Published online 14 September 2016

Nucleic Acids Research, 2017, Vol. 45, No. 1 ***81–91***
doi: 10.1093/nar/gkw813

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

Jonas Ibn-Salem[1,2], Enrique M. Muro[1,2] and Miguel A. Andrade-Navarro[1,2,*]

[1]Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany and [2]Institute of Molecular Biology, 55128 Mainz, Germany

Received November 27, 2015; Revised August 31, 2016; Accepted September 3, 2016

ABSTRACT

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

INTRODUCTION

Paralog genes arise from gene duplication events during evolution. The resulting sequence similarity between paralog pairs might lead to similar structure and function of encoded proteins (1). Since paralogs often form part of the same protein complexes and pathways, it is advantageous for the cell to coordinate their expression (2).

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 (3–7). These chromatin looping events can be measured by chromatin conformation capture (3C) experiments (8), which use proximity-ligation, and more recently high-throughput sequencing (Hi-C) to measure chromatin-chromatin contact frequencies genome-wide (9).

These interaction maps revealed tissue-invariant chromatin regions, named topologically associating domains (TADs), which have more interactions within themselves than with other regions (10–12). TADs seem to be stable across cell types and conserved between mammals (10,13,14). Regions within TADs show concerted histone chromatin signatures (10,12), gene expression (11,15) and DNA replication timing (16). Furthermore, disruption of TAD boundaries is associated to genetic diseases (17,18).

We wondered if the Hi-C data could reveal evolutionary pressure driving paralogous expansion to favor 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 (10,13), mouse and dog (14), 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 ac-

*To whom correspondence should be addressed. Tel: +49 6131 39 21582; Fax: +49 6131 39 21589; Email: andrade@uni-mainz.de

cording 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 (19,20) 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 (21). This resulted in a total of 19 430 human genes; more than half of those had at least one human paralog gene (Supplementary Figure S1A).

However, many human genes have more than one paralog (Supplementary Figure S1B). To avoid over-representation 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` (22). The number of synonymous mutations between paralogs has been used to approximate the duplication age (23). Therefore our implementation favors 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 1). We observed that modifications of this strategy to select unique paralog genes did not affect essentially the results of our study (e.g. by selecting pairs while maximizing dS; Supplementary Figure S2).

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 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.

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 (24). In this study, the activity of enhancers and genes was correlated within 500 kb 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.

Topologically associating domains

We obtained topologically associating domain (TAD) calls from two recently published Hi-C studies in human cells (10,13). TAD locations mapped to the hg18 genome assembly were converted to hg19 using the UCSC liftOver tool (25). A/B-compartment and sub-compartment annotations were obtained from high-resolution Hi-C experiments in human GM12878 cells (13).

Hi-C interaction maps

Individual chromatin–chromatin contact frequencies from IMR90 cells at 5 kb resolution were retrieved from (13)(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 (26). We only considered pairwise gene interactions if the TSSs of the two genes were located in different bins of the Hi-C matrix with normalized contacts $\geq$0. Capture Hi-C data between promoter regions in human GM12878 cells were downloaded from ArrayExpress (accession: E-MTAB-2323) (27).

Randomization

We analyzed 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-localization 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:

$$w(g_i, g_j) = \frac{f_{\text{paralogs}}(d_{i,j})}{f_{\text{all}}(d_{i,j})} \quad (1)$$

where $f_{\text{paralogs}}(d_{i,j})$ is the observed frequency 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

Table 1. Filtering of human paralog gene pairs

Paralog pairs	Human	Mouse	Dog
All paralog pairs	46 546	110 490	28 293
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

for sampling:

$$p_{\text{dist}}(g_i, g_j) = \frac{w(g_i, g_j)}{\sum_{i,j} w(g_i, g_j)} \quad (2)$$

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 (Supplementary Figure S1D). Assuming that the number of enhancers associated to genes is independent from the distance, we computed sampling probabilities by,

$$p_{\text{dist+eh}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j) \quad (3)$$

where 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:

$$p_{\text{eh}}(n) = \frac{w_{\text{eh}}(n)}{\sum_{i=0}^{N} w_{\text{eh}}(i)} \quad (4)$$

and

$$w_{\text{eh}}(n) = \frac{f_{\text{paralogs}}(n)}{f_{\text{all}}(n)} \quad (5)$$

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.

$$p_{\text{dist+eh+strand}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j) \cdot p_{\text{strand}}(s_{i,j}) \quad (6)$$

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 4.

