Krefting et al. BMC Biology (2018) 16:87
https://doi.org/10.1186/s12915-018-0556-x

BMC Biology

RESEARCH ARTICLE Open Access

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*]

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 stable 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

Background

The three-dimensional structure of eukaryotic genomes is organized in many hierarchical levels [1]. The development of high-throughput experiments to measure pairwise chromatin-chromatin interactions, such as Hi-C [2], enabled the identification of genomic domains of several hundred kilo-bases with increased self-interaction frequencies, described as topologically associating domains (TADs) [3–5]. 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 [4, 6, 7]. Several studies associated disruption of TADs to ectopic regulation of important developmental genes leading to genetic diseases [8–10]. These properties of TADs suggested that they are functional genomic units of gene regulation.

Interestingly, TADs are largely stable across cell types [3, 11] and during differentiation [12]. 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* [5], zebrafish [13], *Caenorhabditis elegans* [14], and yeast [15, 16]. 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

* Correspondence: j.ibn-salem@uni-mainz.de
[1]Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany
[2]Institute of Molecular Biology, 55128 Mainz, Germany

along evolution. More than 54% of TAD boundaries in human cells occur at homologous positions in mouse genomes [3]. Similarly, 45% of contact domains called in mouse B-lymphoblasts were also identified at homologous regions in human lymphoblastoid cells [11]. A single TAD boundary at the six gene loci could be traced back in evolution to the origin of deuterostomes [13]. 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 [17]. 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 [18]. 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 [19, 20] to compare the human genome to 12 other species. These species were selected to have genome assemblies of good quality and to span several hundred million years of evolution. They range from chimpanzee to zebrafish (Fig. 1). The whole-genome data consists of consecutive alignment blocks that are chained and hierarchically ordered into so-called net files as fills [19]. 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

