Vertebrate whole genome duplications shaped the current 3D genome architecture

Caelinn James[1,2,5], Marco Trevisan-Herraz[1,3,5,6], David Juan[4], Daniel Rico[1,6]

[1]Biosciences Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom
[2]Current address: Institute of Evolutionary Biology, The University of Edinburgh, Edinburgh, United Kingdom
[3]Current address: Translational and Clinical Research Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom
[4]Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain
[5]Joint first authors
[6]Corresponding authors: Marco Trevisan-Herraz (marco.trevisan@newcastle.ac.uk) and Daniel Rico (daniel.rico@newcastle.ac.uk)

Summary

Topologically associated domains (TADs) are interaction sub-networks of 3D genomes. TAD boundaries frequently coincide with genome breaks while their deletion is under negative selection, suggesting that TADs facilitate genome rearrangements and evolution. Here, we discovered that genes co-localise by evolutionary age in the human and mouse genomes, resulting in TADs having different proportions of younger and older genes. A major transition in the age co-localisation patterns is observed between the genes born as a result of the vertebrate whole genome duplications (WGDs) or before, and those born afterwards. We also found that primate- and rodent-specific genes are more frequently essential when they are located in 'aged' TADs and interact to genes that have not duplicated since the WGD. Therefore, the evolutionary success of recent genes may increase when located in TADs with established regulatory networks. Our data suggests that TADs can play a role organising ancestral functions and evolutionary novelty.

Keywords

Genome architecture, evolution, gene birth, TADs, essentiality, whole genome duplication.

Introduction

Recent advances in understanding of the modular organisation of chromatin are beginning to shed light on the local context within which genes are situated, thanks to the development of chromosome conformation capture (3C) techniques (Rada-Iglesias et al., 2018; Stadhouders et al., 2019). In 3C data, we can see that chromosomes are composed of sets of neighbouring genes and regulatory regions that tend to preferentially interact with each other in three-dimensional (3D) space, called topologically associated domains (TADs).

Accumulating data suggest that TAD boundaries could amount to important functional elements providing a modular system that might enhance the evolvability of genomes. Recent work has shown that synteny breakpoints between different mammals (Krefting et al., 2018; Lazar et al., 2018) and flies (Liao et al., 2021) are enriched at TAD boundaries. TAD boundaries are evolutionarily constrained (McArthur and Capra, 2021) and their deletion is under negative selection (Fudenberg and Pollard, 2019; Huynh and Hormozdiari, 2019). Interestingly, a comparison of linkage disequilibrium (LD) blocks and TADs revealed that TAD boundaries do not coincide with recombination hotspots, suggesting that recombination might be deleterious at these regions (Whalen and Pollard, 2019). Taken together, these recent works support the Integrative Breakage Model of genome evolution, chromatin conformation actually facilitates genome reorganisation during evolution (Farré et al., 2015).

Taken together, these results indicate that TAD boundaries are conserved but breakable, facilitating genome rearrangements during evolution while preserving intra-TAD functional and regulatory interactions. Indeed, certain classes of paralog protein-coding genes have been shown to co-localise within TADs and partially share regulatory elements in mammals (Ibn-Salem et al., 2017; Long et al., 2022) and clusters of conserved noncoding elements of key developmental genes coincide with TADs in human and flies (Harmston et al., 2017).

A gene is considered essential if it is required for the reproductive success of a cell or an organism (Rancati et al., 2018) or when loss of its function compromises the viability or fitness of the organism (Bartha et al., 2018). Evolutionarily old genes are enriched in essential genes (Bartha et al., 2018; Chen et al., 2012; Rancati et al., 2018) but we still do not understand the role genome architecture may play in the acquisition of essential functions.

Here, we investigated whether protein-coding genes that share a TAD had also originated around the same evolutionary time. We show how gene age is non-randomly distributed in the human and mouse genomes, where genes within the same TAD tend to have similar gene ages. We uncover an intimate relationship between genome architecture and gene evolution and explore its implications for the emergence of essential functions in evolutionarily young genes.