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

$$p_{\text{dist+eh+len}}(g_i, g_j) = p_{\text{dist}}(g_i, g_j) \cdot p_{\text{eh}}(n_i) \cdot p_{\text{eh}}(n_j) \cdot p_{\text{len}}(l_i) \cdot p_{\text{len}}(l_j) \quad (7)$$

whereby $p_{\text{len}}(l)$ for gene length l is computed in the same way as for distances between gene pairs (equation 2) 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 (Supplementary Figure S3), associated enhancers of each gene (Supplementary Figure S4), same strand (Supplementary Figure S5) and gene lengths (Supplementary Figure S6).

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 hypothesized 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 (Supplementary Figure S1A). However, many human genes have several paralogs (Supplementary Figure S1B). From all 46 546 paralog gene pairs we filtered for only one pair per gene ($n = 6256$) and further for non-overlapping pairs on the same chromosome ($n = 1560$) (see 'Materials and Methods' section). We will refer to close paralogs if their TSSs are within 1 Mb of each other ($n = 1114$) and to distal pairs for paralogs with TSSs separated by more than 1 Mb ($n = 446$) (Table 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, Supplementary Figure S1C), which fits the observation from (28) in yeast. Furthermore, paralogs tend to be associated to more enhancers compared to non-paralog genes (on average 3.8 versus 2.5 enhancers per gene, $P = 2.89 \times 10^{-70}$, Supplementary Figure S1D) and the distance to the nearest associated enhancer is significantly shorter ($P = 2.71 \times 10^{-22}$, Supplementary Figure S1E).

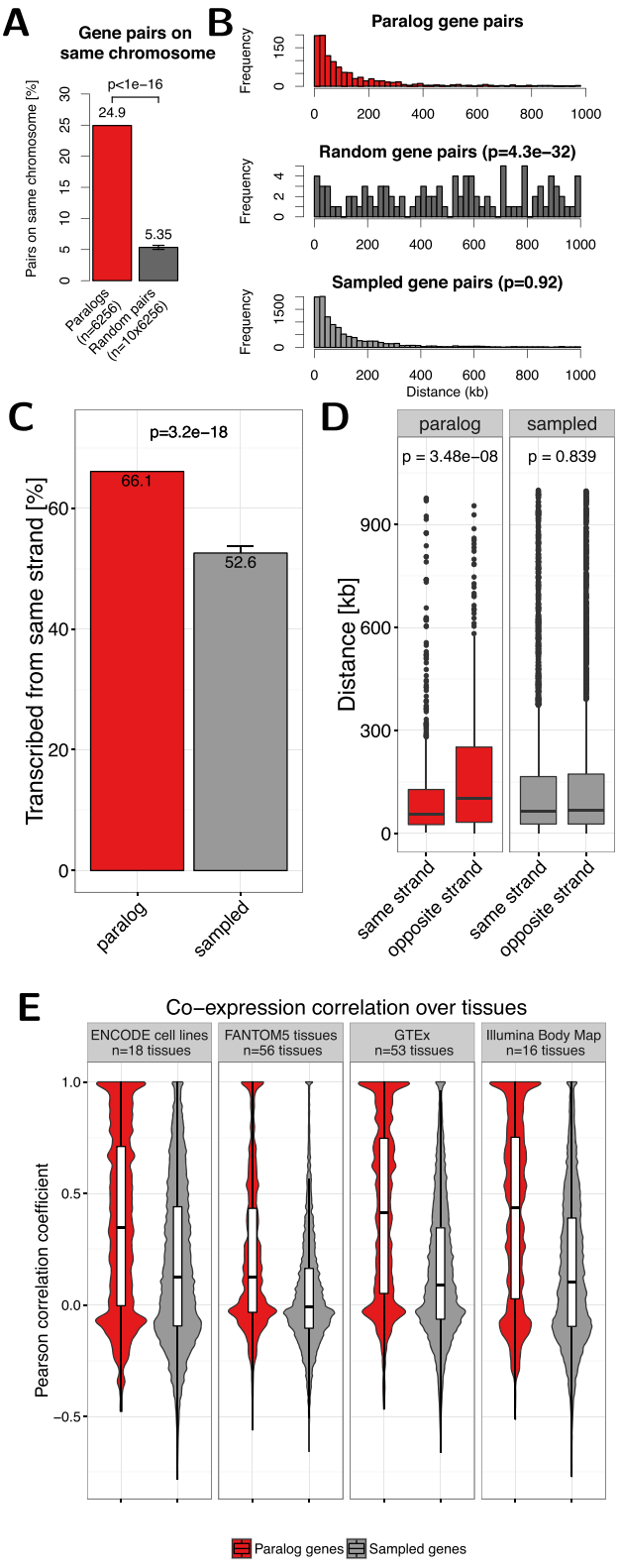

Figure 1. (**A**) Percent of paralog (red) and random (dark gray) gene pairs that are located on the same chromosome. The error bar indicates the standard deviation observed in 10 times replicated random sampling of gene pairs. (**B**) Genomic distance distribution of paralog gene pairs (top), random 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 (gray) 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 datasets between paralog gene pairs (red) and sampled control gene pairs (gray). White boxes show 25th, 50th and 75th percent quantile of the data and the filled areas indicate the density distribution.

