Genome folding in evolution and disease

Jonas Ibn-Salem

January 25, 2018
Version: 0.0.1

Johannes Gutenberg-Universität Mainz

Fachbereich Biologie
Institute of Organismic and Molecular Evolution
Computational Biology and Data Mining Group

Documentation

Genome folding in evolution and disease

Jonas Ibn-Salem

1. Reviewer Prof. Dr. Miguel Andrade-Navarro
Fachbereich Biologie
Johannes Gutenberg-Universität Mainz

2. Reviewer Prof. Dr. Thomas Hankeln
Fachbereich Biologie
Johannes Gutenberg-Universität Mainz

Supervisor Miguel Andrade

January 25, 2018

Jonas Ibn-Salem
Genome folding in evolution and disease
Documentation, January 25, 2018
Reviewers: Prof. Dr. Miguel Andrade-Navarro and Prof. Dr. Thomas Hankeln
Supervisors: Miguel Andrade and

Johannes Gutenberg-Universität Mainz
Computational Biology and Data Mining Group
Institute of Organismic and Molecular Evolution
Fachbereich Biologie
Ackermannweg 4
55128 and Mainz

Contents

Preface

1 Introduction

Structure of the Intorduction

- Cell diversity
- Gene Regulation
- Enhancers
  - Enhancer–promoter interaction
- Techniques to probe the 3D chromatin architecture
  - Imaging based techniques
  - 3C based methods
  - Hi-C
  - ChIA-PET
- Chromatin 3D Structure
  - Hierarchy of chromatin 3D structure
    - Chromosomal Territories
    - A/B-Compartments
    - TADs
    - Chromatin Loops
      - Enhancer-Promoter loops
      - Gene Loops
      - Architectural loops
  - Architectural Proteins
  - Loop extrusion model
- Evolution of chromatin architecture
- Disruption of chromatin architecture

Cell diversity and regulation of gene expression

- Cell diversity
- Gene Regulation

An important mechanism in eukaryotic gene regulation is the binding of transcription factors (TFs) to distal regulatory regions such as enhancers which perform looping interactions to the transcription machinery at gene promoters. Chromatin interactions can be measured by chromatin conformation capture (3C) and its high-throughput variations such as 4C, 5C, 6C, ChIA-PET and Hi-C [Dekker et al., 2013, Lieberman-Aiden et al., 2009].

While genome-wide Hi-C data is still only available for a limited number of cell-types and has limited resolution, it is successfully used to re-discover important features of the three-dimensional chromatin architecture in the nucleus that is organized on different levels. In interphase, chromosomes occupy distinct territories in the nucleus [Cremer and Cremer, 2001] and are further organized in mega-base scale A/B-compartments that show specific preferential interaction patterns and transcriptional activity [Lieberman-Aiden et al., 2009, Rao et al., 2014]. On smaller scales topologically associating domains (TADs) were identified using Hi-C [Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012]. These are regions of several hundred kb, that have more interactions within themselves than with other regions and are separated by boundaries that insulate interactions between loci in different TADs. Interestingly, TADs are largely stable across cell-types and conserved between mammals [Dixon et al., 2012, Rao et al., 2014, Dixon et al., 2015, Vietri Rudan et al., 2015]. However, many cell-type specific chromatin interaction loops occur within TADs [Rao et al., 2014, Dixon et al., 2015]. This indicates a stable and cell-type invariant chromatin architecture on larger scales, such as TADs, and a more dynamic and cell-type specific organization of interactions within TADs that connect enhancers and TF binding sites to regulated genes.

Enhancers

Enhancers where 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] or 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 affects 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 enhancers, but only by combinatorial repression of several weak enhancers [Xie et al., 2017]. (Enhancers are reviewed in [Spitz and Furlong, 2012] and [Andrey and Mundlos, 2017])

Methods to probe the 3D chromatin archtiecture

- Imaging based techniques
- 3C based methods
- Hi-C
- ChIA-PET

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 spacial resolution. Novel super-resolution microscopy approaches such as STORM and PALM have enabled direct visualization of the fin-scale structures of the genome at 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 nuclear architecture of the complete genome. Furthermore, the specific folding patterns observed in microscopy cannot be mapped to genomic coordinates which hinders the integration with other genomic data. However, 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 experimentally measured 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 crosslink two genomic loci that are in close physical proximity in the nuclear space by treating cells with formaldehyde. The crosslinked chromatin is than digested by enzymes to fragment the genomic DNA. Than the fragmented DNA is re-ligated which results in hybrid DNA molecules of restriction fragments that where in close physical proximity during crosslinking but normally originate from different regions in the linear genome sequence [Dekker et al., 2013, Andrey and Mundlos, 2017] (Fig. 1.1A).

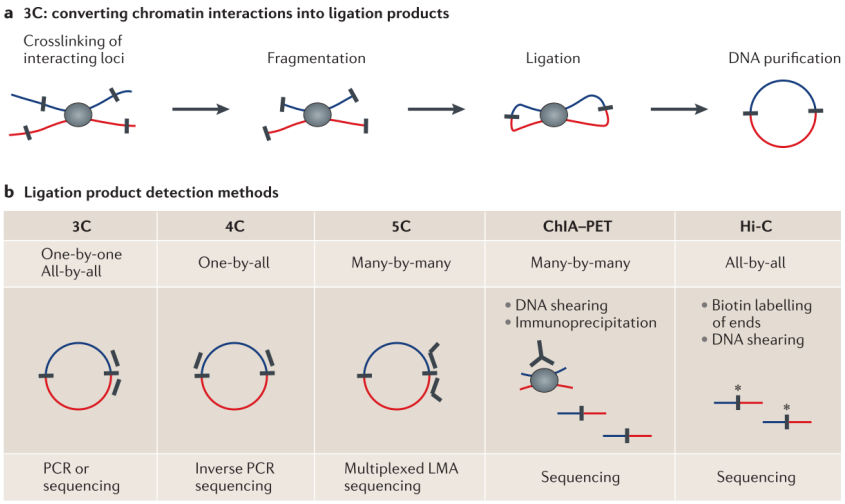

Figure 1.1.: Proximity ligation technologies to measure chromatin interactions (A) By treating cells with formaldehyde chromatin is crosslinked. After fragmentation with restriction enzymes, DNA from two loci in close physical proximity in the nucleus is ligated to a hybrid DNA molecules that is than 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 modified 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 ligated with a desired restriction fragment, e.g. a specific viewpoint of interest. These products are than 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]. Some 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 very 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 than sequenced using paired-end sequencing technology. Each read from the resulting read-pairs is than aligned independently to the reference genome to identify the originating position of the sequenced restriction fragment. Thereby 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 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

- Chromosomal Territories
- A/B-Compartments
- TADs
- Chromatin Loops
  - Enhancer-Promoter loops
  - Gene Loops
  - Architectural loops
- Architectural Proteins
- (Loop extrusion model)
- Nucleosomes / 10nm / 30nm fibers

Chromosomal territories and inter-chromosomal contacts

The eukaryotic genome is highly organized in the interpahse nucelus. 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 order of magnitudes less frequent than 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 monoallelic expression of a single receptor gene per cell [Monahan and Lomvardas, 2015]. While there are specific inter-chromosomal contacts, the genome is non-randomly organized in chromosomes which occupy distinct territories in the spacial nucleus.

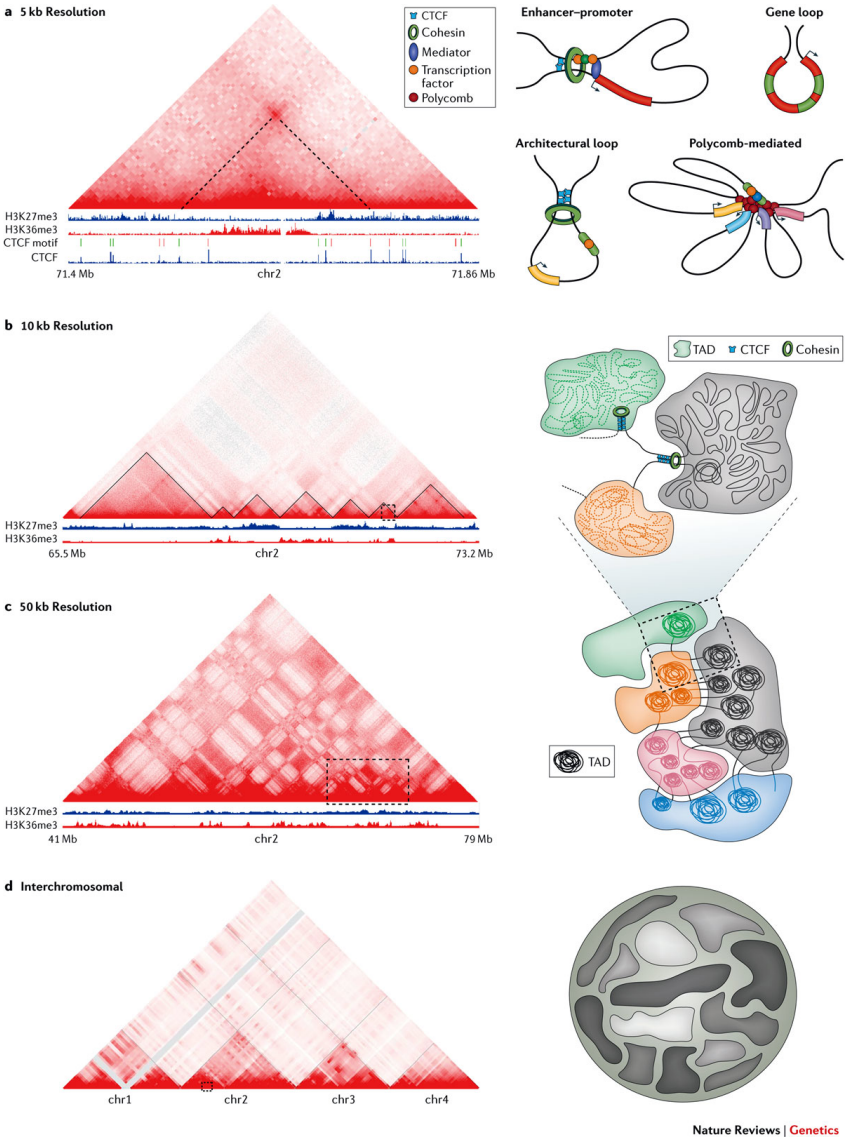

Figure 1.2.: Hierarchical organization of chromatin three-dimensional chromatin architecture (A)
Figure source [Bonev and Cavalli, 2016] https://www.nature.com/articles/nrg.2016.112/figures/2.

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 compartments mayor clusters, referred to as A/B-compartments [Lieberman-Aiden et al., 2009]. Interestingly, regions A-compartments are associated with active histone-modifications and active transcription, whereas B-compartment is associated with heterochromatin, lamina association, and repressed genes.

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 wiht them. These genomic regions were termed topologically associating domains (TADs) [Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012]. They are opperationlly defined as genomic regions with frequent interactions of loci within the domain and decreased interactions across domain boundares [**?**].

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 than identified in regions where DI is changing drastically [Dixon et al., 2012]. Other algorithms compute an insulation score as the extent to which interactions to cross potential TAD boundaries [Crane et al., 2015]. Later the Arrowhead algorithm was introduced 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 differnt length scales, hierarchies of TADs and sub-TADs could be identifed that overlap each other [Filippova et al., 2014, Fraser et al. [2015]].
The different algorithms and parameters used in each 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 [**?**], as well as different normalization strategies for Hi-C contacts introduce variability [Forcato et al., 2017]. Therefore the number and size of TADs varies between different studies and cannot be directly compared.

The first studies on TADs identifed around 3000 TADs in human and mouse genomes with a median size of ~800 kb [Dixon et al., 2012] and ~100 kb in *Drosophila* gneomes [Sexton et al., 2012]. Analysis of 1kb or 5kb resolution Hi-C matrixes resulted in nested “contact domains” with median size of 185 kb (range 40 kb - 3 Mb) in human and mouse cells [Rao et al., 2014].

Importantly, 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]. Annother 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 boudnaries are enriched for binding of “insulator protins”, such as CTCF in mammals and CP190 in Drosophila [Dixon et al., 2012, Sexton et al., 2012]. Furthermore, TAD boundares are associated with active chromatin, such as H3K4me3 and H3K27me3, DNase I hypersensivity, actie transcription, and house-keeping genes [Dixon et al., 2012]. Furthermore, TADs correspond to regions of early and late replication timing [Pope et al., 2014, **?**] and lamina associated domains (LADs) [Dixon et al., 2012]. Importantly, enhancer-promoter interactions seems to be mostly constrained within TADs [Shen et al., 2012]

There is acumulating evidence, that TADs are fundamental units of chromosome organization [Dixon et al., 2016]

- [X] Definition, number, and size
- [X] Algorithms to identify TADs
- Functional features
- [X] CTCF binding at boundaries
- [X] housekeeping genes at boundaries
- enhancer-gene associations within TADs

Chromatin loops

In summary, these findings suggest a hierarchical organization of chromosome architecture. First, dynamic nucleosome contacts from clutches and fibers. These engage in dynamic 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 in tern cluster by their epigenomic type into A/B compartments and coalescence of compartments in the same chromosome forms chromosome territories [**?**].

Dynamics of chromatin structure

- Cell-type invariant TADs

- Conservation of TADs

Changes of genome folding in evoltuion and disease genomes

- Evolution of chromatin architecture
- Disruption of chromatin architecture

Aims of this thesis

In this PhD thesis, I analyze TADs and chromatin interactions with respect to gene expression regulation. More specifically, we find out weather TADs represent only structural units of the genome or also functional building blocks, in which genes regulation is coordinated.

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

- How are duplicated genes during evolution distributed in the three-dimensional genome architecture?
- Can TADs provide regulatory environment for co-regulation of duplicated genes?
- (Are TADs functional genomic units in which genes are co-regulated?)
- (Are paralog genes co-regulated in the 3D chromatin architecture of genomes?)

Are TADs functional buidling blocks of genomes and subjected to selective preasure during evolution?