Results

Gene age co-localisation patterns in human and mouse TADs are associated with whole genome duplications

To study the distributions of gene ages in TADs, we classified human and mouse protein coding genes into 11 age groups using gene trees retrieved from Ensembl (Herrero et al., 2016) depending on the most recent common ancestor of all the extant species in which the gene is present (Juan et al., 2013; Rigau et al., 2019) (see **Figure 1**, **Table 1** and **Methods**). We consider the age of a duplicate gene as the last time it was involved in a gene duplication event, while gene ages of singleton genes were assigned from the last common ancestor to all the genes in their family according to the gene trees (**Figure 1A**). In both cases, the gene age corresponds to the latest genomic gene birth event (duplications, insertions, etc) in which the gene locus was involved (Capra et al., 2013).

Ten of the age groups correspond to the same ancestors in humans and mice (from FungiMetazoa to Eutheria) while the youngest group in each species corresponds to gene births that are Primate-specific (in humans) or Rodent-specific (in mice). For each age group in humans, we computed the frequencies of finding genes of every other age in the same TAD, using TAD maps available from human embryonic stem cells (hESCs) (Ma et al., 2015).

To test the significance of this age co-localisation, we calculated the p-values for the probability of finding gene pairs having each combination of ages (i.e. the probability of obtaining a gene of age *i* in the same TAD by chance when selecting a gene of age *j*). To calculate these p-values, we considered the null hypothesis to be the distribution of these probabilities when randomising the ages of the genes (and hence across TADs, see **Figure 1B** and **Methods**). To visualise the p-values we generated a matrix showing the probability of finding each combination of gene ages (**Figure 2**).

The age co-localisation results show that gene ages are not distributed randomly in the TADs of the human genome (**Figure 2, left**, and **Figure S1**). Surprisingly, our analysis shows that there are two clear groups of old and young gene ages that tend to be co-located with one another. The Euteleostomi genes and those born before this ancestor (Chordata, Bilateria and FungiMetazoa) are significantly co-localised within the same TADs. However, they share TADs with younger genes less often than expected by chance. In contrast, we see significant TAD co-localisation from Tetrapoda to Primate-specific age groups.

To further test the observed pattern of gene age co-localisation, we randomised the age labels of each gene and repeated the analysis. We observed no age co-localisation when age labels are randomly assigned **(Figure 1B, Figure 2**, and **Figure S2)**. This suggests that the co-localisation effect we saw between gene ages FungiMetazoa to Euteleostomi and Tetrapoda to Primate-specific was not random, but instead an effect of genome structure. To rule out that this result was dependent on the TAD definitions used (which can vary dependent on the TAD-calling method and the cell type (Rada-Iglesias et al., 2018; Zufferey et al., 2018), we performed the same analysis using independent TAD maps from human B-cell derived lymphoblastoid GM12878 cells (Rao et al., 2014), monocytes, neutrophils and

naive CD4 T-cells (Javierre et al., 2016). In all these cases, very similar two age co-localisation groups were also observed (**Figures S3** and **S4**).

We know that there has been at least one whole genome duplication (WGD) event (but most likely two) in the ancestor of vertebrates and the majority of genes duplicated during this event have not duplicated since in mammals (Makino and McLysaght, 2010; Sacerdot et al., 2018; Van de Peer et al., 2009). Previous works refer to genes born after these WGDs (in our case, corresponding to most of the Euteleostomi genes) as 'Ohnologs', in homage to Susumu Ohno's genome duplications theory (Ohno, 1970). The gene dosage of Ohnologs is known to be particularly constrained (Makino and McLysaght, 2010). More recently, it has been shown that the expansion of many gene families after the WGDs was frequently associated with gene specialisation evidenced by restricted expression and recruitment of novel, tissue-specific regulatory elements (Marlétaz et al., 2018). This suggests that the WGDs contributed to increasing the complexity of gene regulatory landscapes in vertebrates (Marlétaz et al., 2018) to a level close to other highly specialised and diverse metazoans such as Hexapoda, that have undergone similar WGD events (Li et al., 2018).