Since most genome duplication events in humans emerge through tandem duplications (29), 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 ($P < 10^{-16}$, Figure 1A).

We further analyzed 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 ($P = 4.3 \times 10^{-32}$, Figure 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%, $P = 3.2 \times 10^{-18}$, Figure 1C).

Furthermore, we observed that paralogs in the same strand are closer to each other on the chromosome than pairs in opposite strands ($P = 3.48 \times 10^{-8}$, Figure 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 analyzed by the ENCODE consortium by RNA-seq (30). 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 (Figure 1E).

We repeated the analysis with three other independent gene expression datasets from FANTOM5 ($n = 56$ tissues) (31), GTEx ($n = 53$ tissues) (32) and the Illumina Body Map ($n = 16$ tissues), which we retrieved from the EBI Expres-

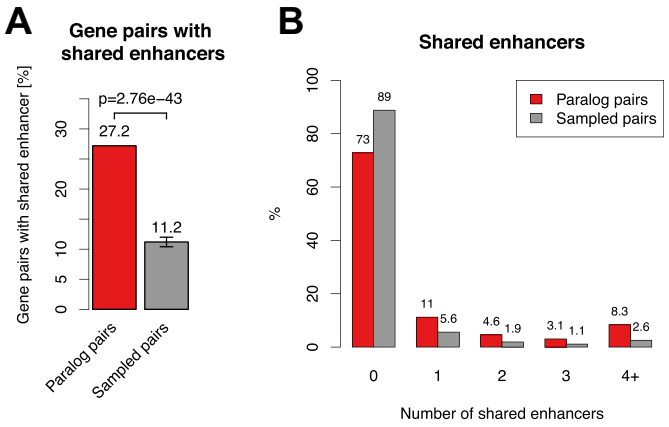

Figure 2. Shared enhancers among paralog gene pairs. (**A**) Percent of close paralog (red) and sampled control (gray) gene pairs with at least one shared enhancer. (**B**) Percent of gene pairs versus number of shared enhancers for paralog and sampled control gene pairs.

sion Atlas (33). In all datasets we found more positively correlated paralog pairs compared to the sampled gene pairs (Figure 1E). This shows that many paralogs are expressed with high coordination in a tissue specific manner.

Paralog genes share enhancers

We hypothesized 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 (Figure 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 (Supplementary Figure S4, 'Materials and Methods' section).

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 ($P = 2.76 \times 10^{-43}$, Figure 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; Supplementary Figure S7). Next, we compared the percent of gene pairs with shared enhancers as a function of the number of shared enhancers between paralogs and sampled gene pairs. We observed that paralog pairs are enriched for higher number of shared enhancers compared to the sampled gene pairs (Figure 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 topologically 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 (10,13). Genes within the same TAD are therefore likely to have common gene regulatory programs (11,15).

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 (10,13). 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 (Supplementary Figure S8). TADs from (13) 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 (10) are of 1 Mb on average and are defined as non-overlapping genomic intervals.

We hypothesized 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 (Figure 3A). In seven out of nine datasets 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 analyzed in (13). 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 Figure 3A).

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 (9,13,34). 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 ($P = 0.00965$, Figure 3B). Next we tested the same for recently distinguished sub-compartment types from high-resolution Hi-C interaction maps (13). Again, paralogs are enriched to be located within the same sub-compartment type (38.8 versus 32.4%, $P = 0.00547$, Figure 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 (27,35)) we hypothesized 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 (9) we compared the number of normalized Hi-C contacts between TSS of distal paralog genes (that have pro-

