_adjunct_files/1/257584-2.tsv

Table S3 Accession numbers and download URLs for data sets used in data type comparisons. https://www.biorxiv.org/highwire/filestream/79233/field_highwire_adjunct_files/2/257584-3.tsv

Supplementary Figures

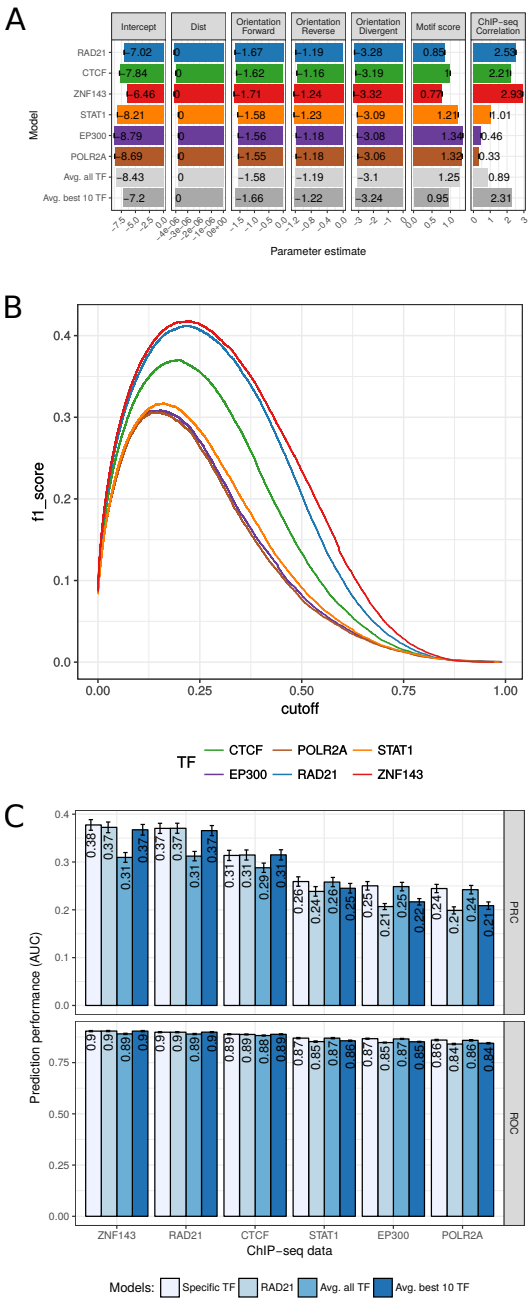

Figure D.1.: **7C model parameters and optimal cut-offs for binary prediction. (A)** Parameter values of the logistic regression model in 7C for different features (columns), separated for different models (rows). Average of model parameters of model training in 10-fold cross-validation is shown with error bars indicating the standard deviations. While the first six rows represent the models with the indicated TF ChIP-seq data and the genomic features, "Avg. all TF" is the average across all 124 TFs analyzed and "Avg. best 10 TF" is the average across the best ten performing TF models. **(B)** Prediction performance as f1 score (y-axis) for different cutoffs on the prediction probability p for the six selected models. **(C)** Prediction performance as auPRC (top) and auROC (bottom) of four different models (colors) on ChIP-seq data for six selected TFs (x-axis). 'Specific TF' is the model fitted using the ChIP-seq data indicated on the x-axis, 'RAD21' is the model trained on RAD21 ChIP-seq data, 'Avg. all TF' is a model averaged across all 124 models of analyzed TFs, and 'Avg. best 10 TF' is the averaged model across the 10 best performing models.

Contribution to individual publications

E

Table E.1.: Contribution to individual publications. Contributions in percent from all authors for each contribution role and publication. Author contribution definitions according to the CRediT Taxonomy (http://journals.plos.org/plosone/s/authorship).

Contributor Role	Role Definition
Conceptualization	Ideas; formulation or evolution of overarching research goals and aim
Data Curation	Management activities to annotate (produce metadata), scrub data ar
Formal Analysis	Application of statistical, mathematical, computational, or other form
Funding Acquisition	Acquisition of the financial support for the project leading to this pub
Investigation	Conducting a research and investigation process, specifically performi
Methodology	Development or design of methodology; creation of models
Project Administration	Management and coordination responsibility for the research activity
Resources	Provision of study materials, reagents, materials, patients, laboratory
Software	Programming, software development; designing computer programs;
Supervision	Oversight and leadership responsibility for the research activity plann
Validation	Verification, whether as a part of the activity or separate, of the overa
Visualization	Preparation, creation and/or presentation of the published work, spec
Writing Original Draft Preparation	Creation and/or presentation of the published work, specifically writi
Writing Review & Editing	Preparation, creation and/or presentation of the published work by th

Bibliography

Lelli, K. M., Slattery, M., and Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics*, 46:43–68, 2012. doi: 10.1146/annurev-genet-110711-155437.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K. B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi: 10.1038/nature11247.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518 (7539):317–30, 2015. doi: 10.1038/nature14248.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, a. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61, 2014. doi: 10.1038/nature12787.

Long, H., Prescott, S., and Wysocka, J. Ever-changing landscapes: Transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, 2016. doi: 10.1016/j.cell.2016.09.018.

Banerji, J., Rusconi, S., and Schaffner, W. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308, 1981. doi: 10.1016/0092-8674(81)90413-X.

Shlyueva, D., Stampfel, G., and Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(March):272–86, 2014. doi: 10.1038/nrg3682.

Song, L. and Crawford, G. E. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, 2010(2):pdb.prot5384, 2010. doi: 10.1101/pdb.prot5384.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013. doi: 10.1038/nmeth.2688.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–6, 2010. doi: 10.1073/pnas.1016071107.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L. A., Sloane-Stanley, J. A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R. J., Smith, A. J. H., Wood, W. G., and Higgs, D. R. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903, 2016. doi: 10.1038/ng.3605.

Will, A. J., Cova, G., Osterwalder, M., Chan, W.-L., Wittler, L., Brieske, N., Heinrich, V., de Villartay, J.-P., Vingron, M., Klopocki, E., Visel, A., Lupiáñez, D. G., and Mundlos, S. Composition and dosage of a multipartite enhancer cluster control developmental expression of ihh (indian hedgehog). *Nature Genetics*, (August), 2017. doi: 10.1038/ng.3939.

Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Molecular Cell*, 66(2):285–299.e5, 2017. doi: 10.1016/J.MOLCEL.2017.03.007.

Spitz, F. and Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 13(9):613–26, 2012. doi: 10.1038/nrg3207.

Andrey, G. and Mundlos, S. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 144(20):3646–3658, 2017. doi: 10.1242/dev.148304.

Bonev, B. and Cavalli, G. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661–678, 2016. doi: 10.1038/nrg.2016.112.

Stevens, T. J., Lando, D., Basu, S., Liam, P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., Shaughnessy-kirwan, A. O., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M. G. S., Lehner, B., Croce, L. D., Wutz, A., and Hendrich, B. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, pages 1–21, 2017. doi: 10.1038/nature21429.

Flyamer, I. M., Gassler, J., Imakaev, M., Ulyanov, S. V., Abdennur, N., Razin, S. V., Mirny, L., and Tachibana-Konwalski, K. Single-cell hi-c reveals unique chromatin reorganization at oocyte-tozygote transition. *Nature Publishing Group*, 544 (7648):1–17, 2017. doi: 10.1038/nature21711.

Sati, S. and Cavalli, G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126(1):33–44, 2017. doi: 10.1007/s00412-016-0593-6.

Schmitt, A. D., Hu, M., and Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, 2016. doi: 10.1038/nrm.2016.104.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–11, 2002. doi: 10.1126/science.1067799.

Hoffman, E. A., Frey, B. L., Smith, L. M., and Auble, D. T. Formaldehyde crosslinking: a tool for the study of chromatin complexes. *The Journal of biological chemistry*, 290(44):26404–11, 2015. doi: 10.1074/jbc.R115.651679.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403, 2013. doi: 10.1038/nrg3454.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation captureon-chip (4c). *Nature Genetics*, 38(11):1348–1354, 2006. doi: 10.1038/ng1896.

Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R. H., and de Laat, W. Variegated gene expression caused by cell-specific long-range dna interactions. *Nature Cell Biology*, 13(8):944–951, 2011. doi: 10.1038/ncb2278.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16 (10):1299–1309, 2006. doi: 10.1101/gr.5571506.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012. doi: 10.1038/nature11279.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009. doi: 10.1038/nature08497.

Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-l., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G., and Ruan, Y. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, pages 1–17, 2015. doi: 10.1016/j.cell.2015.11.024.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. a., Lander, E. S., and Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–93, 2009. doi: 10.1126/science.1181369.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. Topological domains in mammalian genomes identified by analy-

sis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. doi: 10.1038/nature11082.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80, 2014. doi: 10.1016/j.cell.2014.11.021.

Pombo, A. and Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, 12(March), 2015. doi: 10.1038/nrm3965.

Sexton, T. and Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell*, 160(6):1049–1059, 2015. doi: 10.1016/j.cell.2015.02.040.

Bouwman, B. A. and de Laat, W. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology*, 16(1):154, 2015. doi: 10.1186/s13059-015-0730-1.

Dekker, J. and Mirny, L. The 3d genome as moderator of chromosomal communication. *Cell*, 164(6):1110–1121, 2016. doi: 10.1016/j.cell.2016.02.007.

Dixon, J., Gorkin, D., and Ren, B. Chromatin domains: The unit of chromosome organization. *Molecular Cell*, 62(5):668–680, 2016. doi: 10.1016/j.molcel.2016.05.018.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-l., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., and Young, R. A. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, 351(6280):1454–8, 2016. doi: 10.1126/science.aad9024.

Merkenschlager, M. and Nora, E. P. Ctcf and cohesin in genome folding and transcriptional gene regulation. *Annual review of genomics and human genetics*, 17 (April):17–43, 2016. doi: 10.1146/annurev-genom-083115-022339.

Ruiz-Velasco, M. and Zaugg, J. B. Structure meets function: How chromatin organisation conveys functionality. *Current Opinion in Systems Biology*, 1(February): 129–136, 2017. doi: 10.1016/j.coisb.2017.01.003.

Cremer, T. and Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(April):292–301, 2001.

Branco, M. R. and Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, 4(5):780–788, 2006. doi: 10.1371/journal.pbio.0040138.

Roukos, V., Voss, T. C., Schmidt, C. K., Lee, S., Wangsa, D., and Misteli, T. Spatial dynamics of chromosome translocations in living cells. *Science*, 341(6146):660–664, 2013. doi: 10.1126/science.1237150.

Roukos, V. and Misteli, T. The biogenesis of chromosome translocations. *Nature cell biology*, 16(4):293–300, 2014. doi: 10.1038/ncb2941.

de Laat, W. and Grosveld, F. Inter-chromosomal gene regulation in the mammalian cell nucleus. *Current opinion in genetics & development*, 17(5):456–464, 2007. doi: 10.1016/j.gde.2007.07.009.

Monahan, K. and Lomvardas, S. Monoallelic expression of olfactory receptors. *Annual Review of Cell and Developmental Biology*, 31(1):annurev–cellbio–100814–125308, 2015. doi: 10.1146/annurev-cellbio-100814-125308.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–5, 2012. doi: 10.1038/nature11049.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–72, 2012. doi: 10.1016/j.cell.2012.01.010.

Ay, F. and Noble, W. S. Analysis methods for studying the 3d architecture of the genome. *Genome Biology*, 16(1):183, 2015. doi: 10.1186/s13059-015-0745-7.

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J., and Meyer, B. J. Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature*, 2015. doi: 10.1038/nature14450.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, 9(1): 14, 2014. doi: 10.1186/1748-7188-9-14.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh,

M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., Dostie, J., Pombo, A., and Nicodemi, M. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, 11:1–14, 2015. doi: 10.15252/msb.

Dali, R. and Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*, 45(6):2994–3005, 2016. doi: 10.1093/nar/gkx145.

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. Comparison of computational methods for hi-c data analysis. *Nature Methods*, (May):14–19, 2017. doi: 10.1038/nmeth.4325.

Cremer, T. and Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889, 2010. doi: 10.1101/cshperspect.a003889.

Gibcus, J. H. and Dekker, J. The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–82, 2013. doi: 10.1016/j.molcel.2013.02.011.

Eagen, K., Hartl, T., and Kornberg, R. Stable chromosome condensation revealed by chromosome conformation capture. *Cell*, 163(4):934–946, 2015. doi: 10.1016/j.cell.2015.10.026.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. a., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., and Gilbert, D. M. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, 2014. doi: 10.1038/nature13986.

Dileep, V., Ay, F., Sima, J., Vera, D. L., Noble, W. S., and Gilbert, D. M. Topologically associating domains and their long-range contacts are established during early g1 coincident with the establishment of the replication-timing program. *Genome research*, 25(8):1104–13, 2015. doi: 10.1101/gr.183699.114.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., and Ren, B. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–20, 2012. doi: 10.1038/nature11243.

Ghavi-Helm, Y., Klein, F. a., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. M. Enhancer loops appear stable during development and are

associated with paused polymerase. *Nature*, 512:96–100, 2014. doi: 10.1038/nature13417.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 24(3):390–400, 2014. doi: 10.1101/gr.163519.113.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome. *Cell*, 128(6):1231–1245, 2007. doi: 10.1016/J.CELL.2006.12.048.

Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., and Barta, E. Motif oriented high-resolution analysis of chip-seq data reveals the topological order of ctcf and cohesin proteins on dna. *BMC Genomics*, 17(1):637, 2016. doi: 10.1186/s12864-016-2940-7.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell reports*, 10(8):1297–309, 2015. doi: 10.1016/j.celrep.2015.02.004.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M., Ren, B., Krainer, A., Maniatis, T., and Wu, Q. Crispr inversion of ctcf sites alters genome topology and enhancer/promoter function. *Cell*, 162(4):900–910, 2015. doi: 10.1016/j.cell.2015.07.038.

de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. L., and de Laat, W. Ctcf binding polarity determines chromatin looping. *Molecular Cell*, 60(4):676–684, 2015. doi: 10.1016/j.molcel.2015.09.023.

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):201518552, 2015. doi: 10.1073/pnas.1518552112.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. Formation of chromosomal domains by loop extrusion. *Cell Reports*, 15(9):

2038–2049, 2016. doi: 10.1016/j.celrep.2016.04.085.

Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A. G., Flicek, P., Dekker, J., and Merkenschlager, M. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Research*, 23 (12):2066–2077, 2013. doi: 10.1101/gr.161620.113.

Zuin, J., Dixon, J. R., van der Reijden, M. I. J. a., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. a., van IJcken, W. F. J., Grosveld, F. G., Ren, B., and Wendt, K. S. Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3):996–1001, 2014. doi: 10.1073/pnas.1317788111.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S. M., Schroth, G. P., Tanay, A., and Hadjur, S. Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO journal*, 32(24):3119–29, 2013. doi: 10.1038/emboj.2013.237.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. a., and Dekker, J. Organization of the mitotic chromosome. *Science (New York, N.Y.)*, 342 (6161):948–53, 2013. doi: 10.1126/science.1236083.

Dekker, J. Two ways to fold the genome during the cell cycle: insights obtained with chromosome conformation capture. *Epigenetics & Chromatin*, 7(1):25, 2014. doi: 10.1186/1756-8935-7-25.

Chubb, J. R., Boyle, S., Perry, P., and Bickmore, W. A. Chromatin motion is constrained by association with nuclear compartments in human cells. *Current Biology*, 12(6):439–445, 2002. doi: 10.1016/S0960-9822(02)00695-4.

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., and Cremer, T. Chromosome order in hela cells changes during mitosis and early g1, but is stably maintained during subsequent interphase stages. *The Journal of cell biology*, 160(5):685–97, 2003. doi: 10.1083/jcb.200211103.

Deng, Y., Gao, L., Wang, B., and Guo, X. Hposim: An r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS ONE*, 10(2):1–12, 2015. doi: 10.1371/journal.pone.0115692.

Solovei, I., Wang, A. S., Thanisch, K., Schmidt, C. S., Krebs, S., Zwerger, M., Cohen, T. V., Devys, D., Foisner, R., Peichl, L., Herrmann, H., Blum, H., Engelkamp, D.,

Stewart, C. L., Leonhardt, H., and Joffe, B. Lbr and lamin a/c sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell*, 152(3): 584–98, 2013. doi: 10.1016/j.cell.2013.01.009.

Rutledge, M. T., Russo, M., Belton, J.-M., Dekker, J., and Broach, J. R. The yeast genome undergoes significant topological reorganization in quiescence. *Nucleic Acids Research*, page gkv723, 2015. doi: 10.1093/nar/gkv723.