If the marked gene age co-localisation we observed in humans is associated with the WGDs that occurred at the origin of the ancestral vertebrate genome, we would expect to see a similar pattern in the mouse genome. We performed the same analysis using mouse genes and TADs from mouse embryonic stem cells (mESCs) (Bonev et al., 2017). Strikingly, a very similar significant pattern was observed in the mouse (**Figure 1, bottom-left**) that disappeared when the age labels were randomised (**Figure 1, bottom-right**). It is particularly remarkable that Primate- and Rodent-specific genes show the same pattern although they represent independent gene birth events.

In summary, the results in both species show that there is a major transition in the TAD co-localisation patterns prior to Sarcopterygii, coinciding with the time when the two rounds of WGD are thought to have occurred (Sacerdot et al., 2018). Interestingly, Ohnolog pairs have been shown to exhibit higher spatial proximity in 3D nuclear organisation than other paralog pairs (Xie et al., 2016). Taken together with our own data, these observations suggest that the constraints of the WGD are decisively influencing the genome architecture in mammalian genomes.

Gene age co-localisation is stronger in TADs than in fixed-size partitions of the human genome

We need to consider that the co-localisation of gene ages could also be associated with gene proximity in the linear genome, and the observed association with TADs could be simply due to the fact that genes within the same TAD are close to each other. To test this alternative hypothesis, we divided the human genome into windows of fixed size, using different sizes (ranging from 20 Kb to 50 Mb), and repeated the analysis for each window size (**Figure 3**). For each window size, we generated a corresponding matrix of p-values and generated a heatmap. A visual comparison of the results obtained suggests that fixed windows in the size range of TADs (300 Kb to 1Mb) show similar patterns of co-localisation, but are less evident than real TADs (**Figure 3A**).

To quantitatively compare the results of different windows and real TADs, we created an extreme pattern of co-localisation where genes perfectly cluster in either the four older (FungiMetazoa-Euteleostomi) or seven younger (Sarcopterigii-Primates) age groups. Then, we calculated the Euclidean distance between each real matrix of p-values and this extreme reference pattern of two clusters. This approach showed that the matrix with colocalisation p-values obtained with the TADs has a shorter Euclidean distance than any fixed-size windows, indicating that the TAD structure is possibly playing a role in the observed gene age distribution (**Figure 3B**).

Effect of TAD insulation, size and gene density in gene age distributions and age group co-localisation

TAD boundaries can show different degrees of insulation (Chang et al., 2020) and boundaries with higher insulation are generally associated with more CTCF binding (Cubeñas-Potts and Corces, 2015; Dixon et al., 2012). We investigated if the gene age co-localisation was related to the insulation of TADs. For this, we used the TAD boundary insulation scores calculated by Gong and co-workers using the high-resolution HiC data of the lymphoblastoid cell line GM12878 (Gong et al., 2018). We calculated a TAD insulation score by averaging the insulation scores of their boundaries and then divided the TADs into five groups based on their insulation (from very low to very high, **Figure 4**). We found that the gene age co-localisation pattern was different depending on the TAD insulation strength, which was almost lost in the two groups with the lowest insulation strength (**Figure 4A**). Interestingly, the strongest two-group pattern was observed in those TADs with intermediate insulation.

The gene age co-localisation patterns can also be affected by other factors that might vary among these five groups of TADs, such as the size of the TADs or their number of genes. Therefore, we decided to investigate the characteristics of TADs with differing insulation strengths. We observed that more insulated the TADs tend to be smaller (**Figure 4B**) but have higher gene density (**Figure 4C**). Next, we analysed how the group of older genes (Euteleostomi genes and those born before: Chordata, Bilateria and FungiMetazoa) and youngest genes born during the Primates lineage varied among the TAD insulation groups. Although both Old and Young genes tend to be more abundant in more insulated TADs (**Figure 4D**), their relative proportion changes (**Figure 4E**): Young genes are relatively more enriched in more insulated TADs. In summary, these results show that the more insulated TADs are, the more gene density they have in general, but also have a relative higher proportion of Young genes. The combination of these properties probably contributes to the differences in gene age co-localisation patterns associated with the TAD insulation strength.