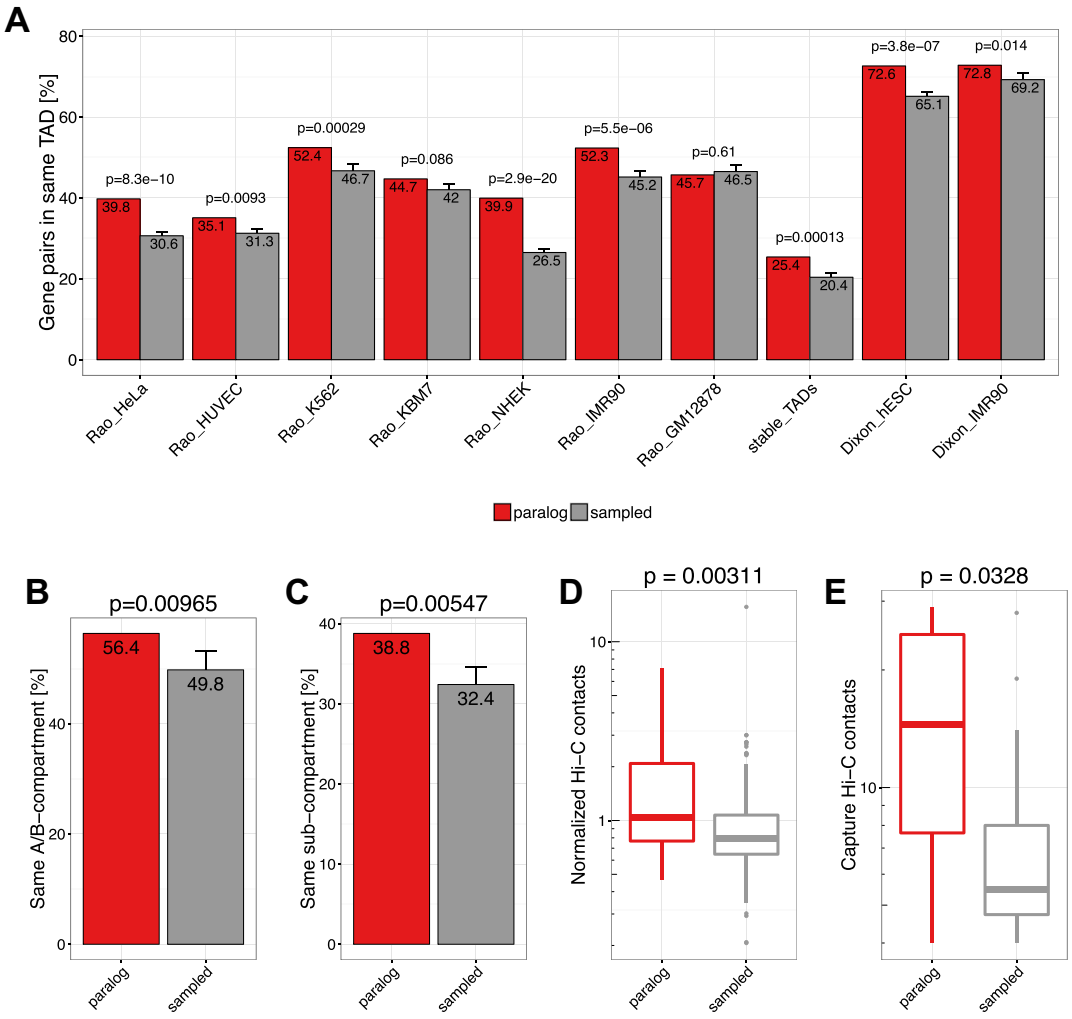

Figure 3. (**A**) Co-localization of close paralog genes within the same TAD compared against sampled gene pairs for TAD datasets from different cell types and studies. The first seven bars show values for TADs called in HeLa, HUVEC, K562, KBM7, NHEK, IMR90 and GM12878 cells by (13). The eighth bar shows 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 (10). 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 (13). 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 (13). Pairs located in the same subcompartment interval were excluded. (**D**) Normalized Hi-C contact frequencies between TSSs of distal paralog gene pairs (n=30, median=1.04, average=1.86) and sampled background gene pairs (n=300, median=0.788, average=0.968). (**E**) Promoter capture-C contact frequencies between distal paralog gene pairs (n=6, median=15.5, average=16.2) and sampled background gene pairs (n=43, median=5, average=6.95).