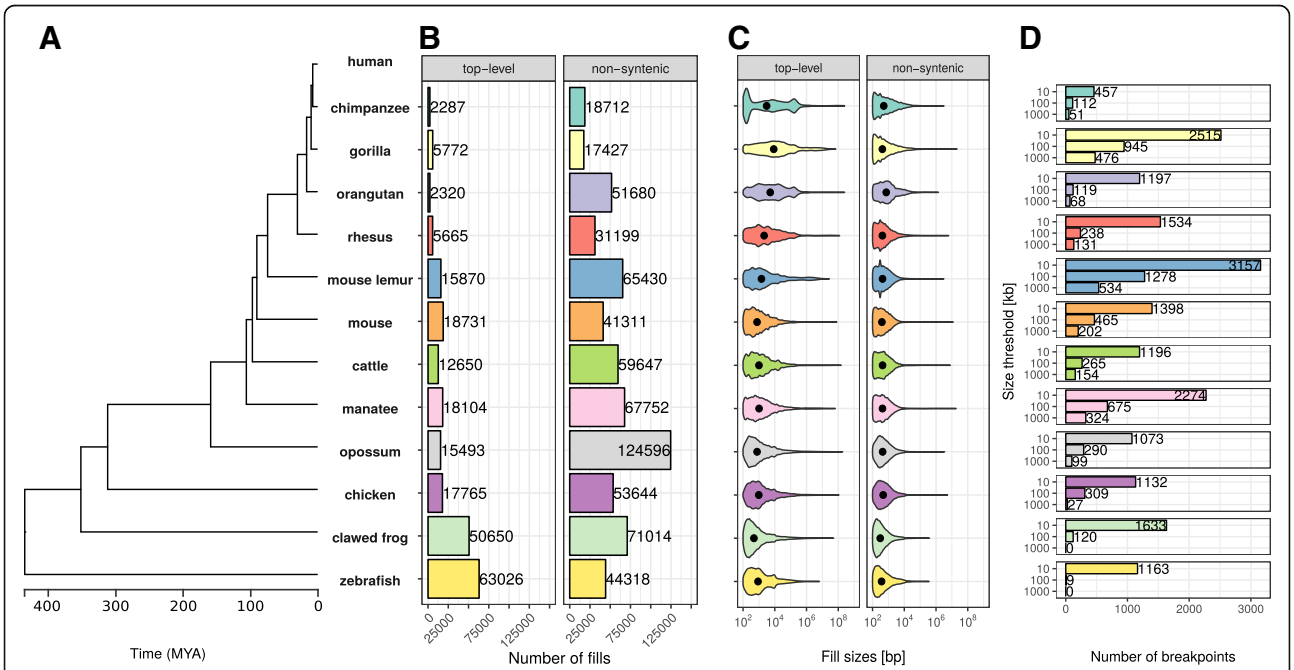

Fig. 1 Number and size distributions of fill sizes of whole-genome alignments between human and 12 other species. **a** Phylogenetic tree with estimated divergence times according to http://timetree.org/ **b** Number of syntenic alignment blocks (fills) between human (hg38) and 12 other species. Top-level 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). **c** Size distribution of top-level (top panel) and non-syntenic (bottom panel) fills as violin plot. **d** 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)

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. In an additional refinement step, we removed false positive breakpoints that are located between close fills mapping on the same chromosome and same orientation in the query species (see the "Methods" section for details).

First, we analyzed the number and size distributions of top-level and non-syntenic fills between human and other species (Fig. 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 and zebrafish, tend to have more but smaller fills (mean length ≤ 1 kb, Fig. 1b, c). However, we also observe many small non-syntenic fills in closely related species, likely arising from transposon insertions [21]. 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. 1d). 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.

Comparing identified breakpoints with syntenic gene pairs

A classical analysis to detect evolutionary rearrangement is to compare adjacent gene pairs with their ortholog genes in another species. If the orthologs are also adjacent and with the same orientation to each other, the human genes are considered syntenic and rearranged if not. Such synteny-based approaches use protein sequences to calculate homology and are therefore likely more accurate in terms of homology. However, the restriction to coding sequences makes them unable to identify the exact breakpoint location in intergenic DNA between non-syntenic genes.

We reasoned that a subset of here identified breakpoints that are located between adjacent genes with unique one-to-one orthologs in a target species can be validated by testing the gene pairs for synteny. To this end, we retrieved for all human genes one-to-one orthologs in 11 species and considered human gene pairs syntenic, if their orthologs are in the other genome on the same chromosome, within close distance, and with the same orientation to each other as the human genes. We calculated a positive predicted value (PPV) of breakpoint identification as the fraction of the non-syntenic gene pairs with breakpoints from all gene pairs (syntenic and non-syntenic) with breakpoint (Additional file 1: Figure S1). The PPV varies depending on species and size thresholds used and has a median of 0.959. Together with a median false positive rate (FPR) of only 0.0169%, this indicates that our approach to identify evolutionary rearrangement breakpoints from whole-genome alignment data is reliable and has high accuracy when compared to gene synteny.

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 3062 TADs identified in human embryonic stem cells (hESC) [3] and 9274 contact domains from high-resolution in situ Hi-C in human B-lymphoblastoid cells (GM12878) [11]. 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. 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 then 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. 2c). For virtually all species and size thresholds analyzed, we found breakpoints significantly enriched at boundaries of TADs and contact domains (Fig. 2c, Additional file 2: Figure S2). Depletion was only observed for some combinations of species and size thresholds which have only very few breakpoints (see Fig. 1c). Furthermore, we compared the distance of each breakpoint to the closest TAD boundary and observed nearly always significantly shorter distances for actual