- Are TADs conserved during evolution or disrupted by rearrangements?
- Are changes of TADs during evolution associated with changes in gene expression profiles?
- (Are TADs stable units that are often transmitted as a whole than disrupted by rearrangements?)

Is the disruption of TADs by rearrangements also associated to genetic diseases?

- Can TADs be used to interpret position effects of rearrangements in genetic diseases?
- Can chromatin interaction data and TADs be integrated with phenotype data to predict pathomechanism of balanced chromosomal rearrangements?

Can chromatin looping interactions be predicted by ChIP-seq and sequence features?

- Are there signals form TF ChIP-seq data at chromatin looping anchors that predict long-range contacts?
- Does the gnomic sequence encode features that are predictive for chromatin looping interactions?
- Can we provide a computational method to predict chromatin looping interactions in specific cell-types and conditions of interest?

Structure of this thesis

These questions are addressed and discussed in the following for chapters. First, we focus on duplicated genes in the human genome. Because of their related sequence and function, shared evolutionary history, and close co-localization in the genome they represent an interesting model to study how genome folding is related to regulation of gene expression during evolution (Chapter Paralog genes). Furthermore, we make use of deeply sequenced genomes of other vertebrates to systematically investigate whether TADs represent conserved building blocks of genomes and whether rearrangements are associated with altered gene expression programs (Chapter TAD evolution). Next, we will address disruption of chromatin organization by analyzing disease associated rearrangement breakpoints from whole-genome sequenced patients of various genetic diseases to explain miss-regulation by disruption of TADs and chromatin contacts in diseases (Chapter Position effect). Finally, we will make use of recent insights in chromatin loop formation to provide a computational tool to predict chromatin loops from largely available genomic data, with the aim to facilitate association of TF binding sites in enhancers to regulated genes in many cell-types and condition, for which Hi-C like data is not available (Chapter Loop prediction).

Paralog genes in the 3D genome architecture

2

This chapter is published in Nucleac Acid Research [Ibn-Salem et al., 2017]. The surce code for the complete analysis is available at GitHub: https://github.com/ibn-salem/paralog_regulation

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., 2009a, 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 ≥ 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).

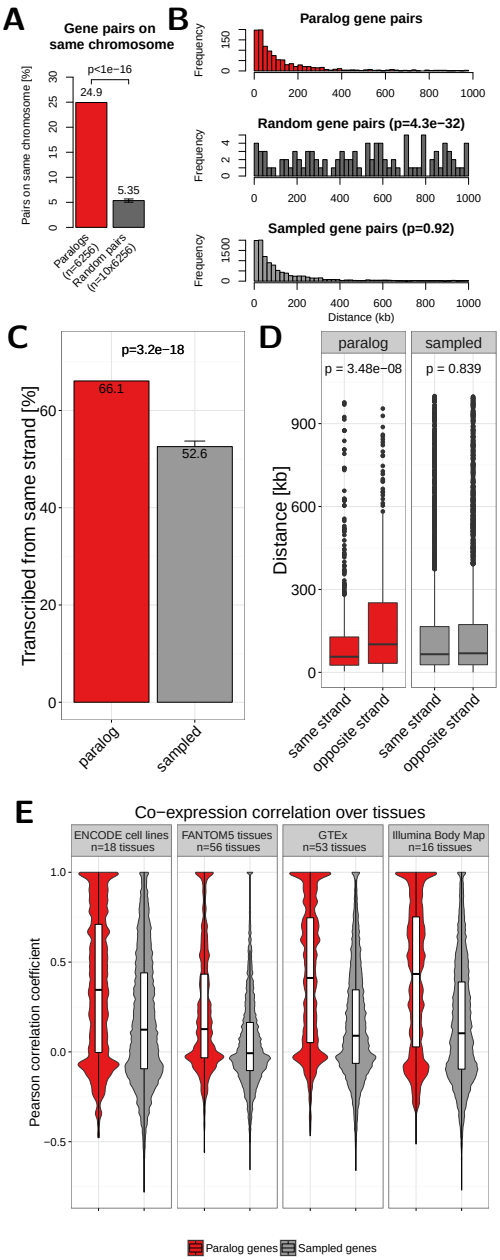

Figure 2.1.: **(A)** Percent of paralog (red) and random (dark grey) gene pairs that are located on the same chromosome. The error bar indicates the standard deviation observed in 10 times replicated random sampling of gene pairs. **(B)** Genomic distance distribution of paralogs gene pairs (top), random 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..

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.

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).

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 com-

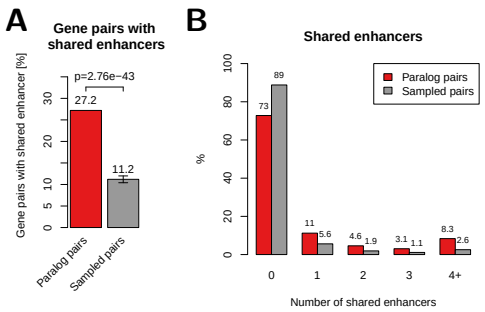

Figure 2.2.: Shared enhancers among paralog gene pairs. **(A)** Percent of close paralog (red) and sampled control (grey) gene pairs with at least one shared enhancer. **(B)** Percent of gene pairs versus number of shared enhancers for paralog and sampled control gene pairs.

pared 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).

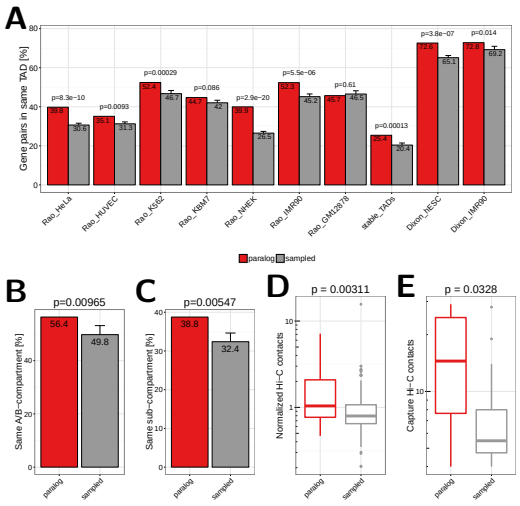

Figure 2.3.: **(A)** Co-localization of close paralog genes within the same TAD compared against sampled gene pairs for TAD data sets from different cell types and studies. The first seven bars show values for TADs called in HeLa, HUVEC, K562, KBM7, NHEK, IMR90, and GM12878 cells by [Rao et al., 2014]. 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 subcompartment 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).

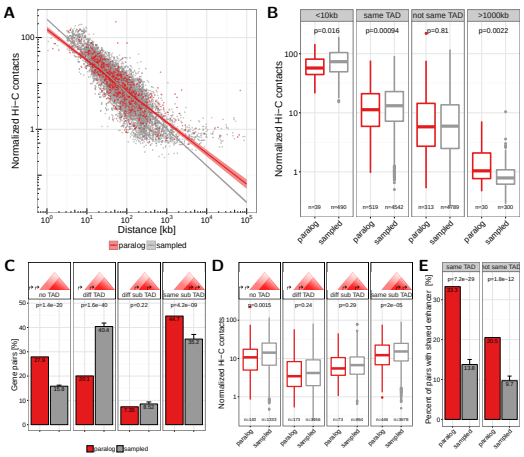

Figure 2.4.: **(A)** Normalized Hi-C contacts by genomic distance between paralog (red) and sampled (grey) gene pairs. Lines show linear regression fit separately for paralogs (red) and sampled (grey) pairs with 95% confidence intervals in shaded areas. **(B)** Normalized Hi-C contacts between pairs 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).

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 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.

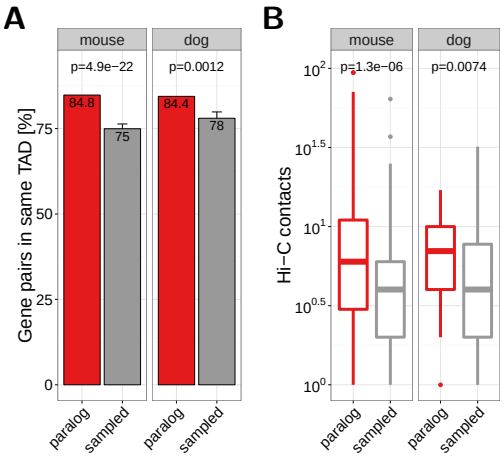

Figure 2.5.: **(A)** Co-occurrence of close paralog genes with the same TAD in mouse (left panel) and dog (right panel). **(B)** Hi-C contacts between promoter of distal gene pairs in Hi-C experiments in liver cells from mouse (left panel) and dog (right panel). Hi-C data and TAD calls were taken from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

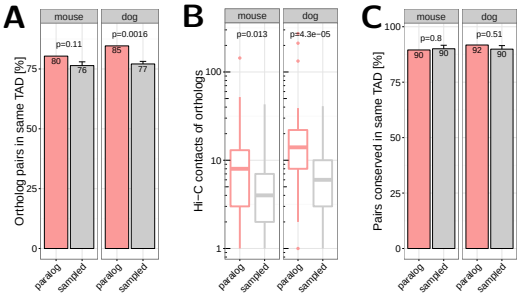

Figure 2.6.: One-to-one orthologs of human paralog genes in mouse and dog genome. **(A)** Percent of mouse (left) and dog (right) orthologs of human paralog pairs that are in the same TAD in the mouse and dog genome, respectively. **(B)** Normalized Hi-C contacts between promoters of one-to-one orthologs 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.

DISCUSSION

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 en-

code 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

Evolutionary stability of topologically associating domains is associated with conserved gene regulation

Jan Krefting[1,2], Miguel A. Andrade-Navarro[1,2] and Jonas Ibn-Salem[1,2,#]

[1]Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany

[2]Institute of Molecular Biology, 55128 Mainz, Germany

corresponding author

https://doi.org/10.1101/231431

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

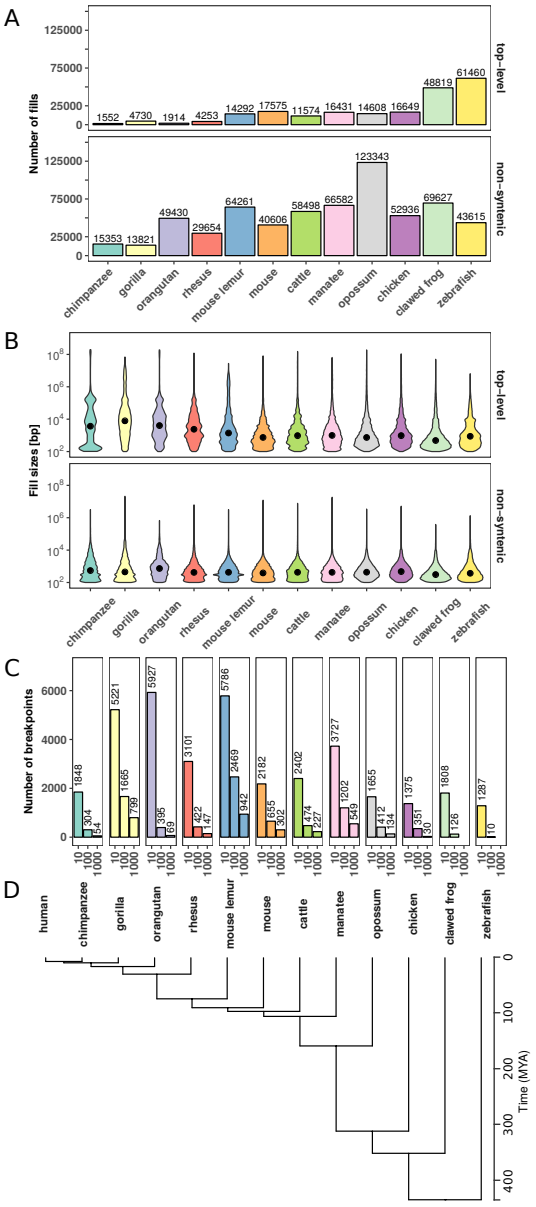

Figure 3.1.: **Number and size distributions of fill sizes of whole-genome alignments between human and 12 other species.** **(A)** Number of syntenic alignment blocks (fills) between human (hg38) and 12 other species. Top-level fills are the largest and highest scoring chains and occur at the top level 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. 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

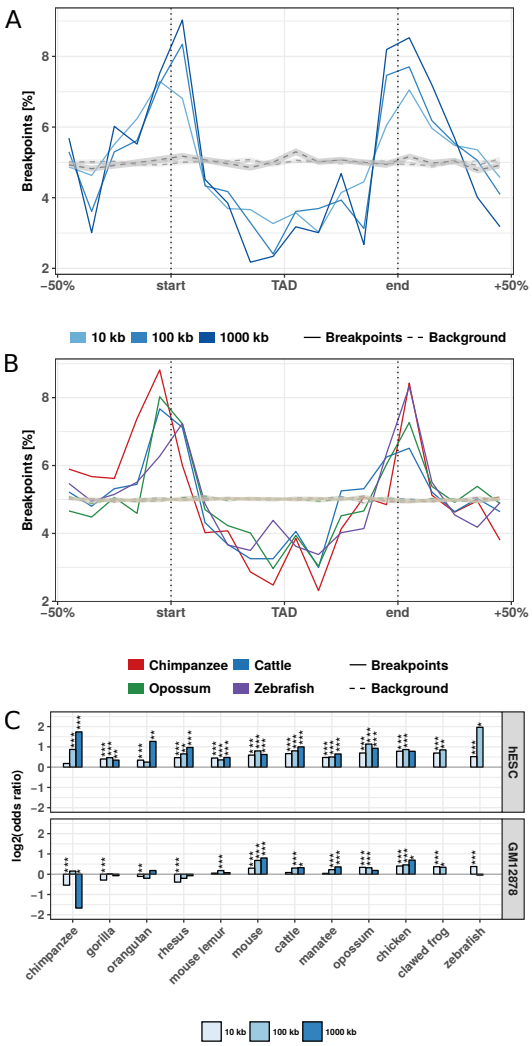

Figure 3.2.: **Evolutionary rearrangements are enriched at TAD boundaries. (A)** Distribution of evolutionary rearrangement breakpoints between human and mouse around hESC TADs. Each TAD and 50% of its adjacent sequence was subdivided into 20 bins of equal size, the breakpoints were assigned to the bins 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