moters 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 5 kb bin-size resolution (13) and observed significantly more normalized chromatin interactions between paralog gene pairs compared to sampled control gene pairs ($P = 0.00311$, Figure 3D). We furthermore used an independent dataset of high resolution promoter–promoter interactions measured by capture Hi-C (27). Again, we observed a strong enrichment of promoter–promoter interactions between distal paralogs compared to control genes pairs ($P = 0.0328$, Figure 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 (Figure 4A). To analyze this in more detail we focused on only those pairs on the same chromosome that have a TSS distance of at least 10 kb but less than 1 Mb. This is the distance range of most paralog pairs and allows to separate genes in Hi-C interaction maps and TADs (Supplementary Figure S9A). Consequently, we observe paralogs more often in the same TAD in eight out of nine datasets for this distance range (Supplementary Figure S9B). For these pairs we observe significant lower Hi-C contact frequencies if pairs are within the same IMR90 TAD (13) as compared to sampled genes ($P = 0.00094$) but not if pairs are in different TADs

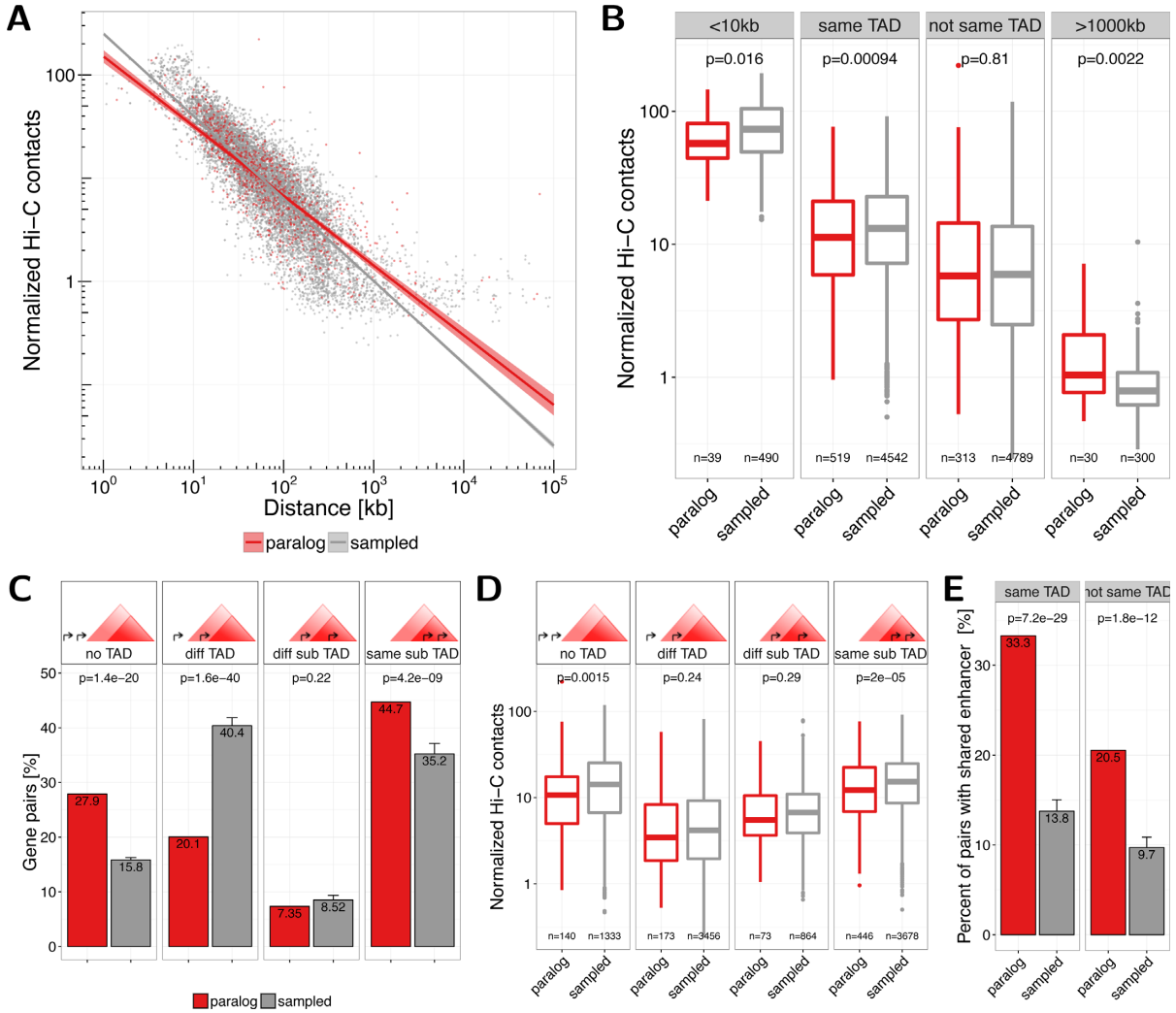

Figure 4. (**A**) Normalized Hi-C contacts by genomic distance between paralog (red) and sampled (gray) gene pairs. Lines show linear regression fit separately for paralogs (red) and sampled (gray) pairs with 95% confidence intervals in shaded areas. (**B**) Normalized Hi-C contacts between pairs of paralogs (red) and sampled gene pairs (gray) for the groups: <10 kb genomic distance, located in the same TAD, not in the same TAD and with genomic distance >1000 kb. (**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 (gray) pairs. TADs from IMR90 cells from (13) were used, which nested in contrast to TAD calls from (10). (**D**) Normalized Hi-C contacts between pairs of paralogs (red) and sampled gene pairs (gray) 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 (gray) separated for pairs in the same IMR90 TAD (left) or not (right).

($P = 0.81$, Figure 4B). We got comparable results when analyzing the Capture Hi-C data the same way (Supplementary Figure S9C). Next, we tested whether this can be explained by the nested sub-TAD structure of TADs called from high-resolution Hi-C in IMR90 (13). 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 ($P = 4.2 \times 10^{-9}$, Figure 4C). However, although not always significant, paralogs have fewer Hi-C contacts than sampled gene pairs in all of these groups (Figure 4D). In addition, close paralogs within the same TAD share more enhancers than close paralogs not being in the same TAD (Figure 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 (Supplementary Figure S10).