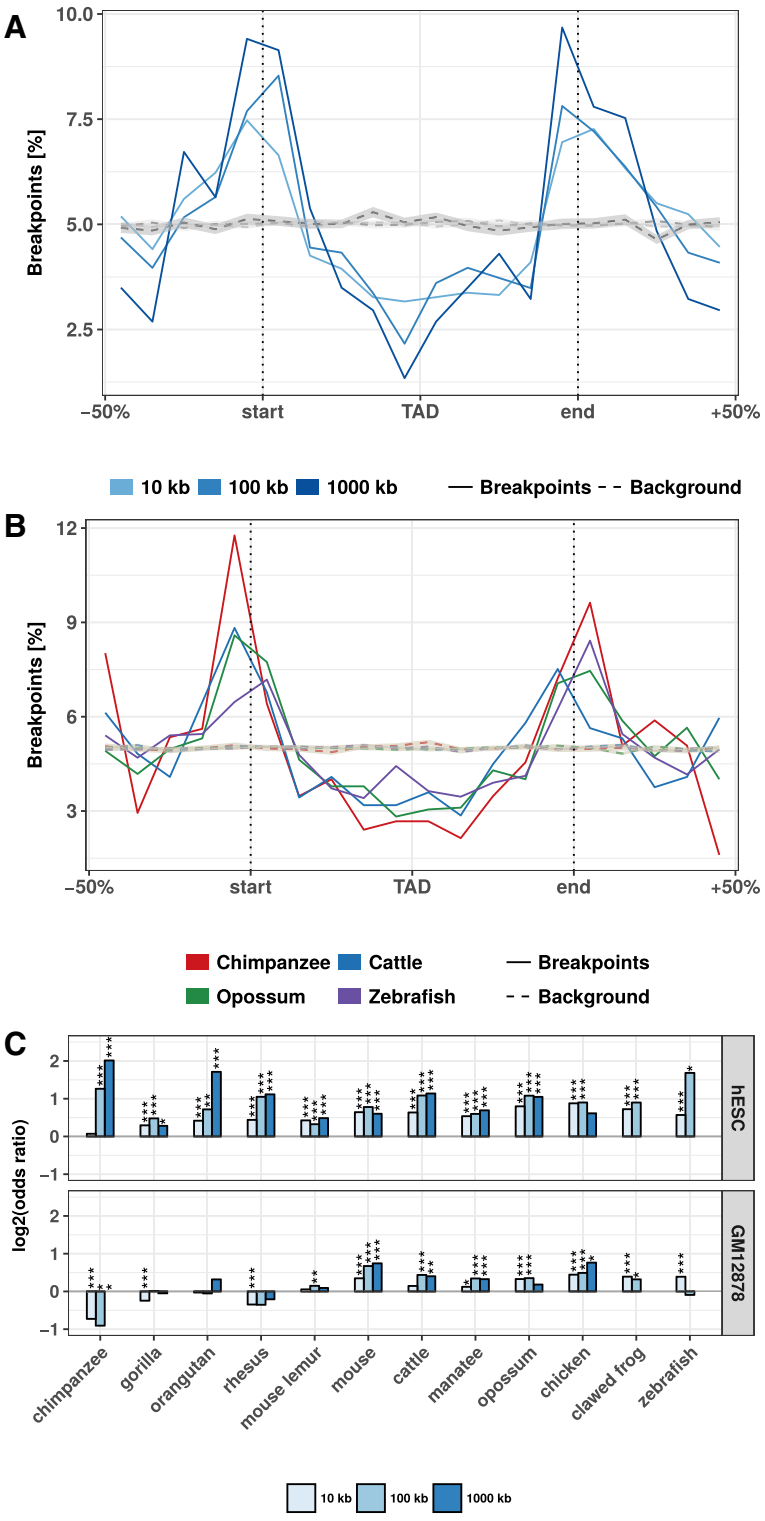

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 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 [3] (top) and contact domains in human GM12878 cells from [11] (bottom), respectively. Asterisks indicate significance of the enrichment using Fisher's exact test (*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$)

breakpoints compared to random controls (Additional file 3: Figure S3). Overall, the enrichment was stronger for TADs in hESC compared to the 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 [11]. 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 [22], 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) [23]. 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 [23]. In the human genome, CNEs cluster around developmental genes in so-called genomic regulatory blocks (GRBs) [24]. It has been shown recently that many GRBs coincide with TADs in human and *Drosophila* genomes [25]. 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. 3a). This is consistent with previous studies in *Drosophila* and fish where CNE arrays often correspond to syntenic blocks [26, 27]. 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 [25]. 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. 3b). However, also the non-GRB-TADs that have less than 20% overlap with GRBs are enriched for rearrangements at TAD boundaries. 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 strong conservation of TAD regions is not fully resolved. A mechanistic explanation could be that certain chromatin features at TAD boundaries promote or prevent DNA double-strand breaks (DSBs) [22, 28]. Alternatively, selective pressure might act against the disruption of TADs due to their functional importance, for example in developmental gene regulation [22]. TADs constitute a structural framework determining possible interactions between promoters and cis-regulatory sequences while prohibiting the influence of other sequences [6, 9]. 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 [9, 29]. 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 (Additional file 4: Table S1) [30]. 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.