Chandra, T., Ewels, P. A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S. W., Kirschner, K., Thuret, J. Y., Andrews, S., Fraser, P., and Reik, W. Global reorganization of the nuclear landscape in senescent cells. *Cell Reports*, 10(4):471–484, 2015. doi: 10.1016/j.celrep.2014.12.055.

Criscione, S. W., De Cecco, M., Siranosian, B., Zhang, Y., Kreiling, J. A., Sedivy, J. M., and Neretti, N. Reorganization of chromosome architecture in replicative cellular senescence. *Science advances*, 2(2):e1500882, 2016. doi: 10.1126/sciadv.1500882.

Joshi, O., Wang, S.-Y., Kuznetsova, T., Atlasi, Y., Peng, T., Fabre, P. J., Habibi, E., Shaik, J., Saeed, S., Handoko, L., Richmond, T., Spivakov, M., Burgess, D., and Stunnenberg, H. G. Dynamic reorganization of extremely long-range promoter-promoter interactions between two states of pluripotency. *Cell Stem Cell*, 17(6): 748–757, 2015. doi: 10.1016/J.STEM.2015.11.010.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. Polycomb-dependent regulatory contacts between distant hox loci in drosophila. *Cell*, 144(2):214–26, 2011. doi: 10.1016/j.cell.2010.12.026.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. a., and Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015. doi: 10.1038/nature14222.

Lin, Y. C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C. K., and Murre, C. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate b cell fate. *Nature Immunology*, 13(12):1196–1204, 2012. doi: 10.1038/ni.2432.

Le Dily, F., Bau, D., Pohl, a., Vicent, G. P., Serra, F., Soronellas, D., Castellano, G., Wright, R. H. G., Ballare, C., Filion, G., Marti-Renom, M. a., and Beato, M. Distinct structural transitions of chromatin topological domains correlate with

coordinated hormone-induced gene regulation. *Genes & Development*, 28(19): 2151–2162, 2014. doi: 10.1101/gad.241422.114.

Gómez-Marín, C., Tena, J. J., Acemel, R. D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., Bovolenta, P., Nobrega, M. a., Carvajal, J., and Gómez-Skarmeta, J. L. Evolutionary comparison reveals that diverging ctcf sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences*, 112(24): 201505463, 2015. doi: 10.1073/pnas.1505463112.

Hsieh, T.-H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O. Mapping nucleosome resolution chromosome folding in yeast by micro-c. *Cell*, pages 1–12, 2015. doi: 10.1016/j.cell.2015.05.048.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., and Grewal, S. I. S. Cohesin-dependent globules and heterochromatin shape 3d genome architecture in s. pombe. *Nature*, 2014. doi: 10.1038/nature13833.

Zhu, B., Zhang, W., Zhang, T., Liu, B., and Jiang, J. Genome-wide prediction and validation of intergenic enhancers in arabidopsis using open chromatin signatures. *The Plant cell*, 27(9):2415–26, 2015. doi: 10.1105/tpc.15.00537.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., and Wiehe, T. The chromatin insulator ctcf and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences*, 109(43):17507–17512, 2012. doi: 10.1073/pnas.1111941109.

Le, T. B. K., Imakaev, M. V., Mirny, L. a., and Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science (New York, N.Y.)*, 342(6159):731–4, 2013. doi: 10.1126/science.1242059.

Dowen, J., Fan, Z., Hnisz, D., Ren, G., Abraham, B., Zhang, L., Weintraub, A., Schuijers, J., Lee, T., Zhao, K., and Young, R. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, 2014. doi: 10.1016/j.cell.2014.09.030.

Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., and Reinberg, D. Transcription. ctcf establishes discrete functional chromatin domains at the hox clusters during differentiation. *Science (New York, N.Y.)*, 347(6225): 1017–21, 2015. doi: 10.1126/science.1262088.

Spielmann, M., Brancati, F., Krawitz, P. M., Robinson, P. N., Ibrahim, D. M., Franke, M., Hecht, J., Lohan, S., Dathe, K., Nardone, A. M., Ferrari, P., Landi, A., Wittler, L., Timmermann, B., Chan, D., Mennen, U., Klopocki, E., and Mundlos, S. Homeotic arm-to-leg transformation associated with genomic rearrangements at the pitx1 locus. *American journal of human genetics*, 91(4):629–35, 2012. doi: 10.1016/j.ajhg.2012.08.014.

Spielmann, M. and Mundlos, S. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *BioEssays : news and reviews in molecular, cellular and developmental biology*, pages 1–11, 2013. doi: 10.1002/bies.201200178.

Ibn-Salem, J., Köhler, S., Love, M. I., Chung, H.-R., Huang, N., Hurles, M. E., Haendel, M., Washington, N. L., Smedley, D., Mungall, C. J., Lewis, S. E., Ott, C.-E., Bauer, S., Schofield, P. N., Mundlos, S., Spielmann, M., and Robinson, P. N. Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biology*, 15(9):423, 2014. doi: 10.1186/s13059-014-0423-1.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C. M., Brown, D. L., Brudno, M., Campbell, J., Fitzpatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S.-M. M., Riggs, E. R., Scott, R. H., et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(Database issue):D966–74, 2014. doi: 10.1093/nar/gkt1026.

Firth, H. V., Richards, S. M., Bevan, a. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M., and Carter, N. P. Decipher: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *American journal of human genetics*, 84(4):524–33, 2009. doi: 10.1016/j.ajhg.2009.03.010.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015. doi: 10.1016/j.cell.2015.04.004.

Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F.,

Pombo, A., Vingron, M., Spitz, F., and Mundlos, S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, (Idi):1–15, 2016. doi: 10.1038/nature19800.

Redin, C., Brand, H., Collins, R. L., Kammin, T., Mitchell, E., Hodge, J. C., Hanscom, C., Pillalamarri, V., Seabra, C. M., Abbott, M.-A., Abdul-Rahman, O. A., Aberg, E., Adley, R., Alcaraz-Estrada, S. L., Alkuraya, F. S., An, Y., Anderson, M.-A., Antolik, C., Anyane-Yeboa, K., Atkin, J. F., Bartell, T., Bernstein, J. A., Beyer, E., Blumenthal, I., Bongers, E. M. H. F., Brilstra, E. H., Brown, C. W., Brüggenwirth, H. T., Callewaert, B., et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, 49(1):36–45, 2017. doi: 10.1038/ng.3720.

Northcott, P. a., Lee, C., Zichner, T., Stütz, A. M., Erkek, S., Kawauchi, D., Shih, D. J. H., Hovestadt, V., Zapatka, M., Sturm, D., Jones, D. T. W., Kool, M., Remke, M., Cavalli, F. M. G., Zuyderduyn, S., Bader, G. D., VandenBerg, S., Esparza, L. A., Ryzhova, M., Wang, W., Wittmann, A., Stark, S., Sieber, L., Seker-Cin, H., Linke, L., Kratochwil, F., Jäger, N., Buchhalter, I., Imbusch, C. D., et al. Enhancer hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature*, 2014. doi: 10.1038/nature13379.

Weischenfeldt, J., Dubash, T., Drainas, A. P., Mardin, B. R., Chen, Y., Stütz, A. M., Waszak, S. M., Bosco, G., Halvorsen, A. R., Raeder, B., Efthymiopoulos, T., Erkek, S., Siegl, C., Brenner, H., Brustugun, O. T., Dieter, S. M., Northcott, P. A., Petersen, I., Pfister, S. M., Schneider, M., Solberg, S. K., Thunissen, E., Weichert, W., Zichner, T., Thomas, R., Peifer, M., Helland, A., Ball, C. R., Jechlinger, M., et al. Pan-cancer analysis of somatic copy-number alterations implicates irs4 and igf2 in enhancer hijacking. *Nature Genetics*, (November), 2016. doi: 10.1038/ng.3722.

Spielmann, M. and Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics*, page ddw205, 2016. doi: 10.1093/hmg/ddw205.

Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39:309–338, 2005. doi: 10.1146/annurev.genet.39.073003.114725.

Makova, K. D. K. and Li, W.-H. W. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research*, 13(7):1638–1645, 2003. doi: 10.1101/gr.1133803.

Ptashne, M. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322(6081):697–701, 1986. doi: 10.1038/322697a0.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A., and Blobel, G. a. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–44, 2012. doi: 10.1016/j.cell.2012.03.051.