In summary, we observed that while close paralogs (situated at less than 1 Mb) 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.

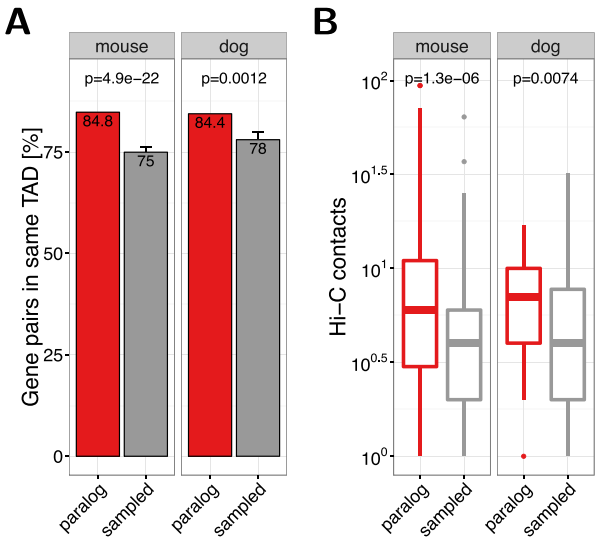

Figure 5. (**A**) Co-occurrence of close paralog genes with the same TAD in mouse (left panel) and dog (right panel). (**B**) Hi-C contacts between promoter of distal gene pairs in Hi-C experiments in liver cells from mouse (left panel; n=66 and n=1005 for paralog and sampled gene pairs, respectively) and dog (right panel; n=21 and n=187 for paralog and sampled gene pairs, respectively). Hi-C data and TAD calls were taken from (14).

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. familiaris*) genomes. Similar as for human data, we found that more than two third of the genes had at least one paralog copy (Supplementary Figure S11A and D), paralog pairs clustered on the same chromosome (Supplementary Figure S11B and E), and had close linear distances (Supplementary Figure S11C and F).

We sampled control gene pairs with the same distance distribution as paralogs for both species separately (Supplementary Figure S11C and F). We used TADs from recently published Hi-C data in liver cells of mouse and dog (14), which have a size distribution comparable to TADs from human cells (Supplementary Figure S8). 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 ($P = 4.9 \times 10^{-22}$) and dog ($P = 0.0012$) than expected by chance (Figure 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 ($P = 1.3 \times 10^{-6}$) and dog ($P = 0.0074$) (Figure 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.

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) (Supplementary Figure S12).

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 (Supplementary Figure S13A and D). We hypothesized 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 (Supplementary Figure S13B and E) and are biased for close linear distances (Supplementary Figure S13C and 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 versus 76%, $P = 0.11$; Figure 6A). Significant enrichment was observed with orthologs in the dog genome (85 versus 77%, $P = 0.0016$; Figure 6A).

For distal human paralogs we quantified the promoter contacts of their orthologs in mouse and dog and found enriched Hi-C contacts in mouse ($P = 0.013$) and dog ($P = 4.3 \times 10^{-5}$; Figure 6B).

These results show that both the co-localization of paralogs in TADs and the contacts between distal paralogs are only weakly 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 (Figure 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.