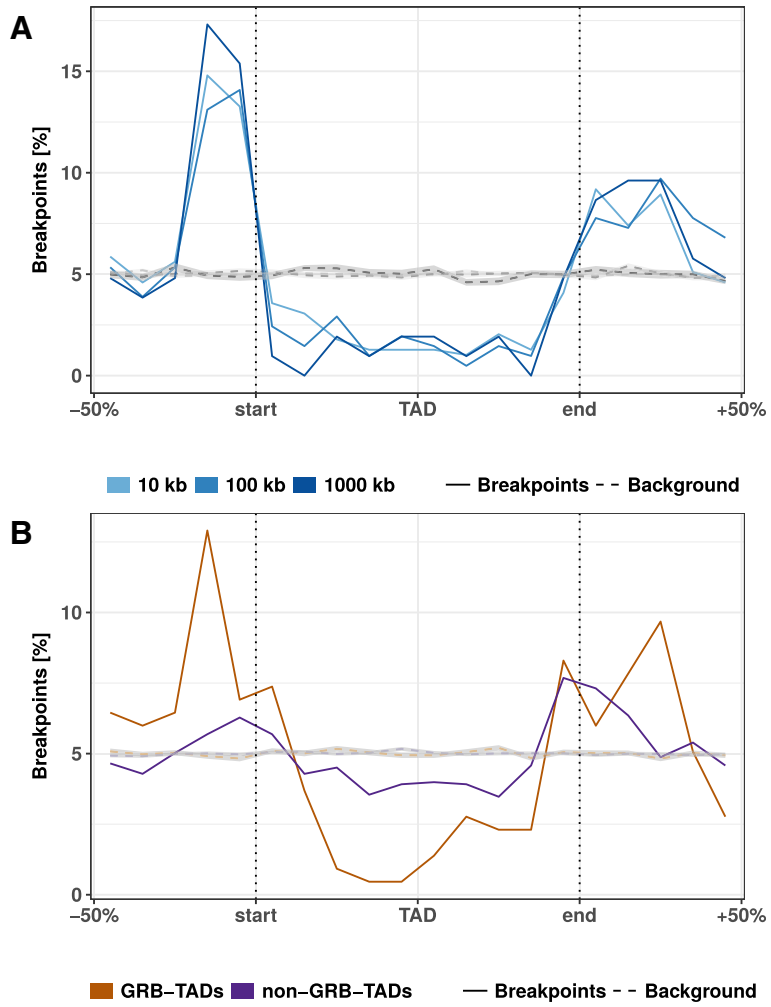

Fig. 3 Rearrangement breakpoint distribution around GRBs and GRB-TADs. **a** Rearrangement breakpoints between mouse and human around 816 GRBs. **b** Breakpoint distribution around GRB-TADs and non-GRB-TADs. GRB-TADs are defined as TADs overlapping more than 80% with GRBs and non-GRB-TADs have less than 20% overlap with GRBs. Breakpoints using a 10 kb fill size threshold are shown

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 (Additional file 5: Table S2), 1525 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 slightly higher expression correlation with their mouse ortholog (median $R = 0.313$) compared to genes outside TADs (mean $R = 0.282$, $p = 0.00023$, Fig. 4a). This indicates higher conservation of gene regulation in TADs and is consistent with the observation of housekeeping genes at TAD boundaries [3] and the role of TADs in providing conserved regulatory environments for gene regulation [25, 31].

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 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 2667 conserved and 94 rearranged TADs. The low number of rearranged TADs is consistent with the depletion of rearrangement breakpoints within TADs in general (Fig. 2). In total, 9500 genes in conserved and 451 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 higher for genes in conserved TADs (median $R = 0.316$) compared to genes in rearranged TADs (median $R = 0.244$) (Fig. 4b). Although the effect size is not very strong, the