Essential young genes are associated with old TADs

The co-localisation of genes by age in the TADs results in the TADs having different proportions of evolutionarily old and young genes (**Figure 5, top row**). Evolutionarily old genes are more frequently expressed across cell types and are more often essential for the organism than younger ones (Bartha et al., 2018; Rancati et al., 2018). It is normally assumed that the longer a gene has been around in a genome, the more likely that its

function is (now) important - otherwise it is likely that the gene would have eventually disappeared. However, we still don't understand how some young genes eventually become essential. To the best of our knowledge, the gene essentiality has never been analysed in the context of both its age and its genome location: could the TAD location of evolutionary young genes influence their probability of becoming essential?

We examined how the proportion of essential genes in a TAD varied depending on the ages of the genes in the TAD. Given that the transition in co-localisation pattern happened around genes aged Sarcopterygii, we classed all genes that were Euteleostomi (WGD) or older as being 'old' genes. We chose to class Primate-specific (human) and Rodent-specific (mouse) genes as 'young' genes, as this is where the lineage of the two species diverges, and so we wanted to examine whether these two groups of genes would behave similarly due to them both being the product of recent independent duplications. We used the database of Online GEne Essentiality (OGEE) (Chen et al., 2017) to determine whether each gene was essential or not. As expected, as a TAD increases in the proportion of young genes, it decreases in the proportion of essential genes, whilst the opposite was true for the proportion of old genes (**Figure 5, middle row**).

Next, we investigated if lineage-specific essential genes were enriched in TADs with other young genes or with older ones. For this, we focused our analyses on two types of TADs: 'Old TADs', in which at least 50% of the genes were aged Euteleostomi or older; and 'Young TADs' which at least 50% of the genes were Primate-specific (in human) or Rodent-specific (in mouse). We grouped the genes in these TADs depending on whether they were young genes in Young TADs, young genes in Old TADs, old genes in Young TADs, and old genes in Old TADs, and calculated the proportion of genes in each group that were essential (**Figure 2, bottom row)**. We then carried out chi-squared tests between each group to determine if there were any significant differences.

We found in both the mouse and the human dataset that young genes in Young TADs were the least likely to be essential than young genes in Old TADs, whilst the frequency of essential old genes was similar in Young and Old TADs. These data suggest that the age of neighbouring genes can help determine whether or not a young gene is essential or functionally important, while for old genes the age of the neighbouring genes does not seem to be informative (**Figure 5, middle row**).

Essential young genes show more interactions with old genes and other essential genes

The enrichment of essential young genes in Old TADs could be associated with its incorporation with established regulatory networks, which would result in an increased number of interactions with older genes. To explore this hypothesis, we analysed publicly available high-resolution promoter-capture HiC (PCHi-C) networks from mouse (Schoenfelder et al., 2015) and human cells (Javierre et al., 2016). Each network consists of nodes that represent the PCHi-C fragments and edges that represent their interactions (Pancaldi et al., 2016). To compare the interactions of young essential genes with those of young non-essential genes we constructed a sub-network from the PCHi-C data in each species keeping only interactions that involve nodes with lineage-specific genes (**Methods**). From there, the subnetwork was divided into two further subnetworks; one subnetwork

where every pair of interacting nodes contained a node with an essential young gene, and a second subnetwork where every pair of interacting nodes contained a node with a non-essential young gene. For both subnetworks, we looked at the age and essentiality of the interactors (**Figure 5, bottom row**).

We found that essential lineage-specific genes have more interactions with old genes than non-essential lineage-specific genes in both mice ($p = 7.7 \times 10^{-90}$) and humans ($p = 9.2 \times 10^{-21}$). Essential Rodent-specific and Primate-specific genes also showed significantly higher interactions with other essential genes, and a similar trend was observed when we looked at interactors that were both old and essential (**Figure 5, bottom row**). These results suggest that, independently of any definition of TADs, 3D interactions with old and/or essential genes are associated with higher probability of a young gene to be essential.