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* (36) or in mammals (37)) and it was previously suspected that the structure of chromatin would have to do with this (38), particularly cis-acting units (37). The discovery and characterization of 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 (29) 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

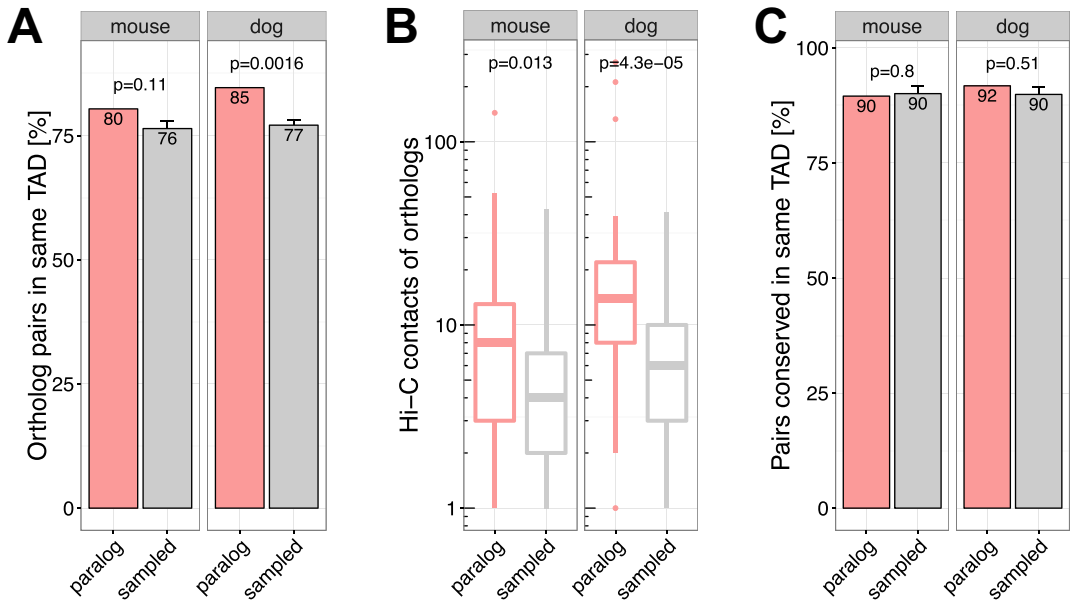

Figure 6. One-to-one orthologs of human paralog genes in mouse and dog genome. (**A**) Percent of mouse (left) and dog (right) orthologs of human paralog pairs that are in the same TAD in the mouse and dog genome, respectively. (**B**) Normalized Hi-C contacts between promoters of one-to-one orthologs of human distal paralogs in the mouse (left; n=21, median=8, average 16.0 for paralogs; n=379, median=4, average=5.39 for sampled) and dog (right; n=21, median=14, average=5.39 for paralogs; n=384, median=6, average=7.26 for sampled) 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 (13) were used.

next enhancer. These differences, which could be partially explained by the observation that paralogs are more often tissue specific (Supplementary Figure S1F), complicated the methodology for choosing meaningful control pairs (see section 'Materials and Methods' section).

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 (39).

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 (Figure 4B), while sharing more enhancers (Figure 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 neighboring positions within the same TAD in human chromosome 17 and encode replaceable subunits of the polycomb repressive complex 1 involved in epigenetic regulation of cell specification (40). 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 (2,41,42) but it was observed that while some paralogs are co-expressed some others have negative correlation across tissues (2). 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 (43). We observed the well-known fact that paralog pairs generally encode for proteins that interact more often than non-paralog proteins (Supplementary Figure S14). But, most importantly, we observed that the chances of close pairs of genes to encode for interacting proteins raise 2.3-fold if they are in the same TAD, while, in contrast, if these genes are paralogs the difference is much smaller (1.2-fold, Supplementary Figure S14). 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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

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

FUNDING

Funding for open access charge: Johannes-Gutenberg University of Mainz.
Conflict of interest statement. None declared.