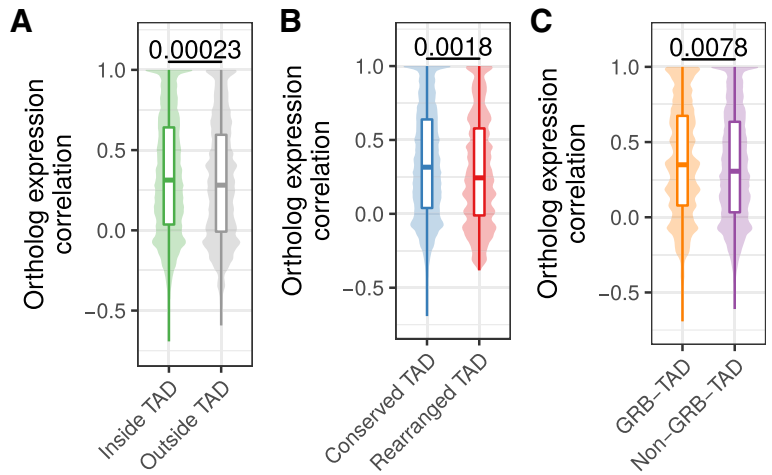

Fig. 4 Ortholog gene expression correlation across tissues in conserved and rearranged TADs. **a** Expression correlation of orthologs across 19 matching tissues in human and mouse for human genes within or outside of hESC TADs. **b** Expression correlation of orthologs across 19 matching tissues in 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

difference is statistically significant ($p = 0.0018$). This shows that disruptions of TADs by evolutionary rearrangements are associated with less conserved gene expression profiles across tissues. We also observed a slightly higher expression correlation for 1003 genes in GRB-TADs compared to 8018 genes in non-GRB TADs (Fig. 4c, $p = 0.0078$).

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 [4, 6, 32]. This might facilitate the cooperation of several sequences that is often needed for the complex spatiotemporal regulation of transcription [33]. The disruption of these enclosed regulatory environments enables the recruitment of other cis-regulatory sequences and might prevent formerly established interactions [22, 34]. The detrimental effects of such events have been shown in the study of diseases [29, 35]. 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 [8, 9, 36]. 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.

Our results are largely consistent with the reported finding that many TADs correspond to clusters of conserved non-coding elements (GRBs) [25]. 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* [26] and fish genomes [27]. However, over 60% of human hESC TADs do not overlap GRBs [25], 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. Alternatively, GRBs detected at lower stringent conservation criteria might be found in some non-GRB TADs. Increased expression correlation of orthologs in conserved TADs suggests that the maintenance of expression regulation is important for many 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 [17]. 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 [17, 37]. However, such an approach assumes equal inter-genic distances, which is violated at TAD boundaries, which have in general higher gene density [3, 38]. 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. For example, the here used gorilla genome gorGor5 was assembled only to contig level and not to whole chromosome level like the other primate genomes and has consequently lower accuracy in breakpoint detection when compared to syntenic genes (Additional file 1: Figure S1).

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 [39] and evolutionary rearrangements [40–42]. Furthermore, induced DSBs and somatic translocation breakpoints are enriched at chromatin loop anchors [28]. 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 is most often disadvantageous for an organism. Structural variations disrupting TADs can lead to miss regulation of neighboring genes as shown for genetic diseases [8, 9, 29, 43] and cancers [44–47].

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 [48]. Since it is also believed that changes in cis-regulatory sequences of developmental genes play a big part in evolutionary innovation [49], 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 [31]. 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. The here reported association of conserved gene expression in intact TADs and divergent expression patterns in rearranged TADs can explain 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 [19]. 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 t of 10 kb, 100 kb, and 1000 kb. The breakpoints were filtered to be located only on chromosomes 1–22, X, and Y. Furthermore, we refined our set of breakpoints to eliminate potential false positives by filtering out breakpoints that are flanked by two different fills of at least threshold size t and that align in the same orientation to the same chromosome in the query species.

Estimating the accuracy of breakpoint detection using gene synteny

We retrieved one-to-one orthologs for all human protein coding genes from ensemble (version aug2017.archive.ensembl.org) for all used species, except manatee for which no ensemble database was available.