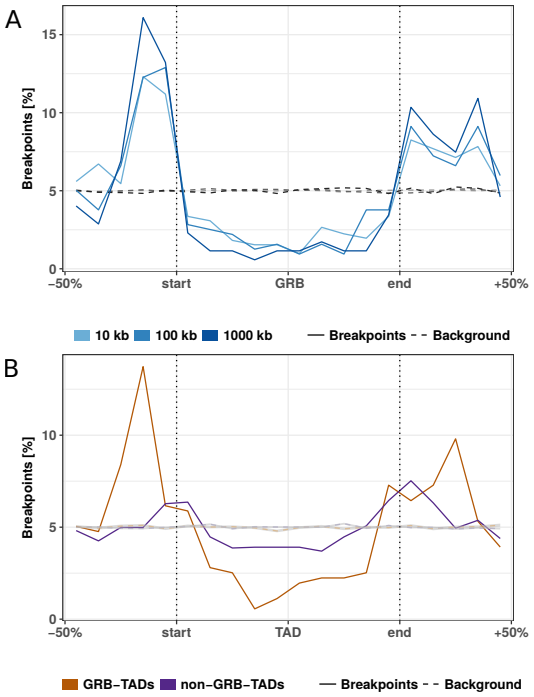

Figure 3.3.: **Rearrangement breakpoint distribution around GRBs and GRB-TADs. (A)** Rearrangement breakpoints between mouse and human around 816 GRBs. **(B)** Breakpoint distribution around GRB-TADs and non-GRB-TADs. GRB-TADs are defined as TADs overlapping more than 80% with GRBs and non-GRB-TADs have 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., 2016]. 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

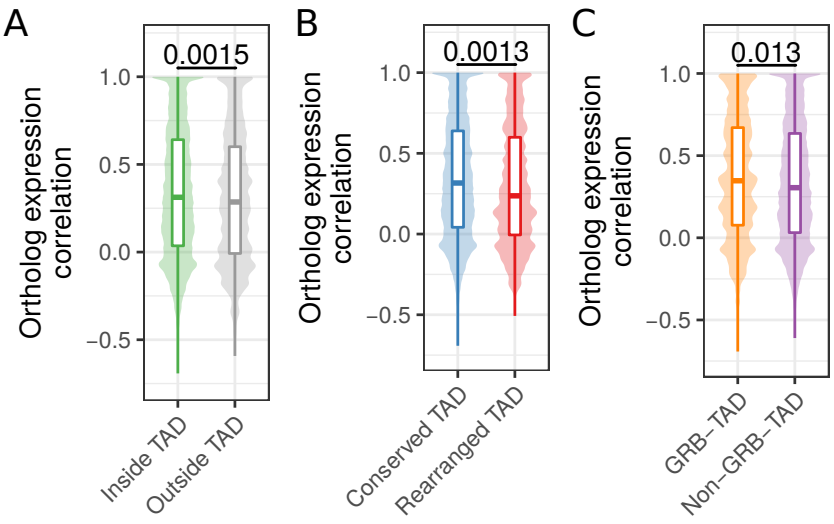

Figure 3.4.: **Ortholog gene expression correlation across tissues in conserved and rearranged TADs.** **(A)** Expression correlation of orthologs across 19 matching tissues in human and mouse for human genes within or outside of hESC TADs. **(B)** Expression correlation of orthologs across 19 matching 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., 2016, 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., 2016, 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 1). Alignment data were

downloaded as net files from UCSC Genome Browser for human genome hg38 and the genomes listed in Table 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].

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., 2009b]. 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 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.

Tables

Table 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

4

Position effects of rearrangements in disease genomes

Cinthya J. Zepeda-Mendoza[1,3,*], Jonas Ibn-Salem[4,*], Tammy Kammin[1], David J. Harris[3,5], Debra Rita[6], Karen W. Gripp[7], Jennifer J. MacKenzie[8], Andrea Gropman[9], Brett Graham[10], Ranad Shaheen[11], Fowzan S. Alkuraya[11,12], Campbell K. Brasington[13], Edward J. Spence[13], Diane Masser-Frye[14], Lynne M. Bird[14,15], Erica Spiegel[16], Rebecca L. Sparkes[17], Zehra Ordulu[18], Michael E. Talkowski[18-24], Miguel A. Andrade-Navarro[4], Peter N. Robinson[25], Cynthia C. Morton[1-3,23,26]

Departments of [1]Obstetrics, Gynecology and Reproductive Biology and [2]Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA

[3]Harvard Medical School, Boston, MA 02115, USA

[4]Johannes Gutenberg University, Mainz 55122, Germany

[5]Boston Children's Hospital, Boston, MA 02115, USA

[6]ACL laboratories. Cytogenetics Lab, Rosemont, IL 60018, USA

[7]Nemours Alfred I. DuPont Hospital for Children; Wilmington, DE 19803, USA

[8]Department of Pediatrics, McMaster University, Hamilton, ON L8S 4L8, Canada

[9]Children's National Medical Center, Washington, DC 20010, USA

[10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

[11]Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh 12713, Saudi Arabia

[12]Department of Anatomy and Cell Biology, College of Medicine, Alfaisal University, Riyadh 11533, Saudi Arabia

[13]Clinical Genetics Division, Department of Pediatrics. Levine Children's Hospital at Carolinas Medical Center. Charlotte, NC 28203, USA

[14]Genetics and Dysmorphology, Rady Children's Hospital San Diego, San Diego, CA 92123, USA

[15]University of California, San Diego, La Jolla, CA 92093, USA

[16]Maternal Fetal Medicine, Columbia University Medical Center, New York, NY 10032, USA

[17]Department of Medical Genetics, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada

Departments of [18]Pathology, [19]Neurology and [20]Psychiatry and [21]Center for Genomic Medicine and, Massachusetts General Hospital, Boston, MA 02114, USA

[22]Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

[23]Program in Medical and Population Genetics and [24]Stanley Center for Psychiatric Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[25]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

[26]Division of Evolution and Genomic Science, School of Biological Sciences, Manchester Academic Health Science Centre, Manchester M13 9NT, UK

*These authors contributed equally to this work

Corresponding author: Cynthia C. Morton, Ph.D. Brigham and Women's Hospital, New Research Building, Rm. 160D, 77 Avenue Louis Pasteur. Boston, MA 02115. Tel: 617-525-4535; Fax: 617-525-4533. Email: cmorton@partners.org

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' phenotypes. 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.[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, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, HapMap2010] 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, **?**] 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, **?**] 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, **?**] 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, **?**]

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 [**?**, Wit2012]), 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] 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,[**?**] and obesity,[Claussnitzer et al., 2015] among others.[Visser et al., 2012, **?**, **?**, Oldridge et al., 2015, Ibn-Salem et al., 2014] 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, **?**] 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, **?**, **?**] 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] [**??**]). 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 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.[**?**] Reported ages of DGAP subjects are from time of enrollment. All reported genomic coordinates use GRCh37.

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.[**?**] 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, **?**, 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[**?**] 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.[**?**] 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.[**?**] 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[**?**] 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.[**?**] 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, **?**] 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 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).

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 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 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 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.[**??**]

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 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 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.[**?**] 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 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 2).Expression assessment for second-tier candidates *PAK3*, *MID2* and *POU3F4* in DGAP176 LCLs did not deviate substantially from their control expression values (Figure S1); further searches into the Genotype-Tissue Expression (GTEx) project[**?**] 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.

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,[**?**] 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;[**?**] 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"[**?**] 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;[**?**] 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,[**?**, Wit2012] 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, **?**, Visser et al., 2012, **?**, 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 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 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,[**?**, Schmid2012] 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 inter-

est. As of now, the "top-ranking" candidates have the highest number of evidence 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 can-

didate 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.

Tables

Table 1.

Description of the 17 analyzed DGAP cases with non-codingBCAs. 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	Reported Karyotype	Disruption of Functional Elemen	Breakpoints within TA
DGAP017	46,X,t(X;10)(p11.2;q24.3)	DHS	2

Subject ID	Reported Karyotype	Disruption of Functional Elemen	Breakpoints w
DGAP111	46,XY,t(16;20)(q11.2;q13.2)dn	CTCF	1
DGAP113	46,XY,t(1;3)(q32.1;q13.2)dn	-	2
DGAP126	46,XX,t(5;10)(p13.3;q21.1)dn	-	2
DGAP138	46,XY,t(1;6)(q23;q13)dn	-	2
DGAP153	46,X,t(X;17)(p11.23;p11.2)dn	-	1
DGAP163	46,XY,t(2;14)(p23;q13)dn	-	2
DGAP176	46,Y,inv(X)(q13q24)mat	DHS, CTCF	2
DGAP249	46,XX,t(2;11)(q33;q23)dn	E, DHS	2
DGAP252	46,XY,t(3;18)(q13.2;q11.2)dn	-	2
DGAP275	46,XX,t(7;12)(p13;q24.33)dn	DHS	1
DGAP287	46,XY,t(10;14)(p13;q32.1)dn	CTCF	2
DGAP288	46,XX,t(6:17)(q13;q21)dn	DHS	2
DGAP315	46,XX,inv(6)(p24q11)dn	-	1
DGAP319	46,XX,t(4;13)(q31.3;q14.3)dn	-	2
DGAP322	46,XY,t(1;18)(q32.1;q22.1)	DHS	1
DGAP329	46,XX,t(2;14)(q21;q24.3)dn	-	1

Figures

Figure 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.

Figure 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*.

Figure 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.

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

This chapter has to be added.

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.

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

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.

A

Supporting Information: Co-regulation of paralog genes in the three-dimensional chromatin architecture

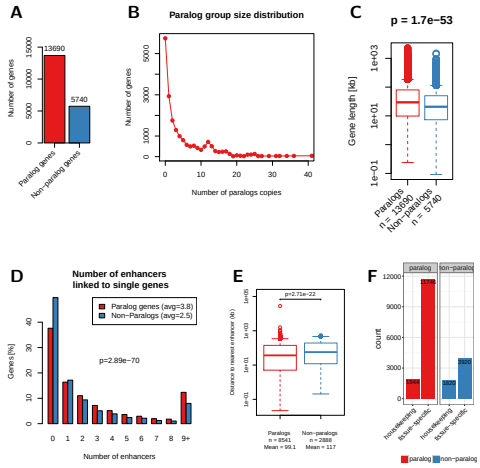

Figure A.1.: **(A)** Number of paralog and non-paralog genes in the human genome. **(B)** Paralog group size distribution in the human genome. **(C)** Gene length of paralog and non-paralog genes. **(D)** Distribution of the number of enhancers linked to single genes compared between paralog genes (red) 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.

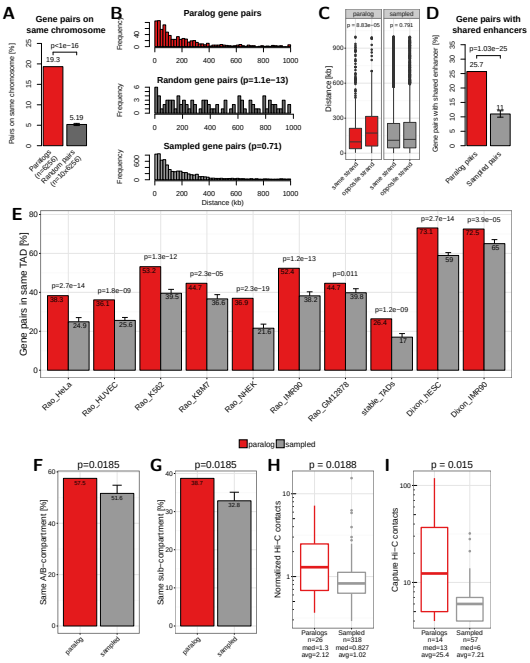

Figure A.2.: Main results of this study by changing the selection of paralog pairs from families with more than two paralogs. Here pairs are selected by maximizing the rate of synonymous mutations between them instead of minimizing, as in the main text. **(A)** Percent of paralogs pairs on the same chromosome 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.

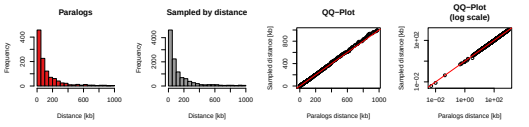

Figure A.3.: **Sampling of gene pairs by distance.** Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column).

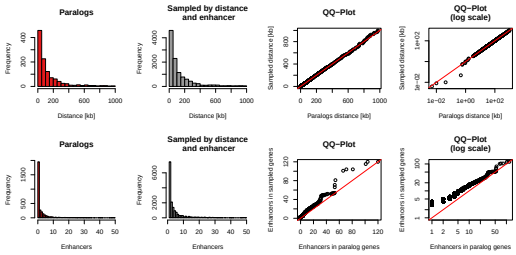

Figure A.4.: Sampling of gene pairs by distance and number of enhancers. Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth column). Bottom row: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

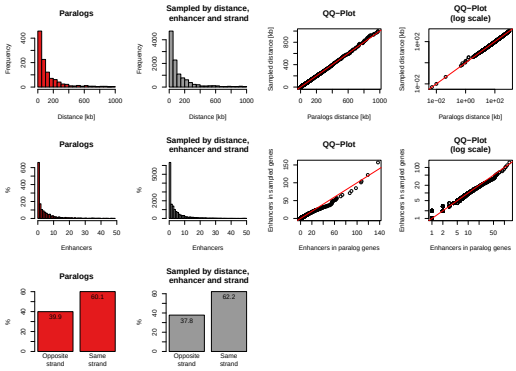

Figure A.5.: Sampling of gene pairs by distance, number of enhancers, and same strand frequency. Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth 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).