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. F., and Fraser, P. Long-range chromatin regulatory interactions in vivo. *Nat Genet*, 32(4):623–626, 2002. doi: 10.1038/ng1051.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., and de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6):1453–1465, 2002.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols*, 4(8):1184–1191, 2009. doi: 10.1038/nprot.2009.97.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. Biomart and bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005. doi: 10.1093/bioinformatics/bti525.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009. doi: 10.1101/gr.073585.107.

Galil, Z. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys*, 18(1):23–38, 1986. doi: 10.1145/6462.6502.

Lan, X. and Pritchard, J. K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science (New York, N.Y.)*, 352(6288):1009–13, 2016. doi: 10.1126/science.aad8411.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D., and Kent, W. J. The ucsc genome browser database: update 2006. *Nucleic Acids Res*, 34(Database issue):D590—-D598, 2006. doi: 10.1093/nar/gkj144.

Knight, P. a. and Ruiz, D. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33:1029–1047, 2013. doi: 10.1093/imanum/drs019.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. a., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. a., Fraser, P., Luscombe, N. M., and Osborne, C. S. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature Genetics*, 47(6), 2015. doi: 10.1038/ng.3286.

He, X. and Zhang, J. Gene complexity and gene duplicability. *Current Biology*, 15 (11):1016–1021, 2005. doi: 10.1016/j.cub.2005.04.035.

Newman, S., Hermetz, K. E., Weckselblatt, B., and Rudd, M. K. Next-generation sequencing of duplication cnvs reveals that most are tandem and some create fusion genes at breakpoints. *The American Journal of Human Genetics*, 96(2): 1–13, 2015. doi: 10.1016/j.ajhg.2014.12.017.

Djebali, S., Davis, C. a., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. a., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. doi: 10.1038/nature11233.

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Lassmann, T., Itoh, M., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. a., Carninci, P., and Hayashizaki, Y. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70, 2014. doi: 10.1038/nature13182.

Ardlie, K. G., DeLuca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., Ward, L. D., Kheradpour, P., Iriarte, B., Meng, Y., Palmer, C. D., Esko, T., Winckler, W., Hirschhorn, J. N., Kellis, M., MacArthur, D. G., Getz, G., Shabalin, A. A., Li, G., Zhou, Y. H., Nobel, A. B., Rusyn, I., Wright, F. A., Lappalainen, T., Ferreira, P. G., et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. doi: 10.1126/science.1262110.

Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A. M.-P., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H. E., and Brazma, A. Expression atlas updatean integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(October 2015):gkv1045, 2015. doi: 10.1093/nar/gkv1045.

Cremer, T., Cremer, M., Hübner, B., Strickfaden, H., Smeets, D., Popken, J., Sterr, M., Markaki, Y., Rippe, K., and Cremer, C. The 4d nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS Letters*, 2015. doi: 10.1016/j.febslet.2015.05.037.

Boutanaev, a. M., Kalmykova, a. I., Shevelyov, Y. Y., Nurminsky, D. I., Smith, M., Evolution, T., Masterpiece, T., Helm, C., and Selection, N. Large clusters of co-expressed genes in the drosophila genome. *Nature*, 420(December):666–669, 2002. doi: 10.1038/nature01191.1.

Purmann, A., Toedling, J., Schueler, M., Carninci, P., Lehrach, H., Hayashizaki, Y., Huber, W., and Sperling, S. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics*, 89(5):580–587, 2007. doi: 10.1016/j.ygeno.2007.01.010.

Sproul, D., Gilbert, N., and Bickmore, W. a. The role of chromatin structure in regulating the expression of clustered genes. *Nature reviews. Genetics*, 6(10): 775–781, 2005. doi: 10.1038/nrg1688.

Becker, M., Mah, N., Zdzieblo, D., Li, X., Mer, A., Andrade-navarro, M. A., and Mu, A. M. Epigenetic mechanisms in cellular reprogramming. In Meissner, A. and Walter, J., editors, *Epigenetics and Human Health*, Epigenetics and Human Health, pages pp 141–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. doi: 10.1007/978-3-642-31974-7.

Huminiecki, L. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Research*, 14(10a):1870–1879, 2004. doi: 10.1101/gr.2705204.

Rogozin, I. B., Managadze, D., Shabalina, S. A., and Koonin, E. V. Gene family level comparative analysis of gene expression n mammals validates the ortholog conjecture. *Genome Biology and Evolution*, 6(4):754–762, 2014. doi: 10.1093/gbe/evu051.

Schaefer, M. H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E. E., and Andrade-Navarro, M. a. Hippie: Integrating protein interaction networks with experiment based quality scores. *PloS one*, 7(2):e31826, 2012. doi: 10.1371/journal.pone.0031826.

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Blüthgen, N., Stadler, M., Tiana, G., Giorgetti, L., Bluthgen, N., Stadler, M., Tiana, G., and Giorgetti, L. Reciprocal insulation analysis of hi-c data shows that tads represent a functionally but not

structurally privileged scale in the hierarchical folding of chromosomes. *Genome Research*, 27(3):gr.212803.116, 2017. doi: 10.1101/gr.212803.116.

Nora, E. P., Dekker, J., and Heard, E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays*, 35(9): 818–828, 2013. doi: 10.1002/bies.201300040.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11484–11489, 2003. doi: 10.1073/pnas.1932072100.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002. doi: 10.1101/gr.229102. ArticlepublishedonlinebeforeprintinMay2002.

Mills, R. E., Bennett, E. A., Iskow, R. C., Luttig, C. T., Tsui, C., Pittard, W. S., and Devine, S. E. Recently mobilized transposons in the human and chimpanzee genomes. *The American Journal of Human Genetics*, 78(4):671–679, 2006. doi: 10.1086/501028.

Farré, M., Robinson, T. J., and Ruiz-Herrera, A. An integrative breakage model of genome architecture, reshuffling and evolution. *BioEssays*, pages n/a–n/a, 2015. doi: 10.1002/bies.201400174.

Polychronopoulos, D., King, J., Nash, A. J., Tan, G., and Lenhard, B. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Research*, (November):1–14, 2017. doi: 10.1093/nar/gkx1074.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., Mourrain, P., Ellingsen, S., Oates, A. C., Thisse, C., Thisse, B., Foucher, I., Adolf, B., Geling, A., Lenhard, B., and Becker, T. S. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, 17(5):545–555, 2007. doi: 10.1101/gr.6086307.

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. Topologically associating domains are ancient features that coincide with metazoan clusters of extreme noncoding conservation. *Nature Communications*, 8(1):441, 2017. doi: 10.1038/s41467-017-00524-5.

Engström, P. G., Sui, S. J. H., Drivenes, Ø., Becker, T. S., and Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research*, 17(12):1898–1908, 2007. doi: 10.1101/gr.6669607.

Dimitrieva, S. and Bucher, P. Ucnebasea database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research*, 41(D1):D101–D109, 2013. doi: 10.1093/nar/gks1092.

Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S. S., Huang, S.-c., Mckinnon, P. J., Aplan, P. D., Pommier, Y., Aiden, E. L., Casellas, R., and Nussenzweig, A. Genome organization drives chromosome fragility. *Cell*, pages 1–15, 2017. doi: 10.1016/j.cell.2017.06.034.

Ibn-Salem, J., Muro, E. M., and Andrade-Navarro, M. A. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Research*, 45(1):81–91, 2017. doi: 10.1093/nar/gkw813.

Schoenfelder, S., Furlan-magaril, M., Mifsud, B., Tavares-cadete, F., Sugar, R., Javierre, B.-m., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., Leproust, E., Osborne, C. S., Mitchell, J. A., Luscombe, N. M., and Fraser, P. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, pages 1–16, 2015. doi: 10.1101/gr.185272.114.Freely.

Montavon, T., Thevenet, L., and Duboule, D. Impact of copy number variations (cnvs) on long-range gene regulation at the hoxd locus. *Proceedings of the National Academy of Sciences of the United States of America*, 109(50):20204–11, 2012. doi: 10.1073/pnas.1217659109.

Zepeda-Mendoza, C. J., Ibn-Salem, J., Kammin, T., Harris, D. J., Rita, D., Gripp, K. W., MacKenzie, J. J., Gropman, A., Graham, B., Shaheen, R., Alkuraya, F. S., Brasington, C. K., Spence, E. J., Masser-Frye, D., Bird, L. M., Spiegel, E., Sparkes, R. L., Ordulu, Z., Talkowski, M. E., Andrade-Navarro, M. A., Robinson, P. N., and Morton, C. C. Computational prediction of position effects of apparently balanced human chromosomal rearrangements. *American journal of human genetics*, 101(2):206–217, 2017. doi: 10.1016/j.ajhg.2017.06.011.