For each species, we filtered the human genes to only those with that have a unique one-to-one ortholog in the respective species and built a dataset of all adjacent gene pairs. For each species s and size threshold t we then considered only the gene pairs with intergenic distance $\leq t$. Each of these gene pairs was then labeled syntenic, if their orthologs in s are adjacent with the same orientation to each other and have an intergenic distance $\leq t$ in the genome of s, or non-syntenic, if not. Furthermore, we considered a gene pair rearranged, if we could identify a breakpoint between human and species s with size threshold t in the intergenic region between the gene pairs, or non-rearranged if not.

We considered these gene pairs as true positives (TP), if non-syntenic and rearranged; false positive (FP), if syntenic and rearranged; true negative (TN), if syntenic and non-rearranged; and false negative (FN), if non-syntenic and non-rearranged. The fraction of breakpoints in syntenic gene pairs was considered as false positives. Furthermore, we computed for each species and size threshold the false positive rate (FPR) as $FPR = FP/(FP + TN)$ and the positive predictive value (PPV) as $PPV = TP/(TP + FP)$.

Topologically associating domains and contact domains

We obtained topologically associating domain (TAD) calls from published Hi-C experiments in human embryonic stem cells (hESC) [3] and contact domains from published in situ Hi-C experiments in human GM12878 cells [11] Genomic coordinates of TADs and contact

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

domains were converted from hg18 and hg19 to hg38 genome assembly using the UCSC liftOver tool [50].

Genomic regulatory blocks (GRBs)

GRBs are clusters of strongly conserved non-coding elements. We downloaded recently published GRB coordinates, which were defined as clusters of non-protein-coding sequences of at least 50 bp with over 70% sequence identity between human (hg19) and chicken (galGal4) genomes [25]. Genomic coordinates of GRBs were converted from hg19 genome assembly to hg38 using the UCSC liftOver tool.

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 [30] were retrieved from the EBI Expression Atlas [51] 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 [52]. 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 aug2017.archive.ensembl.org) and could assign 13,065 to ortholog genes in mouse by allowing only the one-to-one orthology type [53]. 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 [25]. 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/JKrefting/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 [54] packages rtracklayer [55] and GenomicRanges [56]. Gene coordinates and ortholog assignments were retrieved from Ensemble data base (version aug2017.archive.ensembl.org) using the package biomaRt [51]. For data integration and visualization, we used R packages from the tidyverse [52].

Additional files

Additional file 1: Figure S1. Breakpoint identification accuracy as compared to gene synteny. Considered are adjacent pairs of human genes with one-to-one orthologs and intergenic distance below a size threshold. (A) Positive predicted value as the fraction of non-syntenic gene pairs with breakpoint from all considered gene pairs (syntenic and non-syntenic) with breakpoint. (B) False positive rate as the percent of syntenic gene pairs with breakpoint from the sum of syntenic pairs with breakpoint and non-syntenic gene pairs without breakpoint. (PDF 21 kb)

Additional file 2: Figure S2. 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. (PDF 42 kb)

Additional file 3: Figure S3. 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. (PDF 14 kb)

Additional file 4: Table S1. Matching tissues and samples with CAGE expression data in human and mouse. (TSV 2 kb)

Additional file 5: Table S2. Ortholog genes in human and mouse with gene expression correlation across tissues. (TSV 1036 kb)

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

Availability of data and materials
The source code of all analysis is available on GitHub: https://github.com/JKrefting/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 Additional file 4: Table S1.

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. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 11 December 2017 Accepted: 26 July 2018
Published online: 07 August 2018

References
1. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:661–78.
2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
3. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
4. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485:381–5.
5. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148:458–72.
6. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 2014;24:390–400.
7. Zhan Y, Mariani L, Barozzi I, Schulz EG, Bluthgen N, Stadler M, et al. 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 Res. 2017; https://doi.org/10.1101/gr.212803.116.
8. Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang N, Hurles ME, et al. Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. 2014;15:423.
9. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
10. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: how alterations of chromatin domains result in disease. Trends Genet. 2016;xx:1–13.
11. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
12. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.
13. Gómez-Marín C, Tena JJ, Acemel RD, López-Mayorga M, Naranjo S, de la Calle-Mustienes E, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci. 2015;112:201505463.
14. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015; https://doi.org/10.1038/nature14450.
15. Hsieh T-HS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell. 2015;162(4):1–12.
16. Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD, et al. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature. 2014; https://doi.org/10.1038/nature13833.
17. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309.
18. Nora EP, Dekker J, Heard E. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? BioEssays. 2013;35:818–28.
19. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003;100:11484–9.
20. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
21. Mills RE, Bennett EA, Iskow RC, Luttig CT, Tsui C, Pittard WS, et al. Recently mobilized transposons in the human and chimpanzee genomes. Am J Hum Genet. 2006;78:671–9.