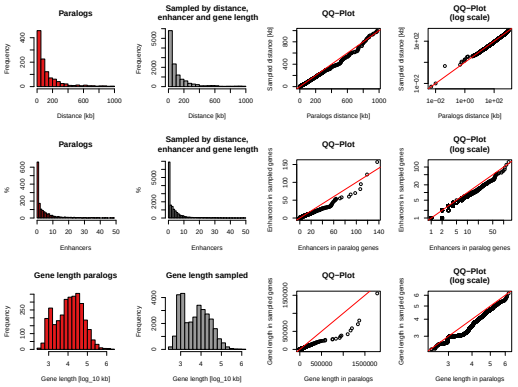

Figure A.6.: Sampling of gene pairs by distance, number of enhancers, and same strand frequency. Top row: Distance distribution of paralog pairs (red) and sampled background gene pairs (grey) and quantile-quantile plot of these two distributions in linear axis (third column) and log scaled axis (fourth 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).

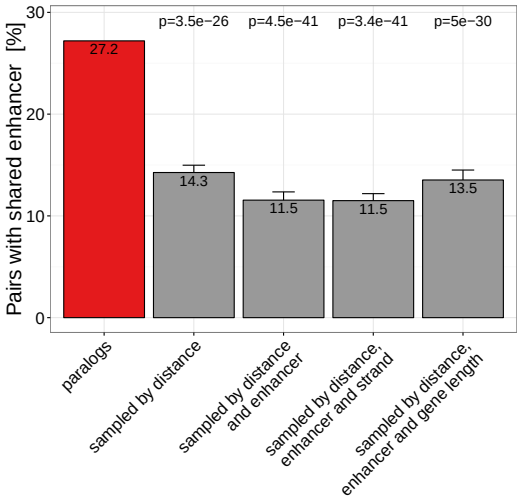

Figure A.7.: Percent of gene pairs with at least one shared enhancer in paralog pairs and four different types of sampled gene pairs. Only pairs with TSS distance $\leq$ 1Mb are considered. Error bars indicate standard variation of ten times replicated sampling.

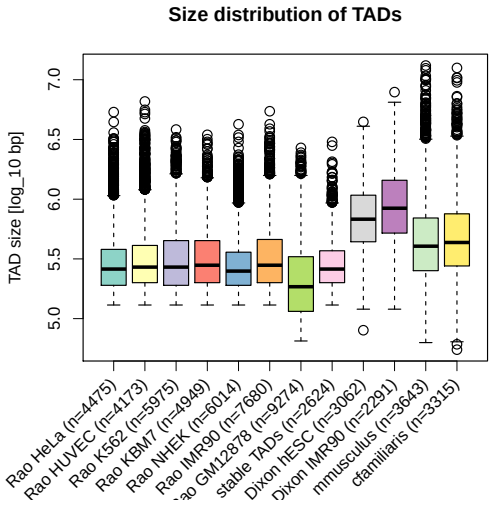

Figure A.8.: **Size distribution of TADs in different cell-types, studies, and species.** Each box shows the size-distribution of one data set of TADs. The labels indicate the study (Rao [Rao et al., 2014], or Dixon [Dixon et al., 2012]), cell type and number of TADs in each data set. The last two boxes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

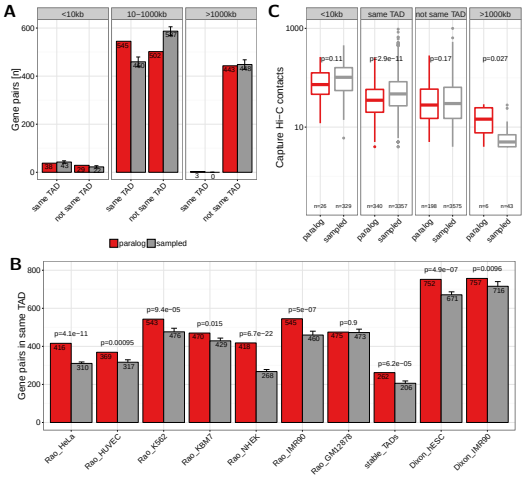

Figure A.9.: **(A)** Number of paralog (red) and sampled (grey) gene pairs that are in the same TAD or not separated in three groups of genomic distances (0-10kb, 10-1000kb and $>$ 1000kb). TADs called from IMR90 cells by [Rao et al., 2014] were used here. **(B)** Co-localization of gene pairs with genomic 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.

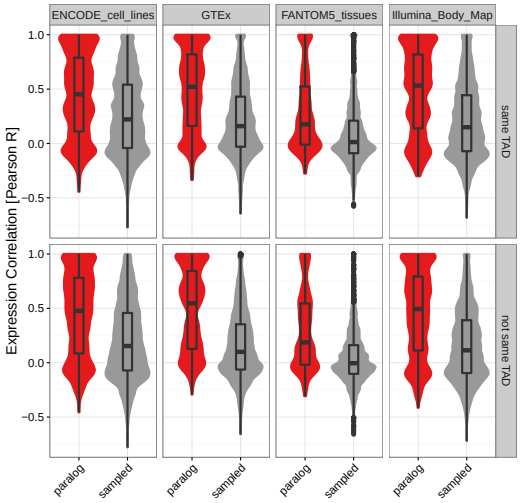

Figure A.10.: Distribution of Pearson correlation coefficients of gene expression values in four independent data sets between close paralog gene pairs (red) and sampled control gene pairs (grey) separated for gene pairs within the same IMR90 TAD (top) or not in the same TAD (bottom). Boxes show 25th,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

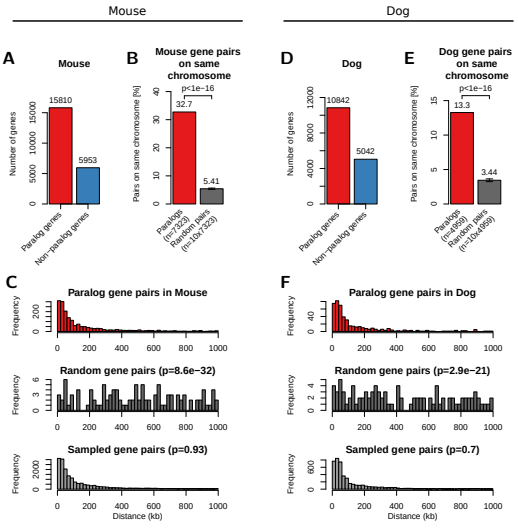

Figure A.11.: Paralog gene pairs in mouse (left) and dog (right) genome cluster on chromosome within short genomic distances. **(A)** Number of genes with paralogs (red) and without (blue) in mouse genomes. **(B)** Percent of filtered mouse paralog pairs on the same chromosome (red) and random gene pairs 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.

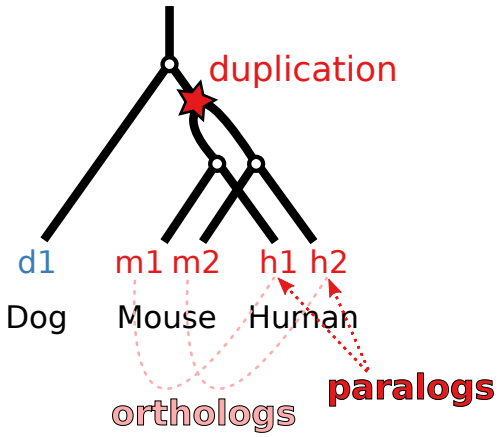

Figure A.12.: Phylogenetic gene tree model of a gene that is duplicated before the separation of mouse and human and consequently leads to two paralogs in mouse and human that are one-to-one orthologs to each other and a single ortholog in the dog genome that cannot be assigned uniquely to a human gene.

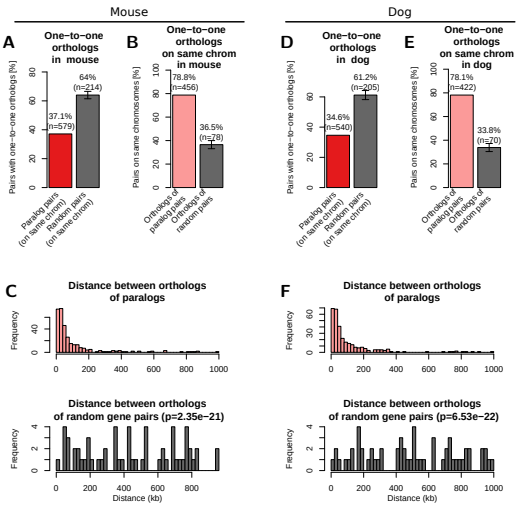

Figure A.13.: One-to-one orthologs of human paralogs in mouse (left) and dog (right) genome. **(A)** Percent of filtered human paralog pairs with one-to-one orthologs for both genes in mouse genome compared to random genes. **(B)** Percent of one-to-one orthologs on the same chromosome in the mouse genome (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.

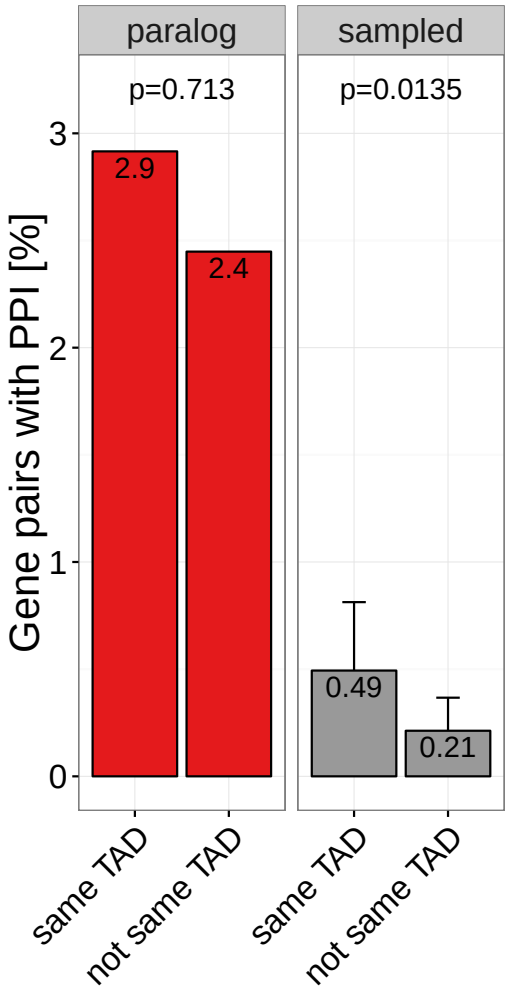

Figure A.14.: Percent of close paralogs (red) and sampled (grey) gene pairs in the same IMR90 TAD (left bar) or not same TAD (right bar) that have a direct protein protein interaction (PPI) with each other in the HIPPIE database [Schaefer et al., 2012].

B

Supplementary Data: Evolutionary stability of topologically associating domains is associated with conserved gene regulation

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

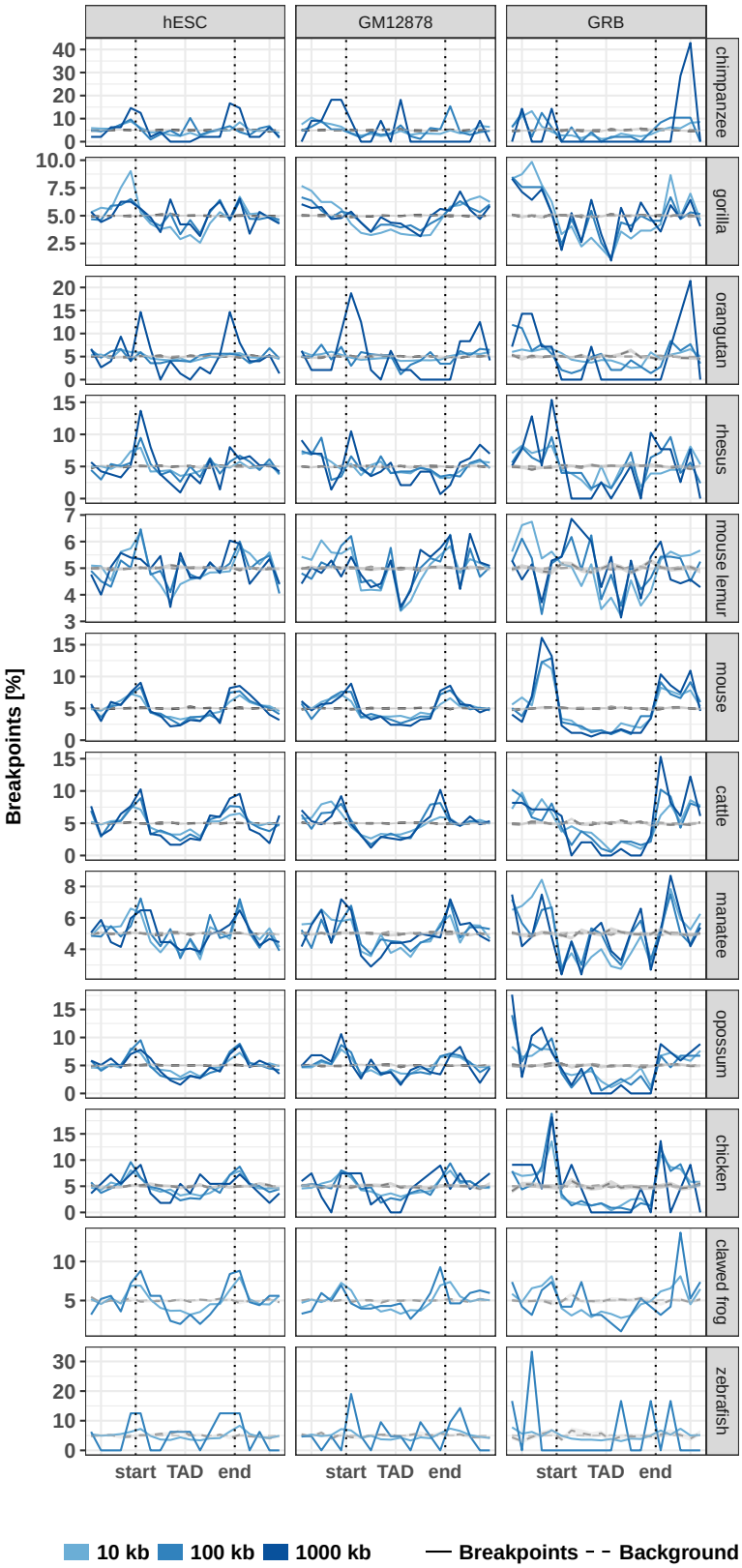

Figure B.1.: **Distribution of evolutionary rearrangement breakpoints between human and 12 vertebrate genomes around domains.** Relative breakpoint numbers from human and different species (horizontal panels) around hESC TADs (left), GM12878 contact domains (center), and GRBs (left). Blue color scale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