Pevzner, P. and Tesler, G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7672–7, 2003. doi: 10.1073/pnas.1330369100.

Hou, C., Li, L., Qin, Z. S., and Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Molecular Cell*, 48(3):471–484, 2012. doi: 10.1016/j.molcel.2012.08.031.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'brien, S. J., Pevzner, P. A., and Lewin, H. A. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science (New York, N.Y.)*, 309(5734):613–7, 2005. doi: 10.1126/science.1111387.

Hinsch, H. and Hannenhalli, S. Recurring genomic breaks in independent lineages support genomic fragility. *BMC evolutionary biology*, 6:90, 2006. doi: 10.1186/1471-2148-6-90.

Acemel, R. D., Maeso, I., and GómezSkarmeta, J. L. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdisciplinary Reviews: Developmental Biology*, pages 1–19, 2017. doi: 10.1002/WDEV.265.

Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008. doi: 10.1016/J.CELL.2008.06.030.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016. doi: 10.1093/database/bav096.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12(2):115–121, 2015. doi: 10.1038/nmeth.3252.

Lawrence, M., Gentleman, R., and Carey, V. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, 2009. doi: 10.1093/bioinformatics/btp328.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013. doi: 10.1371/journal.pcbi.1003118.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4(8):1184–1191, 2009. doi: 10.1038/nprot.2009.97.

Wickham, H. and Grolemund, G. *R for data science : import, tidy, transform, visualize, and model data*.

LeJeune, J., Gautier, M., and Turpin, R. Study of somatic chromosomes from 9 mongoloid children. *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, 248(11):1721–2, 1959.

Ford, C., Jones, K., Polani, P., De Almeida, J., and Briggs, J. A sex-chromosome anomaly in a case of gonadal dysgenesis (turner's syndrome). *The Lancet*, 273 (7075):711–713, 1959. doi: 10.1016/S0140-6736(59)91893-8.

Jacobs, P. A. and Strong, J. A. A case of human intersexuality having a possible xxy sex-determining mechanism. *Nature*, 183(4657):302–303, 1959. doi: 10.1038/183302a0.

Stankiewicz, P. and Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82, 2002. doi: 10.1016/S0168-9525(02)02592-1.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., and Lee, C. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951, 2004. doi: 10.1038/ng1416.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Manér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. Large-scale copy number polymorphism in the human genome. *Science*, 305 (5683):525–528, 2004. doi: 10.1126/science.1098918.

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X., and Frazer, K. A. Common deletions and snps are in linkage disequilibrium in the human genome. *Nature Genetics*, 38 (1):82–85, 2006. doi: 10.1038/ng1695.

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., and Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics*, 38(1):75–81, 2006. doi: 10.1038/ng1697.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12, 2010. doi: 10.1038/nature08516.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M., and Snyder, M. Paired-end mapping reveals extensive structural variation in the human genome. *Science (New York, N.Y.)*, 318(5849):420–6, 2007. doi: 10.1126/science.1149504.

Stankiewicz, P. and Lupski, J. R. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010. doi: 10.1146/annurev-med-100708-204735.

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010. doi: 10.1038/nature09298.

Carvalho, C. M. B. and Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238, 2016. doi: 10.1038/nrg.2015.25.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481, 2009. doi: 10.1146/annurev.genom.9.081307.164217.

Theisen, A. and Shaffer, L. G. Disorders caused by chromosome abnormalities. *The application of clinical genetics*, 3:159–74, 2010. doi: 10.2147/TACG.S8884.

Nambiar, M. and Raghavan, S. C. How does dna break during chromosomal translocations? *Nucleic Acids Research*, 39(14):5813–5825, 2011. doi: 10.1093/nar/gkr223.

Higgins, A. W., Alkuraya, F. S., Bosco, A. F., Brown, K. K., Bruns, G. A., Donovan, D. J., Eisenman, R., Fan, Y., Farra, C. G., Ferguson, H. L., Gusella, J. F., Harris, D. J., Herrick, S. R., Kelly, C., Kim, H. G., Kishikawa, S., Korf, B. R., Kulkarni, S., Lally, E., Leach, N. T., Lemyre, E., Lewis, J., Ligon, A. H., Lu, W., Maas, R. L., MacDonald, M. E., Moore, S. D., Peters, R. E., Quade, B. J., et al. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *American Journal of Human Genetics*, 82(3):712–722, 2008. doi: 10.1016/j.ajhg.2008.01.011.

Kleinjan, D. A. and Van Heyningen, V. Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am. J. Hum. Genet*, 76:8–32, 2005. doi: 10.1086/426833.

Weiler, K. S. and Wakimoto, B. T. Heterochromatin and gene expression in drosophila. *Annual review of genetics*, 29:577–605, 1995. doi: 10.1146/annurev.ge.29.120195.003045.

Zhang, B. and Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19):201506257, 2015. doi: 10.1073/pnas.1506257112.

Fantes, J., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., Jones, S., Bickmore, W., Fukushima, Y., Mannens, M., Danes, S., van Heyningen, V., and Hanson, I. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Human Molecular Genetics*, 4(3):415–422, 1995. doi: 10.1093/hmg/4.3.415.

Kleinjan, D. A., Seawright, A., Schedl, A., Quinlan, R. A., Danes, S., and van Heyningen, V. Aniridia-associated translocations, dnase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of pax6. *Human Molecular Genetics*, 10(19):2049–2059, 2001. doi: 10.1093/hmg/10.19.2049.

Cai, J., Goodman, B. K., Patel, A. S., Mulliken, J. B., Van Maldergem, L., Hoganson, G. E., Paznekas, W. A., Ben-Neriah, Z., Sheffer, R., Cunningham, M. L., Daentl, D. L., and Jabs, E. W. Increased risk for developmental delay in saethre-chotzen syndrome is associated with twist deletions: an improved strategy for twist mutation screening. *Human Genetics*, 114(1):68–76, 2003. doi: 10.1007/s00439-003-1012-7.

Flomen, R. H., Vatcheva, R., Gorman, P. A., Baptista, P. R., Groet, J., Barišić, I., Ligutic, I., and Nižetić, D. Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by haploinsufficiency ofrieg, but also by chromosome breaks 90 kb upstream of this gene. *Genomics*, 47(3):409–413, 1998. doi: 10.1006/GENO.1997.5127.

Trembath, D. G., Semina, E. V., Jones, D. H., Patil, S. R., Qian, Q., Amendt, B. A., Russo, A. F., and Murray, J. C. Analysis of two translocation breakpoints and identification of a negative regulatory element in patients with rieger's syndrome. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 70(2):82–91, 2004. doi: 10.1002/bdra.10154.

Velagaleti, G. V., Bien-Willner, G. A., Northup, J. K., Lockhart, L. H., Hawkins, J. C., Jalal, S. M., Withers, M., Lupski, J. R., and Stankiewicz, P. Position effects due to chromosome breakpoints that map 900 kb upstream and 1.3 mb downstream of sox9 in two patients with campomelic dysplasia. *The American Journal of Human Genetics*, 76(4):652–662, 2005. doi: 10.1086/429252.

Kleinjan, D. and van Heyningen, V. Position effect in human genetic disease. *Human Molecular Genetics*, 7(10):1611–1618, 1998. doi: 10.1093/hmg/7.10.1611.

Lupski, J. R. and Stankiewicz, P. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics*, 1(6):e49, 2005. doi: 10.1371/journal.pgen.0010049.

de Wit, E. and de Laat, W. A decade of 3c technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012. doi: 10.1101/gad.179804.111.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. a., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J., and Corces, V. G. Architectural protein subclasses shape 3d organization of genomes during lineage commitment. *Cell*, 153(6):1281–95, 2013. doi: 10.1016/j.cell.2013.04.053.

Gröschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A. M., Erpelinck, C., van der Velden, V. H. J., Havermans, M., Avellino, R., van Lom, K., Rombouts, E. J., van Duin, M., Döhner, K., Beverloo, H. B., Bradner, J. E., Döhner, H., Löwenberg, B., Valk, P. J. M., Bindels, E. M. J., de Laat, W., and Delwel, R. A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell*, 157(2):369–381, 2014. doi: 10.1016/j.cell.2014.02.019.