REFERENCES

1. Koonin,E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
2. Makova,K.D. and Li,W.-H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.*, **13**, 1638–1645.
3. Ptashne,M. (1986) Gene regulation by proteins acting nearby and at a distance. *Nature*, **322**, 697–701.
4. Deng,W., Lee,J., Wang,H., Miller,J., Reik,A., Gregory,P.D., Dean,A. and Blobel,G.A. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, **149**, 1233–1244.
5. Carter,D., Chakalova,L., Osborne,C.S., Dai,Y.F. and Fraser,P. (2002) Long-range chromatin regulatory interactions in vivo. *Nat. Genet.*, **32**, 623–626.
6. Tolhuis,B., Palstra,R.J., Splinter,E., Grosveld,F. and de Laat,W. (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell*, **10**, 1453–1465.
7. Spitz,F. and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
8. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
9. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
10. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
11. Nora,E.P., Lajoie,B.R., Schulz,E.G., Giorgetti,L., Okamoto,I., Servant,N., Piolot,T., van Berkum,N.L., Meisig,J., Sedat,J. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
12. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458–472.
13. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
14. Vietri Rudan,M., Barrington,C., Henderson,S., Ernst,C., Odom,D.T., Tanay,A. and Hadjur,S. (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.*, **10**, 1297–1309.
15. Le Dily,F., Bau,D., Pohl,a., Vicent,G.P., Serra,F., Soronellas,D., Castellano,G., Wright,R. H.G., Ballare,C., Filion,G. *et al.* (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.*, **28**, 2151–2162.
16. Pope,B.D., Ryba,T., Dileep,V., Yue,F., Wu,W., Denas,O., Vera,D.L., Wang,Y., Hansen,R.S., Canfield,T.K. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.
17. Ibn-Salem,J., Köhler,S., Love,M.I., Chung,H.-R., Huang,N., Hurles,M.E., Haendel,M., Washington,N.L., Smedley,D., Mungall,C.J. *et al.* (2014) Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome Biol.*, **15**, 423.
18. Lupiáñez,D.G., Kraft,K., Heinrich,V., Krawitz,P., Brancati,F., Klopocki,E., Horn,D., Kayserili,H., Opitz,J.M., Laxova,R. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
19. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
20. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
21. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
22. Galil,Z. (1986) Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, **18**, 23–38.

23. Lan,X. and Pritchard,J.K. (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, **352**, 1009–1013.
24. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
25. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598 .
26. Knight,P.a. and Ruiz,D. (2013) A fast algorithm for matrix balancing. *IMA J. Numerical Anal.*, **33**, 1029–1047.
27. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.
28. He,X. and Zhang,J. (2005) Gene complexity and gene duplicability. *Curr. Biol.*, **15**, 1016–1021.
29. Newman,S., Hermetz,K.E., Weckselblatt,B. and Rudd,M.K. (2015) Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. *Am. J. Hum. Genet.*, **96**, 1–13.
30. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
31. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Lassmann,T., Itoh,M., Summers,K.M., Suzuki,H., Daub,C.O. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
32. GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
33. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Füllgrabe,A., Fuentes,A.M.-P., Jupp,S., Koskinen,S. *et al.* (2015) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
34. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Rev. Genet.*, **14**, 390–403.
35. Cremer,T., Cremer,M., Hübner,B., Strickfaden,H., Smeets,D., Popken,J., Sterr,M., Markaki,Y., Rippe,K. and Cremer,C. (2015) The 4D nucleome: evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS Lett.* **589**, 2931–2943.
36. Boutanaev,A.M., Kalmykova,A.I., Shevelyov,Y.Y., Nurminsky,D.I., Smith,M., Evolution,T., Masterpiece,T., Helm,C. and Selection,N. (2002) Large clusters of co-expressed genes in the Drosophila genome. *Nature*, **420**, 666–669.
37. Purmann,A., Toedling,J., Schueler,M., Carninci,P., Lehrach,H., Hayashizaki,Y., Huber,W. and Sperling,S. (2007) Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. *Genomics*, **89**, 580–587.
38. Sproul,D., Gilbert,N. and Bickmore,W.a. (2005) The role of chromatin structure in regulating the expression of clustered genes. *Nature Rev. Genet.*, **6**, 775–781.
39. Lan,X. and Pritchard,J.K. (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, **352**, 1009–1013.
40. Becker,M., Mah,N., Zdzieblo,D., Li,X., Mer,A., Andrade-navarro,M.A. and Mu,A.M. (2015) Epigenetic mechanisms in cellular reprogramming. In: Meissner,A and Walter,J (eds). *Epigenetics and Human Health, Epigenetics and Human Health*. Springer, Heidelberg, pp. 141–166.
41. Huminiecki,L. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.*, **14**, 1870–1879.
42. Rogozin,I.B., Managadze,D., Shabalina,S.A. and Koonin,E.V. (2014) Gene family level comparative analysis of gene expression n mammals validates the ortholog conjecture. *Genome Biol. Evol.*, **6**, 754–762.
43. Schaefer,M.H., Fontaine,J.-F., Vinayagam,A., Porras,P., Wanker,E.E. and Andrade-Navarro,M.a. (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.