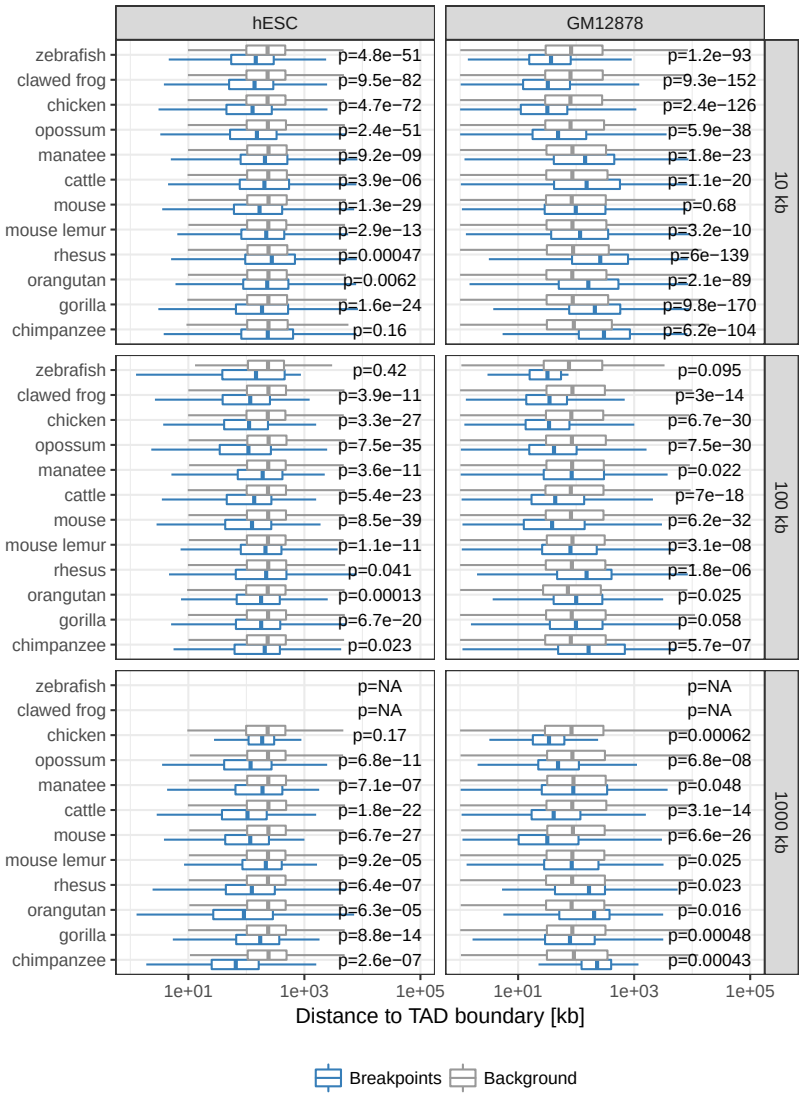

Figure B.2.: Distance between rearrangement breakpoints and random controls to closest TAD boundary. For each species (y-axis) and fill size threshold (vertical panels) the distances from all identified rearrangement breakpoints to its closest TAD boundary (x-axis) are compared between actual rearrangements (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.

Bibliography

Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403, 2013. ISSN 1471-0064. doi: 10.1038/nrg3454. URL http://dx.doi.org/10.1038/nrg3454.

Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid a Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950): 289–93, oct 2009. ISSN 1095-9203. doi: 10.1126/science.1181369. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2858594{&}tool=pmcentrez{&}rendertype=abstracthttp://dx.doi.org/10.1126/science.1181369.

T Cremer and C Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(April):292–301, 2001. URL http://www.nature.com/nrg/journal/v2/n4/abs/nrg0401{_}292a.html.

Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80, dec 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.11.021. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867414014974http://dx.doi.org/10.1016/j.cell.2014.11.021http://www.ncbi.nlm.nih.gov/pubmed/25497547.

Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, apr 2012. ISSN 0028-0836. doi: 10.1038/nature11082. URL http://www.nature.com/doifinder/10.1038/nature11082.

Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–5, may 2012. ISSN 1476-4687. doi: 10.1038/nature11049. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3555144{&}tool=pmcentrez{&}rendertype=abstract.

Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–72, feb 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.01.010. URL http://www.ncbi.nlm.nih.gov/pubmed/22265598.

Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V. Lobanenkov, Joseph R. Ecker, James a. Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015. ISSN 0028-0836. doi: 10.1038/nature14222. URL http://www.nature.com/doifinder/10.1038/nature14222.

Matteo Vietri Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay, and Suzana Hadjur. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell reports*, 10(8):1297–309, mar 2015. ISSN 2211-1247. doi: 10.1016/j.celrep.2015.02.004. URL http://linkinghub.elsevier.com/retrieve/pii/S2211124715001126http://www.ncbi.nlm.nih.gov/pubmed/25732821http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4542312.

Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2):299–308, dec 1981. ISSN 0092-8674. doi: 10.1016/0092-8674(81)90413-X. URL https://www.sciencedirect.com/science/article/pii/009286748190413X.

Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(March):272–86, mar 2014. ISSN 1471-0056. doi: 10.1038/nrg3682. URL http://www.ncbi.nlm.nih.gov/pubmed/24614317http://www.nature.com/doifinder/10.1038/nrg3682.

Lingyun Song and Gregory E Crawford. 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, feb 2010. ISSN 1559-6095. doi: 10.1101/pdb.prot5384. URL http://www.ncbi.nlm.nih.gov/pubmed/20150147http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3627383.

Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 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, dec 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2688. URL http://www.nature.com/articles/nmeth.2688.

Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, Laurie A Boyer, Richard A Young, and Rudolf Jaenisch. 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, dec 2010. ISSN 1091-6490. doi: 10.1073/pnas.1016071107. URL http://www.ncbi.nlm.nih.gov/pubmed/21106759http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3003124.

Deborah Hay, Jim R Hughes, Christian Babbs, James O J Davies, Bryony J Graham, Lars L P Hanssen, Mira T Kassouf, A Marieke Oudelaar, Jacqueline A Sharpe, Maria C Suciu, Jelena Telenius, Ruth Williams, Christina Rode, Pik-Shan Li, Len A Pennacchio, Jacqueline A Sloane-Stanley, Helena Ayyub, Sue Butler, Tatjana Sauka-Spengler, Richard J Gibbons, Andrew J H Smith, William G Wood, and Douglas R Higgs. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8):895–903, 2016. ISSN 1061-4036. doi: 10.1038/ng.3605. URL http://www.nature.com/doifinder/10.1038/ng.3605.

Anja J Will, Giulia Cova, Marco Osterwalder, Wing-Lee Chan, Lars Wittler, Norbert Brieske, Verena Heinrich, Jean-Pierre de Villartay, Martin Vingron, Eva Klopocki, Axel Visel, Darío G Lupiáñez, and Stefan Mundlos. Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nature Genetics*, (August), 2017. ISSN 1061-4036. doi: 10.1038/ng.3939. URL http://www.nature.com/doifinder/10.1038/ng.3939.

Shiqi Xie, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, 66(2):285–299.e5, apr 2017. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2017.03.007. URL https://www.sciencedirect.com/science/article/pii/

S1097276517301740.

François Spitz and Eileen E M Furlong. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 13(9):613–26, sep 2012. ISSN 1471-0064. doi: 10.1038/nrg3207. URL http://www.ncbi.nlm.nih.gov/pubmed/22868264.

Guillaume Andrey and Stefan Mundlos. The three-dimensional genome: regulating gene expression during pluripotency and development. pages 3646–3658, 2017. ISSN 0950-1991. doi: 10.1242/dev.148304. URL http://dev.biologists.org/content/develop/144/20/3646.full.pdf.

Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678, 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.112. URL http://www.nature.com/doifinder/10.1038/nrg.2016.112.

Tim J Stevens, David Lando, Srinjan Basu, P Liam, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O Shaughnessy-kirwan, Julie Cramard, Andre J Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G S Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, and Brian Hendrich. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, pages 1–21, 2017. ISSN 0028-0836. doi: 10.1038/nature21429. URL http://dx.doi.org/10.1038/nature21429.

Ilya M. Flyamer, Johanna Gassler, Maxim Imakaev, Sergey V. Ulyanov, Nezar Abdennur, Sergey V. Razin, Leonid Mirny, and Kikue Tachibana-Konwalski. Single-cell Hi-C reveals unique chromatin reorganization at oocyte-tozygote transition. *Nature Publishing Group*, 544(7648):1–17, 2017. ISSN 0028-0836. doi: 10.1038/nature21711. URL http://dx.doi.org/10.1038/nature21711.

Satish Sati and Giacomo Cavalli. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126(1):33–44, feb 2017. ISSN 0009-5915. doi: 10.1007/s00412-016-0593-6. URL http://link.springer.com/10.1007/s00412-016-0593-6.

Anthony D. Schmitt, Ming Hu, and Bing Ren. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, 2016. ISSN 1471-0072. doi: 10.1038/nrm.2016.104. URL http://www.nature.com/doifinder/10.1038/nrm.2016.104.

Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–11, 2002. ISSN

1095-9203. doi: 10.1126/science.1067799. URL http://www.ncbi.nlm.nih.gov/pubmed/11847345.

Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation captureon-chip (4C). *Nature Genetics*, 38(11):1348–1354, nov 2006. ISSN 1061-4036. doi: 10.1038/ng1896. URL http://www.nature.com/articles/ng1896.

Daan Noordermeer, Elzo de Wit, Petra Klous, Harmen van de Werken, Marieke Simonis, Melissa Lopez-Jones, Bert Eussen, Annelies de Klein, Robert H. Singer, and Wouter de Laat. Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature Cell Biology*, 13(8):944–951, aug 2011. ISSN 1465-7392. doi: 10.1038/ncb2278. URL http://www.nature.com/articles/ncb2278.

J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309, oct 2006. ISSN 1088-9051. doi: 10.1101/gr.5571506. URL http://www.ncbi.nlm.nih.gov/pubmed/16954542http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1581439http://www.genome.org/cgi/doi/10.1101/gr.5571506.

Amartya Sanyal, Bryan R. Lajoie, Gaurav Jain, and Job Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, sep 2012. ISSN 0028-0836. doi: 10.1038/nature11279. URL http://www.nature.com/doifinder/10.1038/nature11279.

Melissa J. Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G. Y. Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N. Ariyaratne, Vinsensius B. Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K. D. Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V. Desai, Jane S. Thomsen, Yew Kok Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G. Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T. Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009. ISSN 0028-0836. doi: 10.1038/nature08497. URL http://www.nature.com/doifinder/10.1038/nature08497.

Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, Simon Zhongyuan Tian, May Penrad-mobayed, Laurent M Sachs, Xiaoan Ruan, Chia-lin Wei, Edison T Liu, Grzegorz M Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, pages 1–17, 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.11.024. URL http://dx.doi.org/10.1016/j.cell.2015.11.024.

Miguel R. Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, 4(5):780–788, 2006. ISSN 15457885. doi: 10.1371/journal.pbio.0040138.

V Roukos, T C Voss, C K Schmidt, S Lee, D Wangsa, and T Misteli. Spatial dynamics of chromosome translocations in living cells. *Science*, 341(6146):660–664, 2013. ISSN 1095-9203. doi: 10.1126/science.1237150. URL http://www.ncbi.nlm.nih.gov/pubmed/23929981.

Vassilis Roukos and Tom Misteli. The biogenesis of chromosome translocations. *Nature cell biology*, 16(4):293–300, 2014. ISSN 1476-4679. doi: 10.1038/ncb2941. URL http://www.ncbi.nlm.nih.gov/pubmed/24691255.

Wouter de Laat and Frank Grosveld. Inter-chromosomal gene regulation in the mammalian cell nucleus. *Current opinion in genetics & development*, 17(5):456–464, 2007. ISSN 0959437X. doi: 10.1016/j.gde.2007.07.009.

Kevin Monahan and Stavros Lomvardas. Monoallelic Expression of Olfactory Receptors. *Annual Review of Cell and Developmental Biology*, 31(1):annurev–cellbio–100814–125308, 2015. ISSN 1081-0706. doi: 10.1146/annurev-cellbio-100814-125308. URL http://www.annualreviews.org/doi/10.1146/annurev-cellbio-100814-125308.

Ferhat Ay and William S. Noble. Analysis methods for studying the 3D architecture of the genome. *Genome Biology*, 16(1):183, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0745-7. URL http://genomebiology.com/2015/16/1/183.

Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 2015. ISSN 0028-0836. doi: 10.1038/nature14450. URL http://www.nature.com/doifinder/10.1038/nature14450.

Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, 9(1):14, jan 2014. ISSN 1748-7188. doi: 10.1186/1748-7188-9-14. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4019371{&}tool=pmcentrez{&}rendertype=abstract.

James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L Moore, Dorothee CA Kraemer, Stuart Aitken, Sheila Q Xie, Kelly J Morris, Masayoshi Itoh, Hideya Kawaji, Ines Jaeger, Yoshihide Hayashizaki, Piero Carninci, Alistair RR Forrest, Josée Dostie, Ana Pombo, and Mario Nicodemi. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, 11:1–14, 2015. ISSN 1744-4292. doi: 10.15252/msb.

Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, (May):14–19, 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4325. URL http://www.nature.com/doifinder/10.1038/nmeth.4325.

Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889, mar 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a003889. URL http://www.ncbi.nlm.nih.gov/pubmed/20300217http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2829961.

Johan H Gibcus and Job Dekker. The hierarchy of the 3D genome. *Molecular cell*, 49(5):773–82, mar 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2013.02.011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3741673{&}tool=pmcentrez{&}rendertype=abstract.

KyleăP. Eagen, TomăA. Hartl, and RogerăD. Kornberg. Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. *Cell*, 163(4):934–946, 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.10.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415013409.