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P.-A., Hsu, Y.-H., Drucker, D. J., Mellgren, G., Hui, C.-C., Hauner, H., and Kellis, M. fto obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine*, 373(10):895–907, 2015. doi: 10.1056/NEJMoa1502214.

Visser, M., Kayser, M., and Palstra, R.-J. Herc2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the oca2 promoter. *Genome research*, 22(3):446–55, 2012. doi: 10.1101/gr.128652.111.

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., Stahl, E. A., Georgakopoulos, A., Ruderfer, D. M., Charney, A., Okada, Y., Siminovitch, K. A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Plenge, R. M., Raychaudhuri, S., Fromer, M., Purcell, S. M., Brennand, K. J., Robakis, N. K., Schadt, E. E., Akbarian, S., and Sklar, P. A role for noncoding variation in schizophrenia. *Cell reports*, 9(4):1417–29, 2014. doi: 10.1016/j.celrep.2014.10.015.

Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., Gasparini, L., Ferrera, D., Canale, C., Guipponi, M., Pennacchio, L. A., Antonarakis, S. E., Brussino, A., and Brusco, A. A large genomic deletion leads to enhancer adoption by the lamin b1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (adld). *Human Molecular Genetics*, 24(11):3143–3154, 2015. doi: 10.1093/hmg/ddv065.

Oldridge, D. A., Wood, A. C., Weichert-Leahey, N., Crimmins, I., Sussman, R., Winter, C., McDaniel, L. D., Diamond, M., Hart, L. S., Zhu, S., Durbin, A. D., Abraham, B. J., Anders, L., Tian, L., Zhang, S., Wei, J. S., Khan, J., Bramlett, K., Rahman, N., Capasso, M., Iolascon, A., Gerhard, D. S., Guidry Auvil, J. M., Young, R. A., Hakonarson, H., Diskin, S. J., Thomas Look, A., and Maris, J. M. Genetic predisposition to neuroblastoma mediated by a lmo1 super-enhancer polymorphism. *Nature*, 528(7582):418–421, 2015. doi: 10.1038/nature15540.

Ordulu, Z., Kammin, T., Brand, H., Pillalamarri, V., Redin, C. E., Collins, R. L., Blumenthal, I., Hanscom, C., Pereira, S., Crandall, B. F., Gerrol, P., Hayden, M. A., Hussain, N., Kanengisser-pines, B., Kantarci, S., Levy, B., Macera, M. J., Quintero-rivera, F., Spiegel, E., Stevens, B., Ulm, J. E., Warburton, D., Wilkins-haug, L. E., Yachelevich, N., Gusella, J. F., and Talkowski, M. E. Structural chromosomal rearrangements require nucleotide-level resolution : Lessons from next-generation

sequencing in prenatal diagnosis. *The American Journal of Human Genetics*, pages 1–19, 2016. doi: 10.1016/j.ajhg.2016.08.022.

Ligon, A. H., Moore, S. D., Parisi, M. A., Mealiffe, M. E., Harris, D. J., Ferguson, H. L., Quade, B. J., and Morton, C. C. Constitutional rearrangement of the architectural factor hmga2: A novel human phenotype including overgrowth and lipomas. *The American Journal of Human Genetics*, 76(2):340–348, 2005. doi: 10.1086/427565.

Kim, W. K. and Marcotte, E. M. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS computational biology*, 4(11):e1000232, 2008. doi: 10.1371/journal.pcbi.1000232.

Lu, W., Quintero-Rivera, F., Fan, Y., Alkuraya, F. S., Donovan, D. J., Xi, Q., Turbe-Doan, A., Li, Q.-G., Campbell, C. G., Shanske, A. L., Sherr, E. H., Ahmad, A., Peters, R., Rilliet, B., Parvex, P., Bassuk, A. G., Harris, D. J., Ferguson, H., Kelly, C., Walsh, C. A., Gronostajski, R. M., Devriendt, K., Higgins, A., Ligon, A. H., Quade, B. J., Morton, C. C., Gusella, J. F., and Maas, R. L. Nfia haploinsufficiency is associated with a cns malformation syndrome and urinary tract defects. *PLoS Genetics*, 3(5):e80, 2007. doi: 10.1371/journal.pgen.0030080.

Talkowski, M., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., Kirby, A., Liu, S., Muddukrishna, B., Ohsumi, T., Shen, Y., Borowsky, M., Daly, M., Morton, C., and Gusella, J. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *The American Journal of Human Genetics*, 88(4):469–481, 2011. doi: 10.1016/J.AJHG.2011.03.013.

Quinn, J. J. and Chang, H. Y. Unique features of long non-coding rna biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, 2016. doi: 10.1038/nrg.2015.10.

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., and Carter, D. R. F. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA (New York, N.Y.)*, 17(5):792–8, 2011. doi: 10.1261/rna.2658311.

Muro, E. M. and Andrade-Navarro, M. A. Pseudogenes as an alternative source of natural antisense transcripts. *BMC Evolutionary Biology*, 10(1):338, 2010. doi: 10.1186/1471-2148-10-338.

Ordulu, Z., Wong, K., Currall, B., Ivanov, A., Pereira, S., Althari, S., Gusella, J., Talkowski, M., and Morton, C. Describing sequencing results of structural chro-

mosome rearrangements with a suggested next-generation cytogenetic nomenclature. *The American Journal of Human Genetics*, 94(5):695–709, 2014. doi: 10.1016/J.AJHG.2014.03.020.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., et al. Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755, 2014. doi: 10.1093/nar/gkt1196.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–27, 2011. doi: 10.1101/gad.17446611.

Huang, N., Lee, I., Marcotte, E. M., and Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genetics*, 6(10):e1001154, 2010. doi: 10.1371/journal.pgen.1001154.

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., Plon, S. E., Ramos, E. M., Sherry, S. T., and Watson, M. S. Clingen the clinical genome resource. *New England Journal of Medicine*, 372(23):2235–2242, 2015. doi: 10.1056/NEJMsr1406261.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012. doi: 10.1038/nature11232.

Flavahan, W. a., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, a. S., Stemmer-Rachamimov, a. O., Suva, M. L., and Bernstein, B. E. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016. doi: 10.1038/nature16490.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. a. Vista enhancer browser–a database of tissue-specific human enhancers. *Nucleic acids research*, 35(Database issue):D88–92, 2007. doi: 10.1093/nar/gkl822.

Zhou, X. and Wang, T. Using the wash u epigenome browser to examine genome-wide sequencing data. In *Current Protocols in Bioinformatics*, volume Chapter 10, page Unit10.10. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012. doi: 10.1002/0471250953.bi1010s40.

Quinlan, A. R. and Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi: 10.1093/bioinformatics/btq033.

Heger, A., Webber, C., Goodson, M., Ponting, C. P., and Lunter, G. Gat: a simulation framework for testing the association of genomic intervals. *Bioinformatics (Oxford, England)*, 29(16):2046–8, 2013. doi: 10.1093/bioinformatics/btt343.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., and Church, D. M. dbvar and dgva: public archives for genomic structural variation. *Nucleic Acids Research*, 41(D1):D936–D941, 2012. doi: 10.1093/nar/gks1213.

Gu, W., Zhang, F., and Lupski, J. R. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4, 2008. doi: 10.1186/1755-8417-1-4.

Cardoso, A. R., Oliveira, M., Amorim, A., and Azevedo, L. Major influence of repetitive elements on disease-associated copy number variants (cnvs). *Human Genomics*, 10(1):30, 2016. doi: 10.1186/s40246-016-0088-9.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., et al. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, 2013. doi: 10.1038/ng.2653.

Vetro, A., Dehghani, M. R., Kraoua, L., Giorda, R., Beri, S., Cardarelli, L., Merico, M., Manolakos, E., Parada-Bustamante, A., Castro, A., Radi, O., Camerino, G., Brusco, A., Sabaghian, M., Sofocleous, C., Forzano, F., Palumbo, P., Palumbo, O., Calvano, S., Zelante, L., Grammatico, P., Giglio, S., Basly, M., Chaabouni, M., Carella, M., Russo, G., Bonaglia, M. C., and Zuffardi, O. Testis development in the absence of sry: chromosomal rearrangements at sox9 and sox3. *European Journal of Human Genetics*, 23(8):1025–1032, 2015. doi: 10.1038/ejhg.2014.237.