22. Farré M, Robinson TJ, Ruiz-Herrera A. An Integrative Breakage Model of genome architecture, reshuffling and evolution. BioEssays. 2015:n/a.
23. Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. Conserved non-coding elements: developmental gene regulation meets genome organization. Nucleic Acids Res. 2017;45(22):12611-12624.
24. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 2007;17:545–55.
25. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun. 2017; 8:441.
26. Engström PG, Sui SJH, Drivenes Ø, Becker TS, Lenhard B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 2007;17:1898–908.
27. Dimitrieva S, Bucher P. Genomic context analysis reveals dense interaction network between vertebrate ultraconserved non-coding elements. Bioinformatics. 2012;28:i395–401.
28. Canela A, Maman Y, Jung S, Wong N, Callen E, Day A, et al. Genome organization drives chromosome fragility. Cell. 2017;170(3):1–15.
29. Redin C, Brand H, Collins RL, Kammin T, Mitchell E, Hodge JC, et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat Genet. 2016; https://doi.org/10.1038/ng.3720.
30. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
31. Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. 2017;45: 81–91.
32. Schoenfelder S, Furlan-magaril M, Mifsud B, Tavares-cadete F, Sugar R, Javierre B, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25:582-597.
33. Andrey G, Mundlos S. The three-dimensional genome: regulating gene expression during pluripotency and development. 2017;144:3646–3658. doi: https://doi.org/10.1242/dev.148304.
34. Montavon T, Thevenet L, Duboule D. Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. Proc Natl Acad Sci U S A. 2012;109:20204–11.
35. Zepeda-Mendoza CJ, Ibn-Salem J, Kammin T, Harris DJ, Rita D, Gripp KW, et al. Computational prediction of position effects of apparently balanced human chromosomal rearrangements. Am J Hum Genet. 2017;101:206–17.
36. Spielmann M, Brancati F, Krawitz PM, Robinson PN, Ibrahim DM, Franke M, et al. Homeotic arm-to-leg transformation associated with genomic rearrangements at the PITX1 locus. Am J Hum Genet. 2012;91:629–35.
37. Pevzner P, Tesler G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A. 2003;100:7672–7.
38. Hou C, Li L, Qin ZS, Corces VG. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol Cell. 2012;48:471–84.
39. Roukos V, Misteli T. The biogenesis of chromosome translocations. Nat Cell Biol. 2014;16:293–300.
40. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 2005;309:613–7.
41. Hinsch H, Hannenhalli S. Recurring genomic breaks in independent lineages support genomic fragility. BMC Evol Biol. 2006;6:90.
42. Gordon L, Yang S, Tran-Gyamfi M, Baggott D, Christensen M, Hamilton A, et al. Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. Genome Res. 2007; 17:1603–13.
43. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538:265–269.
44. Hnisz D, Weintraub AS, Day DS, Valton A, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.
45. Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature. 2014;511:428-434.
46. Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stütz AM, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet. 2016;49:65-74.
47. Akdemir KC, Li Y, Verhaak RG, Beroukhim R, Cambell P, Chin L, et al. Spatial Genome Organization as a framework for somatic alterations in human cancer. bioRxiv. 2017;
48. Acemel RD, Maeso I, Gómez-Skarmeta JL. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. Wiley Interdiscip Rev Dev Biol. 2017;6:e265.
49. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.
50. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC genome browser database: update 2006. Nucleic Acids Res. 2006; 34(Database issue):D590–8.
51. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.
52. Wickham H, Grolemund G. R for data science: import, tidy, transform, visualize, and model data. 1st ed. Sebastopol: O'Reilly Media; 2017.
53. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, et al. Ensembl comparative genomics resources. Database. 2016;2016 https://doi.org/10.1093/database/bav096.
54. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015;12:115–21.
55. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.
56. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118.