Matthias Merkenschlager and Elphège P. Nora. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annual review of genomics and human genetics*, 17(April):17–43, 2016. ISSN 1527-8204. doi: 10.1146/annurev-genom-083115-022339. URL http://www.ncbi.nlm.nih.gov/pubmed/27089971http://www.annualreviews.org/doi/10.1146/annurev-genom-083115-022339.

Benjamin D. Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, Yanli Wang, R. Scott Hansen, Theresa K. Canfield, Robert E.

Thurman, Yong Cheng, Günhan Gülsoy, Jonathan H. Dennis, Michael P. Snyder, John a. Stamatoyannopoulos, James Taylor, Ross C. Hardison, Tamer Kahveci, Bing Ren, and David M. Gilbert. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, nov 2014. ISSN 0028-0836. doi: 10.1038/nature13986. URL http://www.nature.com/doifinder/10.1038/nature13986.

Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, and Bing Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–20, aug 2012. ISSN 1476-4687. doi: 10.1038/nature11243. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4041622{&}tool=pmcentrez{&}rendertype=abstract.

JesseăR. Dixon, DavidăU. Gorkin, and Bing Ren. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell*, 62(5):668–680, 2016. ISSN 10972765. doi: 10.1016/j.molcel.2016.05.018. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276516301812.

Jonas Ibn-Salem, Enrique M. Muro, and Miguel A. Andrade-Navarro. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Research*, 45(1):81–91, jan 2017. ISSN 13624962. doi: 10.1093/nar/gkw813. URL https://doi.org/10.1093/nar/gkw813https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw813.

Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annual review of genetics*, 39:309–338, 2005. ISSN 0066-4197. doi: 10.1146/annurev.genet.39.073003.114725.

Kateryna D K.D. Makova and Wen-Hsiung W.H. Li. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research*, 13(7):1638–1645, 2003. ISSN 1088-9051. doi: 10.1101/gr.1133803. URL http://genome.cshlp.org/content/13/7/1638.short.

M Ptashne. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322(6081):697–701, 1986. ISSN 0028-0836. doi: 10.1038/322697a0.

Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D Gregory, Ann Dean, and Gerd a Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–44, jun 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.03.051. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3372860{&}tool=pmcentrez{&}rendertype=abstract.

D Carter, L Chakalova, C S Osborne, Y F Dai, and P Fraser. Long-range chromatin regulatory interactions in vivo. *Nat Genet*, 32(4):623–626, 2002. ISSN 10614036. doi: 10.1038/ng1051. URL http://www.ncbi.nlm.nih.gov/pubmed/12426570{%}5Cnhttp://www.nature.com/ng/journal/v32/n4/pdf/ng1051.pdf.

Bas Tolhuis, Robert Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, 10(6):1453–1465, dec 2002.

F. Le Dily, D. Bau, a. Pohl, G. P. Vicent, F. Serra, D. Soronellas, G. Castellano, R. H. G. Wright, C. Ballare, G. Filion, M. a. Marti-Renom, and M. Beato. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development*, 28(19): 2151–2162, oct 2014. ISSN 0890-9369. doi: 10.1101/gad.241422.114. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.241422.114.

Jonas Ibn-Salem, Sebastian Köhler, Michael I Love, Ho-Ryun Chung, Ni Huang, Matthew E Hurles, Melissa Haendel, Nicole L Washington, Damian Smedley, Christopher J Mungall, Suzanna E Lewis, Claus-Eric Ott, Sebastian Bauer, Paul N Schofield, Stefan Mundlos, Malte Spielmann, and Peter N Robinson. Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biology*, 15(9):423, 2014. ISSN 1465-6906. doi: 10.1186/s13059-014-0423-1. URL http://genomebiology.com/2014/15/9/423.

Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, 161(5):1012–1025, may 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.04.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415003773.

Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8):1184–1191, 2009a. ISSN 1754-2189. doi: 10.1038/nprot.2009.97.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16): 3439–3440, 2005. ISSN 13674803. doi: 10.1093/bioinformatics/bti525.

Albert J. Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335, 2009. ISSN 10889051. doi: 10.1101/gr.073585.107.

Zvi Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys*, 18(1):23–38, 1986. ISSN 03600300. doi: 10.1145/6462.6502.

Xun Lan and Jonathan K Pritchard. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science (New York, N.Y.)*, 352(6288):1009–13, may 2016. ISSN 1095-9203. doi: 10.1126/science.aad8411. URL http://biorxiv.org/content/early/2015/05/10/019166http://www.ncbi.nlm.nih.gov/pubmed/27199432http://biorxiv.org/content/early/2016/02/02/019166.abstract.

Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, a Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David a Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R R Forrest, Piero Carninci, Michael Rehli, and Albin Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61, mar 2014. ISSN 1476-4687. doi: 10.1038/nature12787. URL http://www.ncbi.nlm.nih.gov/pubmed/24670763.

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

Philip a. Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33:1029–1047, 2013. ISSN 02724979. doi: 10.1093/imanum/drs019.

Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip a Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George a Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6), 2015. ISSN 1061-4036. doi: 10.1038/ng.3286. URL http://www.nature.com/doifinder/10.1038/ng.3286.

Xionglei He and Jianzhi Zhang. Gene complexity and gene duplicability. *Current Biology*, 15(11):1016–1021, 2005. ISSN 09609822. doi: 10.1016/j.cub.2005.04.035.

Scott Newman, Karen E Hermetz, Brooke Weckselblatt, and M Katharine Rudd. 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. ISSN 0002-9297. doi: 10.1016/j.ajhg.2014.12.017. URL http://dx.doi.org/10.1016/j.ajhg.2014.12.017.

Sarah Djebali, Carrie a. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian a. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489 (7414):101–108, 2012. ISSN 0028-0836. doi: 10.1038/nature11233.

Alistair R R Forrest, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Timo Lassmann, Masayoshi Itoh, Kim M Summers, Harukazu Suzuki, Carsten O Daub, Jun Kawai, Peter Heutink, Winston Hide, Tom C Freeman, Boris Lenhard, Vladimir B Bajic, Martin S Taylor, Vsevolod J Makeev, Albin Sandelin,

David a Hume, Piero Carninci, and Yoshihide Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70, mar 2014. ISSN 1476-4687. doi: 10.1038/nature13182. URL http://www.ncbi.nlm.nih.gov/pubmed/24670764.

Kristin G. Ardlie, David S. DeLuca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, Julian B. Maller, Taru Tukiainen, Monkol Lek, Lucas D. Ward, Pouya Kheradpour, Benjamin Iriarte, Yan Meng, Cameron D. Palmer, Tõnu Esko, Wendy Winckler, Joel N. Hirschhorn, Manolis Kellis, Daniel G. MacArthur, Gad Getz, Andrey A. Shabalin, Gen Li, Yi Hui Zhou, Andrew B. Nobel, Ivan Rusyn, Fred A. Wright, Tuuli Lappalainen, Pedro G. Ferreira, Halit Ongen, Manuel A. Rivas, Alexis Battle, Sara Mostafavi, Jean Monlong, Michael Sammeth, Marta Melé, Ferran Reverter, Jakob M. Goldmann, Daphne Koller, Roderic Guigó, Mark I. McCarthy, Emmanouil T. Dermitzakis, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Dan L. Nicolae, Nancy J. Cox, Timothée Flutre, Xiaoquan Wen, Matthew Stephens, Jonathan K. Pritchard, Zhidong Tu, Bin Zhang, Tao Huang, Quan Long, Luan Lin, Jialiang Yang, Jun Zhu, Jun Liu, Amanda Brown, Bernadette Mestichelli, Denee Tidwell, Edmund Lo, Michael Salvatore, Saboor Shad, Jeffrey A. Thomas, John T. Lonsdale, Michael T. Moser, Bryan M. Gillard, Ellen Karasik, Kimberly Ramsey, Christopher Choi, Barbara A. Foster, John Syron, Johnell Fleming, Harold Magazine, Rick Hasz, Gary D. Walters, Jason P. Bridge, Mark Miklos, Susan Sullivan, Laura K. Barker, Heather M. Traino, Maghboeba Mosavel, Laura A. Siminoff, Dana R. Valley, Daniel C. Rohrer, Scott D. Jewell, Philip A. Branton, Leslie H. Sobin, Mary Barcus, Liqun Qi, Jeffrey McLean, Pushpa Hariharan, Ki Sung Um, Shenpei Wu, David Tabor, Charles Shive, Anna M. Smith, Stephen A. Buia, Anita H. Undale, Karna L. Robinson, Nancy Roche, Kimberly M. Valentino, Angela Britton, Robin Burges, Debra Bradbury, Kenneth W. Hambright, John Seleski, Greg E. Korzeniewski, Kenyon Erickson, Yvonne Marcus, Jorge Tejada, Mehran Taherian, Chunrong Lu, Margaret Basile, Deborah C. Mash, Simona Volpi, Jeffery P. Struewing, Gary F. Temple, Joy Boyer, Deborah Colantuoni, Roger Little, Susan Koester, Latarsha J. Carithers, Helen M. Moore, Ping Guan, Carolyn Compton, Sherilyn J. Sawyer, Joanne P. Demchok, Jimmie B. Vaught, Chana A. Rabiner, and Lockhart. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, may 2015. ISSN 10959203. doi: 10.1126/science.1262110. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1262110http://www.ncbi.nlm.nih.gov/pubmed/25954001http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4547484.

Robert Petryszak, Maria Keays, Y. Amy Tang, Nuno A. Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, Alfonso Muñoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, Oliver Mannion, Laura Huerta, Karine Megy, Catherine Snow,

Eleanor Williams, Mitra Barzine, Emma Hastings, Hendrik Weisser, James Wright, Pankaj Jaiswal, Wolfgang Huber, Jyoti Choudhary, Helen E. Parkinson, and Alvis Brazma. Expression Atlas updatean integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44(October 2015):gkv1045, 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv1045. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1045.

Thomas Cremer, Marion Cremer, Barbara Hübner, Hilmar Strickfaden, Daniel Smeets, Jens Popken, Michael Sterr, Yolanda Markaki, Karsten Rippe, and Christoph Cremer. The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS Letters*, 2015. ISSN 00145793. doi: 10.1016/j.febslet.2015.05.037. URL http://dx.doi.org/10.1016/j.febslet.2015.05.037.

a M Boutanaev, a I Kalmykova, Y Y Shevelyov, D I Nurminsky, Maynard Smith, The Evolution, The Masterpiece, Croom Helm, and Natural Selection. Large clusters of co-expressed genes in the Drosophila genome. *Nature*, 420(December):666–669, 2002. ISSN 0028-0836. doi: 10.1038/nature01191.1.

Antje Purmann, Joern Toedling, Markus Schueler, Piero Carninci, Hans Lehrach, Yoshihide Hayashizaki, Wolfgang Huber, and Silke Sperling. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics*, 89(5):580–587, 2007. ISSN 08887543. doi: 10.1016/j.ygeno.2007.01.010.

Duncan Sproul, Nick Gilbert, and Wendy a Bickmore. The role of chromatin structure in regulating the expression of clustered genes. *Nature reviews. Genetics*, 6 (10):775–781, 2005. ISSN 1471-0056. doi: 10.1038/nrg1688.

Matthias Becker, Nancy Mah, Daniela Zdzieblo, Xiaoli Li, Arvind Mer, Miguel A Andrade-navarro, and Albrecht M Mu. Epigenetic Mechanisms in Cellular Reprogramming. In Alexander Meissner and Jörn Walter, editors, *Epigenetics and Human Health*, Epigenetics and Human Health, pages pp 141–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-642-31973-0. doi: 10.1007/978-3-642-31974-7. URL http://link.springer.com/10.1007/978-3-642-31974-7.

L. Huminiecki. Divergence of Spatial Gene Expression Profiles Following Species-Specific Gene Duplications in Human and Mouse. *Genome Research*, 14(10a): 1870–1879, 2004. ISSN 1088-9051. doi: 10.1101/gr.2705204. URL http://www.genome.org/cgi/doi/10.1101/gr.2705204.

Igor B. Rogozin, David Managadze, Svetlana A. Shabalina, and Eugene V. Koonin. Gene family level comparative analysis of gene expression n mammals validates

the ortholog conjecture. *Genome Biology and Evolution*, 6(4):754–762, 2014. ISSN 17596653. doi: 10.1093/gbe/evu051.

Martin H Schaefer, Jean-Fred Fontaine, Arunachalam Vinayagam, Pablo Porras, Erich E Wanker, and Miguel a Andrade-Navarro. HIPPIE: Integrating protein interaction networks with experiment based quality scores. *PloS one*, 7(2): e31826, jan 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0031826. URL 10.1371/journal.pone.0031826http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3279424{&}tool=pmcentrez{&}rendertype=abstract.

Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 24(3):390–400, mar 2014. ISSN 1549-5469. doi: 10.1101/gr.163519.113. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3941104{&}tool=pmcentrez{&}rendertype=abstract.

Yinxiu Zhan, Luca Mariani, Iros Barozzi, Edda G Schulz, Nils Blüthgen, Michael Stadler, Guido Tiana, Luca Giorgetti, Nils Bluthgen, Michael Stadler, Guido Tiana, and Luca Giorgetti. 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. ISSN 1088-9051. doi: 10.1101/gr.212803.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.212803.116.

Carlos Gómez-Marín, Juan J. Tena, Rafael D. Acemel, Macarena López-Mayorga, Silvia Naranjo, Elisa de la Calle-Mustienes, Ignacio Maeso, Leonardo Beccari, Ivy Aneas, Erika Vielmas, Paola Bovolenta, Marcelo a. Nobrega, Jaime Carvajal, and José Luis Gómez-Skarmeta. 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. ISSN 0027-8424. doi: 10.1073/pnas.1505463112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1505463112.

Tsung-HanăS. Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and OliverăJ. Rando. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, pages 1–12, 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.05.048. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415006388.

Takeshi Mizuguchi, Geoffrey Fudenberg, Sameet Mehta, Jon-Matthew Belton, Nitika Taneja, Hernan Diego Folco, Peter FitzGerald, Job Dekker, Leonid Mirny, Jemima Barrowman, and Shiv I. S. Grewal. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, oct 2014. ISSN

0028-0836. doi: 10.1038/nature13833. URL http://www.nature.com/doifinder/10.1038/nature13833.