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K.,

Krantz, I. D., Kuhn, R. M., Lee, C., Ostell, J. M., Rosenberg, C., Scherer, S. W., Spinner, N. B., Stavropoulos, D. J., Tepperberg, J. H., Thorland, E. C., Vermeesch, J. R., Waggoner, D. J., et al. Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *The American Journal of Human Genetics*, 86(5):749–764, 2010. doi: 10.1016/J.AJHG.2010.04.006.

Pachlopnik Schmid, J., Lemoine, R., Nehme, N., Cormier-Daire, V., Revy, P., Debeurme, F., Debré, M., Nitschke, P., Bole-Feysot, C., Legeai-Mallet, L., Lim, A., de Villartay, J.-P., Picard, C., Durandy, A., Fischer, A., and de Saint Basile, G. Polymerase ϵ1 mutation in a human syndrome with facial dysmorphism, immunodeficiency, livedo, and short stature ("fils syndrome"). *The Journal of experimental medicine*, 209(13):2323–30, 2012. doi: 10.1084/jem.20121303.

Krijger, P. H. L. and de Laat, W. Regulation of disease-associated gene expression in the 3d genome. *Nature Reviews Molecular Cell Biology*, 2016. doi: 10.1038/nrm.2016.138.

Mora, A., Sandve, G. K., Gabrielsen, O. S., and Eskeland, R. In the loop: promoter-enhancer interactions and bioinformatics. *Briefings in bioinformatics*, (July):1–16, 2015. doi: 10.1093/bib/bbv097.

Tiwari, V. K., Cope, L., McGarvey, K. M., Ohm, J. E., and Baylin, S. B. A novel 6c assay uncovers polycomb-mediated higher order chromatin conformations. *Genome research*, 18(7):1171–9, 2008. doi: 10.1101/gr.073452.107.

Sati, S. and Cavalli, G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 2016. doi: 10.1007/s00412-016-0593-6.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., Wiegers, J., Wiegers, T. C., and Mattingly, C. J. The comparative toxicogenomics database: Update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, 2017. doi: 10.1093/nar/gkw838.

Orlando, V., Strutt, H., and Paro, R. Analysis of chromatin structure byin vivo-formaldehyde cross-linking. *Methods*, 11(2):205–214, 1997. doi: 10.1006/METH.1996.0407.

Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., Meijsing, S. H., Hernandez, C., Love, M. I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., and Meijsing, S. H. Chip-exo signal associated with dna-binding motifs provide insights into the genomic bind-

ing of the glucocorticoid receptor and cooperating transcription factors. *Genome research*, 25(6):825–35, 2015. doi: 10.1101/gr.185157.114.

Khan, A., Fornes, O., and Stigliani, A. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, (In print)(November 2017):1–7, 2018. doi: 10.1093/nar/gkx1126.

Saito, T. and Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10 (3):1–21, 2015. doi: 10.1371/journal.pone.0118432.

Ye, B.-Y., Shen, W.-L., Wang, D., Li, P., Zhang, Z., Shi, M.-L., Zhang, Y., Zhang, F.-X., and Zhao, Z.-H. Znf143 is involved in ctcf-mediated chromatin interactions by cooperation with cohesin and other partners. *Molecular Biology*, 50(3):431–437, 2016. doi: 10.1134/S0026893316030031.

Bailey, S. D., Zhang, X., Desai, K., Aid, M., Corradin, O., Cowper-Sal Lari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-Kains, B., Lupien, M., Cowper-Sal·lari, R., Akhtar-Zaidi, B., Scacheri, P. C., Haibe-Kains, B., and Lupien, M. Znf143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications*, 2:6186, 2015. doi: 10.1038/ncomms7186.

Zhang, B., Park, B.-H., Karpinets, T., and Samatova, N. F. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics (Oxford, England)*, 24(7):979–86, 2008. doi: 10.1093/bioinformatics/btn036.

Hansen, P., Hecht, J., Ibrahim, D. M., Krannich, A., Truss, M., and Robinson, P. N. Saturation analysis of chip-seq data for reproducible identification of binding peaks. *Genome research*, 25(9):1391–400, 2015. doi: 10.1101/gr.189894.115.

Pique-Regi, R., Degner, J. F., Pai, A. a., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. *Genome research*, 21(3):447–55, 2011. doi: 10.1101/gr.112623.110.

Yardmc, G. G., Frank, C. L., Crawford, G. E., and Ohler, U. Explicit dnase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic acids research*, pages 1–14, 2014. doi: 10.1093/nar/gku810.

Rhee, H. S. and Pugh, B. F. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–19, 2011. doi: 10.

1016/j.cell.2011.11.013.

He, Q., Johnston, J., and Zeitlinger, J. Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(March 2014):395–401, 2015. doi: 10.1038/nbt.3121.

Sheffield, N. C., Thurman, R. E., Song, L., Safi, A., Stamatoyannopoulos, J. a., Lenhard, B., Crawford, G. E., and Furey, T. S. Patterns of regulatory activity across diverse human cell types predict tissue identity , transcription factor binding , and long-range interactions predict tissue identity , transcription factor binding , and long-range interactions. *Genome Research*, 23:777–788, 2013. doi: 10.1101/gr.152140.112.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., and Cohen, D. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database*, 2017(1):1665–1680, 2017. doi: 10.1093/database/bax028.

O'Connor, T. R. and Bailey, T. L. Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. *Nucleic acids research*, pages 1–11, 2014. doi: 10.1093/nar/gku801.

Roy, S., Siahpirani, A. F., Chasman, D., Knaack, S., Ay, F., Stewart, R., Wilson, M., and Sridharan, R. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Research*, 43(18):8694–8712, 2015. doi: 10.1093/nar/gkv865.

Whalen, S., Truty, R. M., and Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, 48(5):1–10, 2015. doi: 10.1038/ng.3539.

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., Ding, B., Li, N., Zheng, L., and Wang, W. Constructing 3d interaction maps from 1d epigenomes. *Nature Communications*, 7:10812, 2016. doi: 10.1038/ncomms10812.

Schreiber, J., Libbrecht, M., Bilmes, J., and Noble, W. Nucleotide sequence and dnasei sensitivity are predictive of 3d chromatin architecture. *bioRxiv*, (December 2016):1–15, 2017. doi: 10.1101/103614.

Dzida, T., Iqbal, M., Charapitsa, I., Reid, G., Stunnenberg, H., Matarese, F., Grote, K., Honkela, A., and Rattray, M. Predicting stimulation-dependent enhancer-promoter interactions from chip-seq time course data. *PeerJ*, 5:e3742, 2017. doi: 10.7717/peerj.3742.

Zhao, C., Li, X., and Hu, H. Petmodule: a motif module based approach for enhancer target gene prediction. *Scientific Reports*, 6(July):30043, 2016. doi: 10.1038/srep30043.

Naville, M., Ishibashi, M., Ferg, M., Bengani, H., Rinkwitz, S., Krecsmarik, M., Hawkins, T. a., Wilson, S. W., Manning, E., Chilamakuri, C. S. R., Wilson, D. I., Louis, A., Lucy Raymond, F., Rastegar, S., Strähle, U., Lenhard, B., Bally-Cuif, L., van Heyningen, V., FitzPatrick, D. R., Becker, T. S., and Roest Crollius, H. Long-range evolutionary constraints reveal cis-regulatory interactions on the human x chromosome. *Nature Communications*, 6(May 2014):6904, 2015. doi: 10.1038/ncomms7904.

Nikumbh, S. and Pfeifer, N. Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. *BMC Bioinformatics*, 18(1):218, 2017. doi: 10.1186/s12859-017-1624-x.

Brackley, C. A., Brown, J. M., Waithe, D., Babbs, C., Davies, J., Hughes, J. R., Buckle, V. J., and Marenduzzo, D. Predicting the three-dimensional folding of cis -regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biology*, pages 31–36, 2016. doi: 10.1186/s13059-016-0909-0.

Chen, Y., Wang, Y., Xuan, Z., Chen, M., and Zhang, M. Q. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Research*, 44(11):e106–e106, 2016. doi: 10.1093/nar/gkw225.

Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., Wang, P., Ruan, Y., and Corces, V. G. Evolutionarily conserved principles predict 3d chromatin organization. *Molecular Cell*, pages 1–16, 2017. doi: 10.1016/j.molcel.2017.07.022.

Oti, M., Falck, J., Huynen, M. A., and Zhou, H. Ctcf-mediated chromatin loops enclose inducible gene regulatory domains. *BMC Genomics*, 17(1):252, 2016. doi: 10.1186/s12864-016-2516-6.

Liang, J., Lacroix, L., Gamot, A., Cuddapah, S., Queille, S., Lhoumaud, P., Lepetit, P., Martin, P. G. P., Vogelmann, J., Court, F., Hennion, M., Micas, G., Urbach, S., Bouchez, O., Nöllmann, M., Zhao, K., Emberly, E., and Cuvier, O. Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and pol ii pausing. *Molecular Cell*, 53(4):672–681, 2014. doi: 10.1016/j.molcel.2013.12.029.

Mourad, R., Li, L., and Cuvier, O. Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach. *PLOS Computational Biology*, 13(5):e1005538, 2017. doi: 10.1371/journal.pcbi.1005538.

Bonev, B., Cohen, N. M., Szabo, Q., Hugnot, J.-p., Tanay, A., Cavalli, G., Bonev, B., Cohen, N. M., Szabo, Q., Fritsch, L., Papadopoulos, G. L., and Lubling, Y. Multiscale 3d genome rewiring during mouse article multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3):557.e1–557.e24, 2017. doi: 10.1016/j.cell.2017.09.043.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., Assiotis, I., Fenwick, K., Maguire, S. L., Campbell, J., Natrajan, R., Lambros, M., Perrakis, E., Ashworth, A., Fraser, P., and Fletcher, O. Unbiased analysis of potential targets of breast cancer susceptibility loci by capture hi-c. *Genome research*, pages 1854–1868, 2014. doi: 10.1101/gr.175034.114.

Heidari, N., Phanstiel, D. H. D., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P. Genome-wide map of regulatory interactions in the human genome. *Genome research*, 24(12):1905–17, 2014. doi: 10.1101/gr.176586.114.Freely.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–19, 2013. doi: 10.1016/j.cell.2013.03.035.

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., NISC Comparative Sequencing Program, N. C. S., Black, B. L., Visel, A., Pennacchio, L. A., Collins, F. S., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors, and NISC Comparative Sequencing Program Authors. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17921–6, 2013. doi: 10.1073/pnas.1317023110.

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannet, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S., and Young, R. A. Yy1 is a structural regulator of enhancer-promoter loops. *Cell*, (172):1–16, 2018. doi: 10.1016/j.cell.2017.11.008.

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C.-y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A., and Wasserman, W. W. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 44 (November 2015):gkv1176, 2015. doi: 10.1093/nar/gkv1176.

Lun, A. T. L., Perry, M., and Ing-Simmons, E. Infrastructure for genomic interactions : Bioconductor classes for hi-c , chia-pet and related experiments [version 1; referees: 2 approved]. *F1000Research*, 5(950):1–6, 2016. doi: 10.12688/f1000research.8759.1.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17): 2204–2207, 2010. doi: 10.1093/bioinformatics/btq351.

Hansen, P., Hecht, J., Ibn-Salem, J., Menkuec, B. S., Roskosch, S., Truss, M., and Robinson, P. N. Q-nexus: a comprehensive and efficient analysis pipeline designed for chip-nexus. *BMC Genomics*, 17(1):873, 2016. doi: 10.1186/s12864-016-3164-6.

Saito, T. and Rehmsmeier, M. Precrec: Fast and accurate precision-recall and roc curve calculations in r. *Bioinformatics*, 33(1):btw570, 2016. doi: 10.1093/bioinformatics/btw570.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005. doi: 10.1093/bioinformatics/bti623.

Zhou, X., Lowdon, R. F., Li, D., Lawson, H. a., Madden, P. a. F., Costello, J. F., and Wang, T. Exploring long-range genome interactions using the washu epigenome browser. *Nature methods*, 10(5):375–6, 2013. doi: 10.1038/nmeth.2440.

Harmston, N., Ing-Simmons, E., Perry, M., Barešić, A., and Lenhard, B. Genomicinteractions: An r/bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics*, 16(1):963, 2015. doi: 10.1186/s12864-015-2140-x.

Durand, N., Shamim, M., Machol, I., Rao, S., Huntley, M., Lander, E., and Aiden, E. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems*, 3(1):95–98, 2016. doi: 10.1016/j.cels.2016.07.002.

Rennie, S., Dalby, M., van Duin, L., and Andersson, R. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory inter-

actions. *Nature Communications*, 9(1):487, 2018. doi: 10.1038/s41467-017-02798-1.

Dickel, D. E., Ypsilanti, A. R., Pla, R., Zhu, Y., Barozzi, I., Mannion, B. J., Khin, Y. S., Fukuda-Yuzawa, Y., Plajzer-Frick, I., Pickle, C. S., Lee, E. A., Harrington, A. N., Pham, Q. T., Garvin, T. H., Kato, M., Osterwalder, M., Akiyama, J. A., Afzal, V., Rubenstein, J. L. R., Pennacchio, L. A., and Visel, A. Ultraconserved enhancers are required for normal development. *Cell*, 172(3):491–499.e15, 2018. doi: 10.1016/j.cell.2017.12.017.

Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., and Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*, pages 1–31, 2017. doi: 10.1038/s41559-017-0377-2.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–7, 2009. doi: 10.1073/pnas.0903103106.

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puviindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H.-K., Naranjo, S., Acemel, R. D., Manzanares, M., Nagy, A., Cox, N. J., Hui, C.-C., Gomez-Skarmeta, J. L., and Nóbrega, M. A. Obesity-associated variants within fto form long-range functional connections with irx3. *Nature*, 507(7492):371–375, 2014. doi: 10.1038/nature13138.

Javierre, B., Burren, O., Wilder, S., Kreuzhuber, R., Hill, S., Sewitz, S., Cairns, J., Wingett, S., Várnai, C., Thiecke, M., Burden, F., Farrow, S., Cutler, A., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., Stunnenberg, H. G., Todd, J. A., Zerbino, D. R., Stegle, O., Ouwehand, W. H., Frontini, M., Wallace, C., Spivakov, M., and Fraser, P. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384.e19, 2016. doi: 10.1016/j.cell.2016.09.037.

Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal, M. J., Sutton, G. J., Hormozdiari, F., Lu, D., Lee, C., Eskin, E., Voineagu, I., Ernst, J., and Geschwind, D. H. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, pages 1–20, 2016. doi: 10.1038/nature19847.

Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., and Mendelson-cohen, N. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature Publishing Group*, 547(7661):61–67, 2017. doi: 10.1038/nature23001.

Sekelja, M., Paulsen, J., and Collas, P. 4d nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? *Genome Biology*, 17(1):54, 2016. doi: 10.1186/s13059-016-0923-2.

Eisenberg, E. and Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013. doi: 10.1016/j.tig.2013.05.010.

Tang, Z., Berlin, D. S., Toji, L., Toruner, G. A., Beiswanger, C., Kulkarni, S., Martin, C. L., Emanuel, B. S., Christman, M., and Gerry, N. P. A dynamic database of microarray-characterized cell lines with various cytogenetic and genomic backgrounds. *G3 (Bethesda, Md.)*, 3(7):1143–9, 2013. doi: 10.1534/g3.113.006577.

Lachke, S. A., Higgins, A. W., Inagaki, M., Saadi, I., Xi, Q., Long, M., Quade, B. J., Talkowski, M. E., Gusella, J. F., Fujimoto, A., Robinson, M. L., Yang, Y., Duong, Q. T., Shapira, I., Motro, B., Miyoshi, J., Takai, Y., Morton, C. C., and Maas, R. L. The cell adhesion gene pvrl3 is associated with congenital ocular defects. *Human Genetics*, 131(2):235–250, 2012. doi: 10.1007/s00439-011-1064-z.

Anger, G. J., Crocker, S., McKenzie, K., Brown, K. K., Morton, C. C., Harrison, K., and MacKenzie, J. J. X-linked deafness-2 (dfnx2) phenotype associated with a paracentric inversion upstream of pou3f4. *American Journal of Audiology*, 23(1):1, 2014. doi: 10.1044/1059-0889(2013/13-0018).

Jenkins, E. C., Curcuru-Giordano, F. M., Krishna, S. G., and Cantarella, J. De novo occurrence of 46,xx,t(4;13) (q31;q14) in a mentally retarded girl. *Annales de genetique*, 18(2):117–20, 1975.

Frizell, E. R., Sutphen1, R., Diamond Jr., F. B., Sherwood, M., and Overhauser, J. t(1;18)(q32.1;q22.1) associated with genitourinary malformations. *Clinical Genetics*, 54(4):330–333, 1998. doi: 10.1034/j.1399-0004.1998.5440411.x.