Elphège P. Nora, Job Dekker, and Edith Heard. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? *BioEssays*, 35 (9):818–828, 2013. ISSN 02659247. doi: 10.1002/bies.201300040.

W James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. 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. ISSN 0027-8424. doi: 10.1073/pnas.1932072100. URL http://www.ncbi.nlm.nih.gov/pubmed/14500911{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC208784.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome research*, 12(6):996–1006, jun 2002. ISSN 1088-9051. doi: 10.1101/gr.229102.Articlepublishedonlinebeforeprintin May2002. URL http://www.ncbi.nlm.nih.gov/pubmed/12045153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC186604.

Ryan E Mills, E Andrew Bennett, Rebecca C Iskow, Christopher T Luttig, Circe Tsui, W Stephen Pittard, and Scott E Devine. Recently Mobilized Transposons in the Human and Chimpanzee Genomes. *The American Journal of Human Genetics*, 78(4):671–679, 2006. ISSN 00029297. doi: 10.1086/501028. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929707637045.

Marta Farré, Terence J. Robinson, and Aurora Ruiz-Herrera. An Integrative Breakage Model of genome architecture, reshuffling and evolution. *BioEssays*, pages n/a–n/a, 2015. ISSN 02659247. doi: 10.1002/bies.201400174. URL http://doi.wiley.com/10.1002/bies.201400174.

Dimitris Polychronopoulos, JamesăW.ăD. King, Alexander J. Nash, Ge Tan, and Boris Lenhard. Conserved non-coding elements: developmental gene regulation meets genome organization. *Nucleic Acids Research*, (November):1–14, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1074. URL http://academic.oup.com/nar/article/doi/10.1093/nar/gkx1074/4599184.

Hiroshi Kikuta, Mary Laplante, Pavla Navratilova, Anna Z. Komisarczuk, Pär G. Engström, David Fredman, Altuna Akalin, Mario Caccamo, Ian Sealy, Kerstin Howe, Julien Ghislain, Guillaume Pezeron, Philippe Mourrain, Staale Ellingsen, Andrew C. Oates, Christine Thisse, Bernard Thisse, Isabelle Foucher, Birgit Adolf,

Andrea Geling, Boris Lenhard, and Thomas S. Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Research*, 17(5):545–555, 2007. ISSN 10889051. doi: 10.1101/gr.6086307.

Nathan Harmston, Elizabeth Ing-Simmons, Ge Tan, Malcolm Perry, Matthias Merkenschlager, and Boris Lenhard. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nature Communications*, 8(1):441, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00524-5. URL http://www.nature.com/articles/s41467-017-00524-5.

Pär G. Engström, Shannan J Ho Sui, Øyvind Drivenes, Thomas S. Becker, and Boris Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research*, 17(12):1898–1908, 2007. ISSN 10889051. doi: 10.1101/gr.6669607.

Slavica Dimitrieva and Philipp Bucher. UCNEbasea database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research*, 41(D1):D101–D109, jan 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1092. URL http://academic.oup.com/nar/article/41/D1/D101/1057253/UCNEbasea-database-of-ultraconserved-noncoding.

Andres Canela, Yaakov Maman, Seolkyoung Jung, Nancy Wong, Elsa Callen, Amanda Day, Kyong-Rim Kieffer-Kwon, Aleksandra Pekowska, Hongliang Zhang, Suhas S.P. Rao, Su-chen Huang, Peter J. Mckinnon, Peter D. Aplan, Yves Pommier, Erez Lieberman Aiden, Rafael Casellas, and André Nussenzweig. Genome Organization Drives Chromosome Fragility. *Cell*, pages 1–15, jul 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.06.034. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867417307183.

Claire Redin, Harrison Brand, Ryan L Collins, Tammy Kammin, Elyse Mitchell, Jennelle C Hodge, Carrie Hanscom, Vamsee Pillalamarri, Catarina M Seabra, Mary-Alice Abbott, Omar A Abdul-Rahman, Erika Aberg, Rhett Adley, Sofia L Alcaraz-Estrada, Fowzan S Alkuraya, Yu An, Mary-Anne Anderson, Caroline Antolik, Kwame Anyane-Yeboa, Joan F Atkin, Tina Bartell, Jonathan A Bernstein, Elizabeth Beyer, Ian Blumenthal, Ernie M H F Bongers, Eva H Brilstra, Chester W Brown, Hennie T Brüggenwirth, Bert Callewaert, Colby Chiang, Ken Corning, Helen Cox, Edwin Cuppen, Benjamin B Currall, Tom Cushing, Dezso David, Matthew A Deardorff, Annelies Dheedene, Marc D'Hooghe, Bert B A de Vries, Dawn L Earl, Heather L Ferguson, Heather Fisher, David R FitzPatrick, Pamela Gerrol, Daniela Giachino, Joseph T Glessner, Troy Gliem, Margo Grady, Brett H

Graham, Cristin Griffis, Karen W Gripp, Andrea L Gropman, Andrea Hanson-Kahn, David J Harris, Mark A Hayden, Rosamund Hill, Ron Hochstenbach, Jodi D Hoffman, Robert J Hopkin, Monika W Hubshman, A Micheil Innes, Mira Irons, Melita Irving, Jessie C Jacobsen, Sandra Janssens, Tamison Jewett, John P Johnson, Marjolijn C Jongmans, Stephen G Kahler, David A Koolen, Jerome Korzelius, Peter M Kroisel, Yves Lacassie, William Lawless, Emmanuelle Lemyre, Kathleen Leppig, Alex V Levin, Haibo Li, Hong Li, Eric C Liao, Cynthia Lim, Edward J Lose, Diane Lucente, Michael J Macera, Poornima Manavalan, Giorgia Mandrile, Carlo L Marcelis, Lauren Margolin, Tamara Mason, Diane Masser-Frye, Michael W McClellan, Cinthya J Zepeda Mendoza, Björn Menten, Sjors Middelkamp, Liya R Mikami, Emily Moe, Shehla Mohammed, Tarja Mononen, Megan E Mortenson, Graciela Moya, Aggie W Nieuwint, Zehra Ordulu, Sandhya Parkash, Susan P Pauker, Shahrin Pereira, Danielle Perrin, Katy Phelan, Raul E Piña Aguilar, Pino J Poddighe, Giulia Pregno, Salmo Raskin, Linda Reis, William Rhead, Debra Rita, Ivo Renkens, Filip Roelens, Jayla Ruliera, Patrick Rump, Samantha L P Schilit, Ranad Shaheen, Rebecca Sparkes, Erica Spiegel, Blair Stevens, Matthew R Stone, Julia Tagoe, Joseph V Thakuria, Bregje W van Bon, Jiddeke van de Kamp, Ineke van Der Burgt, Ton van Essen, Conny M van Ravenswaaij-Arts, Markus J van Roosmalen, Sarah Vergult, Catharina M L Volker-Touw, Dorothy P Warburton, Matthew J Waterman, Susan Wiley, Anna Wilson, Maria de la Concepcion A Yerena-de Vega, Roberto T Zori, Brynn Levy, Han G Brunner, Nicole de Leeuw, Wigard P Kloosterman, Erik C Thorland, Cynthia C Morton, James F Gusella, and Michael E Talkowski. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature Genetics*, 2016. ISSN 1061-4036. doi: 10.1038/ng.3720. URL http://www.nature.com/doifinder/10.1038/ng.3720.

Stefan Schoenfelder, Mayra Furlan-magaril, Borbala Mifsud, Filipe Tavares-cadete, Robert Sugar, Biola-maria Javierre, Takashi Nagano, Yulia Katsman, Moorthy Sakthidevi, Steven W Wingett, Emilia Dimitrova, Andrew Dimond, Lucas B Edelman, Sarah Elderkin, Kristina Tabbada, Elodie Darbo, Simon Andrews, Bram Herman, Andy Higgs, Emily Leproust, Cameron S Osborne, Jennifer A Mitchell, Nicholas M Luscombe, and Peter Fraser. 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.

Thomas Montavon, Laurie Thevenet, and Denis Duboule. 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. ISSN 1091-6490. doi: 10.1073/pnas.1217659109. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3528568{&}tool=pmcentrez{&}rendertype=abstract.

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, and Cynthia C. Morton. Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements. *American journal of human genetics*, 101(2):206–217, aug 2017. ISSN 1537-6605. doi: 10.1016/j.ajhg.2017.06.011. URL http://www.sciencedirect.com/science/article/pii/S000292971730246Xhttp://www.ncbi.nlm.nih.gov/pubmed/28735859.

Malte Spielmann, Francesco Brancati, Peter M Krawitz, Peter N Robinson, Daniel M Ibrahim, Martin Franke, Jochen Hecht, Silke Lohan, Katarina Dathe, Anna Maria Nardone, Paola Ferrari, Antonio Landi, Lars Wittler, Bernd Timmermann, Danny Chan, Ulrich Mennen, Eva Klopocki, and Stefan Mundlos. Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. *American journal of human genetics*, 91(4):629–35, oct 2012. ISSN 1537-6605. doi: 10.1016/j.ajhg.2012.08.014. URL http://www.ncbi.nlm.nih.gov/pubmed/23022097.

Pavel Pevzner and Glenn Tesler. 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. ISSN 0027-8424. doi: 10.1073/pnas.1330369100. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=164646{&}tool=pmcentrez{&}rendertype=abstract{%}5Cnhttp://www.pnas.org/content/100/13/7672.

Chunhui Hou, Li Li, Zhaohui S. Qin, and Victor G. Corces. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, 48(3):471–484, 2012. ISSN 10972765. doi: 10.1016/j.molcel.2012.08.031.

William J Murphy, Denis M Larkin, Annelie Everts-van der Wind, Guillaume Bourque, Glenn Tesler, Loretta Auvil, Jonathan E Beever, Bhanu P Chowdhary, Francis Galibert, Lisa Gatzke, Christophe Hitte, Stacey N Meyers, Denis Milan, Elaine A Ostrander, Greg Pape, Heidi G Parker, Terje Raudsepp, Margarita B Rogatcheva, Lawrence B Schook, Loren C Skow, Michael Welge, James E Womack, Stephen J O'brien, Pavel A Pevzner, and Harris A Lewin. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science (New York, N.Y.)*, 309(5734):613–7, 2005. ISSN 1095-9203. doi: 10.1126/science.1111387. URL http://www.ncbi.nlm.nih.gov/pubmed/16040707.

Hanno Hinsch and Sridhar Hannenhalli. Recurring genomic breaks in independent lineages support genomic fragility. *BMC evolutionary biology*, 6:90, 2006. ISSN 1471-2148. doi: 10.1186/1471-2148-6-90. URL http://www.ncbi.nlm.nih.gov/pubmed/17090315.

Martin Franke, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerković, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, (Idi):1–15, 2016. ISSN 0028-0836. doi: 10.1038/nature19800. URL http://www.nature.com/doifinder/10.1038/nature19800.

Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H Porteus, Job Dekker, and Richard A Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, 351(6280): 1454–8, mar 2016. ISSN 1095-9203. doi: 10.1126/science.aad9024.

Paul a. Northcott, Catherine Lee, Thomas Zichner, Adrian M. Stütz, Serap Erkek, Daisuke Kawauchi, David J. H. Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, David T. W. Jones, Marcel Kool, Marc Remke, Florence M. G. Cavalli, Scott Zuyderduyn, Gary D. Bader, Scott VandenBerg, Lourdes Adriana Esparza, Marina Ryzhova, Wei Wang, Andrea Wittmann, Sebastian Stark, Laura Sieber, Huriye Seker-Cin, Linda Linke, Fabian Kratochwil, Natalie Jäger, Ivo Buchhalter, Charles D. Imbusch, Gideon Zipprich, Benjamin Raeder, Sabine Schmidt, Nicolle Diessl, Stephan Wolf, Stefan Wiemann, Benedikt Brors, Chris Lawerenz, Jürgen Eils, Hans-Jörg Warnatz, Thomas Risch, Marie-Laure Yaspo, Ursula D. Weber, Cynthia C. Bartholomae, Christof von Kalle, Eszter Turányi, Peter Hauser, Emma Sanden, Anna Darabi, Peter Siesjö, Jaroslav Sterba, Karel Zitterbart, David Sumerauer, Peter van Sluis, Rogier Versteeg, Richard Volckmann, Jan Koster, Martin U. Schuhmann, Martin Ebinger, H. Leighton Grimes, Giles W. Robinson, Amar Gajjar, Martin Mynarek, Katja von Hoff, Stefan Rutkowski, Torsten Pietsch, Wolfram Scheurlen, Jörg Felsberg, Guido Reifenberger, Andreas E. Kulozik, Andreas von Deimling, Olaf Witt, Roland Eils, Richard J. Gilbertson, Andrey Korshunov, Michael D. Taylor, Peter Lichter, Jan O. Korbel, Robert J. Wechsler-Reya, and Stefan M. Pfister. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, jun 2014. ISSN 0028-0836. doi: 10.1038/nature13379. URL http://www.nature.com/doifinder/10.1038/nature13379.

Joachim Weischenfeldt, Taronish Dubash, Alexandros P Drainas, Balca R Mardin, Yuanyuan Chen, Adrian M Stütz, Sebastian M Waszak, Graziella Bosco, Ann Rita Halvorsen, Benjamin Raeder, Theocharis Efthymiopoulos, Serap Erkek, Christine Siegl, Hermann Brenner, Odd Terje Brustugun, Sebastian M Dieter, Paul A Northcott, Iver Petersen, Stefan M Pfister, Martin Schneider, Steinar K Solberg, Erik Thunissen, Wilko Weichert, Thomas Zichner, Roman Thomas, Martin Peifer, Aslaug Helland, Claudia R Ball, Martin Jechlinger, Rocio Sotillo, Hanno Glimm, and Jan O Korbel. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nature Genetics*, (November), 2016. ISSN 1061-4036. doi: 10.1038/ng.3722. URL http://www.nature.com/doifinder/10.1038/ng.3722.

Rafael D. Acemel, Ignacio Maeso, and José Luis GómezSkarmeta. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. *Wiley Interdisciplinary Reviews: Developmental Biology*, pages 1–19, 2017. ISSN 1759-7692. doi: 10.1002/WDEV.265. URL http://onlinelibrary.wiley.com/doi/10.1002/wdev.265/full.

Sean B. Carroll. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, 134(1):25–36, jul 2008. ISSN 0092-8674. doi: 10.1016/J.CELL.2008.06.030. URL https://www.sciencedirect.com/science/article/pii/S0092867408008179.

Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, Stephen M. J. Searle, Ridwan Amode, Simon Brent, William Spooner, Eugene Kulesha, Andrew Yates, and Paul Flicek. Ensembl comparative genomics resources. *Database*, 2016:bav096, feb 2016. ISSN 1758-0463. doi: 10.1093/database/bav096. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bav096.

Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3252. URL http://www.nature.com/doifinder/10.1038/nmeth.3252.

M. Lawrence, R. Gentleman, and V. Carey. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842, jul 2009. ISSN

1367-4803. doi: 10.1093/bioinformatics/btp328. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp328.

Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, aug 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003118. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3738458{&}tool=pmcentrez{&}rendertype=abstract.

Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191, aug 2009b. ISSN 1754-2189. doi: 10.1038/nprot.2009.97. URL http://www.nature.com/articles/nprot.2009.97.

Hadley Wickham and Garrett Grolemund. *R for data science : import, tidy, transform, visualize, and model data*. ISBN 1491910399.

A John Iafrate, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9):949–951, sep 2004. ISSN 1061-4036. doi: 10.1038/ng1416. URL http://www.nature.com/articles/ng1416.

Feng Zhang, Wenli Gu, Matthew E Hurles, and James R Lupski. Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481, sep 2009. ISSN 1527-8204. doi: 10.1146/annurev.genom.9.081307.164217. URL http://dx.doi.org/10.1146/annurev.genom.9.081307.164217.

Aaron Theisen and Lisa G Shaffer. Disorders caused by chromosome abnormalities. *The application of clinical genetics*, 3:159–74, dec 2010. ISSN 1178-704X. doi: 10.2147/TACG.S8884. URL http://www.dovepress.com/disorders-caused-by-chromosome-abnormalities-peer-reviewed-article-TACGhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3681172{&}tool=pmcentrez{&}rendertype=abstract.

Dirk A Kleinjan and Veronica Van Heyningen. Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease. *Am. J. Hum. Genet*, 76:8–32, 2005. ISSN 00029297. doi: 10.1086/426833.

Karen S Weiler and Barbara T Wakimoto. Heterochromatin and gene expression in Drosophila. *Annual review of genetics*, 29:577–605, 1995. ISSN 0066-

4197. doi: 10.1146/annurev.ge.29.120195.003045. URL http://www.ncbi.nlm.nih.gov/pubmed/8825487.

Bin Zhang and Peter G. Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences*, 112(19): 201506257, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1506257112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1506257112.

Malte Spielmann and Stefan Mundlos. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics*, page ddw205, 2016. ISSN 0964-6906. doi: 10.1093/hmg/ddw205. URL http://www.hmg.oxfordjournals.org/lookup/doi/10.1093/hmg/ddw205.

Judy Fantes, Bert Redeker, Matthew Breen, Shelagh Boyle, John Brown, Judy Fletcher, Sinead Jones, Wendy Bickmore, Yoshimitsu Fukushima, Marcel Mannens, Sarah Danes, Veronica van Heyningen, and Isabel Hanson. Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Human Molecular Genetics*, 4(3):415–422, mar 1995. ISSN 0964-6906. doi: 10.1093/hmg/4.3.415. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/4.3.415.

Juanliang Cai, Barbara K. Goodman, Ankita S. Patel, John B. Mulliken, Lionel Van Maldergem, George E. Hoganson, William A. Paznekas, Ziva Ben-Neriah, Ruth Sheffer, Michael L. Cunningham, Donna L. Daentl, and Ethylin Wang Jabs. 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, dec 2003. ISSN 0340-6717. doi: 10.1007/s00439-003-1012-7. URL http://link.springer.com/10.1007/s00439-003-1012-7.

Rachel H Flomen, Radost Vatcheva, Patricia A Gorman, Pedro R Baptista, Juergen Groet, Ingeborg Barišić, Ivo Ligutic, and Dean Nižetić. 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, feb 1998. ISSN 0888-7543. doi: 10.1006/GENO.1997.5127. URL https://www.sciencedirect.com/science/article/pii/S0888754397951272.

Gopalrao V.N. Velagaleti, Gabriel A. Bien-Willner, Jill K. Northup, Lillian H. Lockhart, Judy C. Hawkins, Syed M. Jalal, Marjorie Withers, James R. Lupski, and Pawel Stankiewicz. 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, apr

2005. ISSN 0002-9297. doi: 10.1086/429252. URL https://www.sciencedirect.com/science/article/pii/S0002929707628766.

D. Kleinjan and Veronica van Heyningen. Position effect in human genetic disease. *Human Molecular Genetics*, 7(10):1611–1618, sep 1998. ISSN 14602083. doi: 10.1093/hmg/7.10.1611. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/7.10.1611.

Jennifer E Phillips-Cremins, Michael E G Sauria, Amartya Sanyal, Tatiana I Gerasimova, Bryan R Lajoie, Joshua S K Bell, Chin-Tong Ong, Tracy a Hookway, Changying Guo, Yuhua Sun, Michael J Bland, William Wagstaff, Stephen Dalton, Todd C McDevitt, Ranjan Sen, Job Dekker, James Taylor, and Victor G Corces. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–95, jun 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.04.053. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3712340{&}tool=pmcentrez{&}rendertype=abstract.

Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*, 373(10):895–907, 2015. ISSN 0028-4793. doi: 10.1056/NEJMoa1502214. URL http://www.nejm.org/doi/10.1056/NEJMoa1502214.

Mijke Visser, Manfred Kayser, and Robert-Jan Palstra. 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, mar 2012. ISSN 1549-5469. doi: 10.1101/gr.128652.111. URL http://www.ncbi.nlm.nih.gov/pubmed/22234890http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3290780.

Derek A. Oldridge, Andrew C. Wood, Nina Weichert-Leahey, Ian Crimmins, Robyn Sussman, Cynthia Winter, Lee D. McDaniel, Maura Diamond, Lori S. Hart, Shizhen Zhu, Adam D. Durbin, Brian J. Abraham, Lars Anders, Lifeng Tian, Shile Zhang, Jun S. Wei, Javed Khan, Kelli Bramlett, Nazneen Rahman, Mario Capasso, Achille Iolascon, Daniela S. Gerhard, Jaime M. Guidry Auvil, Richard A. Young, Hakon Hakonarson, Sharon J. Diskin, A. Thomas Look, and John M. Maris. Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature*, 528(7582):418–421, 2015. ISSN 0028-0836. doi: 10.1038/nature15540. URL http://www.nature.com/doifinder/10.1038/nature15540.

Anne W. Higgins, Fowzan S. Alkuraya, Amy F. Bosco, Kerry K. Brown, Gail A.P. Bruns, Diana J. Donovan, Robert Eisenman, Yanli Fan, Chantal G. Farra, Heather L. Ferguson, James F. Gusella, David J. Harris, Steven R. Herrick, Chantal Kelly, Hyung Goo Kim, Shotaro Kishikawa, Bruce R. Korf, Shashikant Kulkarni, Eric Lally, Natalia T. Leach, Emma Lemyre, Janine Lewis, Azra H. Ligon, Weining Lu, Richard L. Maas, Marcy E. MacDonald, Steven D.P. Moore, Roxanna E. Peters, Bradley J. Quade, Fabiola Quintero-Rivera, Irfan Saadi, Yiping Shen, Jay Shendure, Robin E. Williamson, and Cynthia C. Morton. Characterization of Apparently Balanced Chromosomal Rearrangements from the Developmental Genome Anatomy Project. *American Journal of Human Genetics*, 82(3):712–722, 2008. ISSN 00029297. doi: 10.1016/j.ajhg.2008.01.011.

Azra H. Ligon, Steven D.P. Moore, Melissa A. Parisi, Matthew E. Mealiffe, David J. Harris, Heather L. Ferguson, Bradley J. Quade, and Cynthia C. Morton. 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, feb 2005. ISSN 0002-9297. doi: 10.1086/427565. URL https://www.sciencedirect.com/science/article/pii/S0002929707625853.

Wan Kyu Kim and Edward M Marcotte. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. *PLoS computational biology*, 4(11):e1000232, nov 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000232. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2583957{&}tool=pmcentrez{&}rendertype=abstract.

Jeffrey J. Quinn and Howard Y. Chang. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47–62, jan 2016. ISSN 1471-0056. doi: 10.1038/nrg.2015.10. URL http://www.nature.com/articles/nrg.2015.10.

Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle. Ensembl 2014. *Nucleic Acids Research*, 42(D1):

D749–D755, jan 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1196. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1196.

Ni Huang, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. Characterising and Predicting Haploinsufficiency in the Human Genome. *PLoS Genetics*, 6(10): e1001154, oct 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001154. URL http://dx.plos.org/10.1371/journal.pgen.1001154.

Heidi L. Rehm, Jonathan S. Berg, Lisa D. Brooks, Carlos D. Bustamante, James P. Evans, Melissa J. Landrum, David H. Ledbetter, Donna R. Maglott, Christa Lese Martin, Robert L. Nussbaum, Sharon E. Plon, Erin M. Ramos, Stephen T. Sherry, and Michael S. Watson. ClinGen The Clinical Genome Resource. *New England Journal of Medicine*, 372(23):2235–2242, jun 2015. ISSN 0028-4793. doi: 10.1056/NEJMsr1406261. URL http://www.nejm.org/doi/10.1056/NEJMsr1406261.

Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O Dorschner, R Scott Hansen, Patrick a Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John a Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, sep 2012. ISSN 1476-4687. doi: 10.1038/nature11232. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3721348{&}tool=pmcentrez{&}rendertype=abstract.

Varun Narendra, Pedro P Rocha, Disi An, Ramya Raviram, Jane A Skok, Esteban O Mazzoni, and Danny Reinberg. Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science (New York, N.Y.)*, 347(6225):1017–21, 2015. ISSN 1095-9203. doi: 10.1126/science.1262088. URL http://www.ncbi.nlm.nih.gov/pubmed/25722416.

W a Flavahan, Y Drier, B B Liau, S M Gillespie, a S Venteicher, a O Stemmer-Rachamimov, M L Suva, and B E Bernstein. Insulator dysfunction and oncogene

activation in IDH mutant gliomas. *Nature*, 529(7584):110–114, 2016. ISSN 0028-0836. doi: 10.1038/nature16490. URL http://www.ncbi.nlm.nih.gov/pubmed/26700815.

Andreas Heger, Caleb Webber, Martin Goodson, Chris P Ponting, and Gerton Lunter. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics (Oxford, England)*, 29(16):2046–8, aug 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt343. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3722528{&}tool=pmcentrez{&}rendertype=abstract.

Sebastian Köhler, Sandra C. Doelken, Christopher J. Mungall, Sebastian Bauer, Helen V. Firth, Isabelle Bailleul-Forestier, Graeme C M Black, Danielle L. Brown, Michael Brudno, Jennifer Campbell, David R. Fitzpatrick, Janan T. Eppig, Andrew P. Jackson, Kathleen Freson, Marta Girdea, Ingo Helbig, Jane A. Hurst, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, Peter N. Robinson, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole De Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, Peter N. Robinson, Johanna J??hn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole De Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, Peter N. Robinson, Johanna Jähn, Laird G. Jackson, Anne M. Kelly, David H. Ledbetter, Sahar Mansour, Christa L. Martin, Celia Moss, Andrew Mumford, Willem H. Ouwehand, Soo-Mi Mi Park, Erin Rooney Riggs, Richard H. Scott, Sanjay Sisodiya, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole de Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, Peter N. Robinson, Steven Van Vooren, Ronald J. Wapner, Andrew O M Wilkie, Caroline F. Wright, Anneke T. Vulto-van Silfhout, Nicole De Leeuw, Bert B A De Vries, Nicole L. Washingthon, Cynthia L. Smith, Monte Westerfield, Paul

Schofield, Barbara J. Ruef, Georgios V. Gkoutos, Melissa Haendel, Damian Smedley, Suzanna E. Lewis, and Peter N. Robinson. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(Database issue):D966–74, jan 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1026. URL http://dx.doi.org/10.1093/nar/gkt1026.

Helen V Firth, Shola M Richards, a Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American journal of human genetics*, 84(4):524–33, apr 2009. ISSN 1537-6605. doi: 10.1016/j.ajhg.2009.03.010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2667985{&}tool=pmcentrez{&}rendertype=abstract.

Zehra Ordulu, Tammy Kammin, Harrison Brand, Vamsee Pillalamarri, Claire E Redin, Ryan L Collins, Ian Blumenthal, Carrie Hanscom, Shahrin Pereira, Barbara F Crandall, Pamela Gerrol, Mark A Hayden, Naveed Hussain, Bibi Kanengisser-pines, Sibel Kantarci, Brynn Levy, Michael J Macera, Fabiola Quintero-rivera, Erica Spiegel, Blair Stevens, Janet E Ulm, Dorothy Warburton, Louise E Wilkins-haug, Naomi Yachelevich, James F Gusella, and Michael E Talkowski. 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. ISSN 0002-9297. doi: 10.1016/j.ajhg.2016.08.022. URL http://dx.doi.org/10.1016/j.ajhg.2016.08.022.

HannahăK. Long, SaraăL. Prescott, and Joanna Wysocka. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, 167 (5):1170–1187, 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.09.018. URL http://linkinghub.elsevier.com/retrieve/pii/S009286741631251X.

Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, and Netta Mendelson-cohen. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature Publishing Group*, 547(7661):61–67, 2017. ISSN 0028-0836. doi: 10.1038/nature23001. URL http://dx.doi.org/10.1038/nature23001.

Monika Sekelja, Jonas Paulsen, and Philippe Collas. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? *Genome Biology*, 17(1):54, 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0923-2. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0923-2.

Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013. ISSN 01689525. doi: 10.1016/j.tig.2013.05.010. URL http://dx.doi.org/10.1016/j.tig.2013.05.010.