Zinc Finger Protein 141 (ZNF141) is a representative example of a young human essential gene (based on CRISPR-Cas9 screening data (Chen et al., 2017)) located in a TAD rich in old genes; out of the 27 genes in the TAD, 22 (81.5%) are aged Eutelostomi or older (**Figure 6**). This TAD is in the telomeric region on chromosome 4p16.3 and contains another 4 essential genes, three of which are directly interacting with ZNF141 in hESC DNAse HiC data (Ma et al., 2015): MAEA (Bilateria), GAK (Euteleostomi) and ZNF721 (Primates). Interestingly, mutations in ZNF141 have been shown to be associated with postaxial polydactyly type A malformation characterised by well-formed functionally developed 5th digit duplication in hands and/or feet (Kalsoom et al., 2013).

Discussion

We have found that the TADs in human and mouse genomes coincide with clusters of genes that are close in their evolutionary age. The local expansion of duplicating gene families is probably a key factor contributing to the co-localisation of younger genes. It is possible that the structure of TADs that are rich in old genes is under negative selection and the birth of new genes there is detrimental. This might have been important at least right after the WGDs and then promoted by the local expansions of genes afterwards, which tend to accumulate in error-prone late replicating regions of the genome (Juan et al., 2013). Another possibility is that the lower rates of mutations and copy number variants in early replicating TADs, rich in old genes, reduces the probability of genes being born there, independently of negative selection biases. Early replicating TADs are expected to have more efficient DNA repair (Gaboriaud and Wu 2019) to reduce the probability of having damaging mutations; the low frequency of young genes in these TADs there might be a secondary effect of a lower mutation rate in these regions. Future availability of TADs in a wider diversity of species will allow to investigate the conservation of these TADs and their chromatin interactions (Álvarez-González et al., 2022; Li et al., 2022).

The enrichment of lineage-specific essential genes in Old TADs could be associated to different factors: the higher probability of the chromatin in these TADs to be open, functional relationships with the same essential processes where their neighbours are involved or new regulatory and/or structural roles keeping the integrity of the 3D structure. Some of these lineage-specific essential genes could indeed have originated from parental genes that were already essential in the ancestral species. Moreover, enhancers have been shown to be a

possible substrate for the origin of new genes (Carelli et al., 2018; Majic and Payne, 2020) and it is also possible that the essentiality retrieved from CRISPR screens is actually capturing the essentiality of the regulatory role of the genomic regions. Indeed, promoters with enhancer function (ePromoters) of interferon signalling response have been shown to be more conserved than other non-enhancer genes belonging to the same pathway (Santiago-Algarra et al., 2021).

Finally, our finding has a simple but very important implication: TADs can be classified by the age of their genes. This can be useful to study the evolution of different biological processes, such as regulatory modules of development, cell type ontogenies or cancer. Here, we explored the relationship between TAD age, gene age and gene essentiality, showing that young essential genes tend to be located in TADs rich in genes that are older than them. This could help in prioritising the study of genes of unknown function by taking into account the age of their neighbours.

Methods

Dating ages in genes

An age was assigned to mouse and human genes using Ensembl gene trees (Herrero et al, 2016) as described before (Juan et al., 2013; Rigau et al., 2019) and illustrated in Figure 1A. Our definition of gene age tries to date the last time a gene was involved in a gene birth event. In the case of genes that have been duplicated (according to the Ensembl trees), we consider the node of their most recent duplication - accordingly, daughter genes will have the same age (if none them are involved in more recent duplication). In the case of singleton genes (i.e., those without detectable duplicates), we consider their gene age as the last common ancestor of all species that have orthologs; in other words, the gene age is the gene family birth event (for a review about different ways to date gene ages see (Capra et al, 2013).

The age groups Simiiformes, Catarrhini, Hominoidea, Hominidae, HomoPanGorilla and Homo sapiens were collapsed into Primate-specific genes; while Rodent-specific merges Glires, Rodentia, Sciurognathi, Murinae and Mus musculus (here, we decided to name this group as Rodent-specific because only the birth of 14 mouse genes were dated as Glires). Ten of the age groups correspond to the same ancestors in humans and mice (from FungiMetazoa to Eutheria) while the youngest group in each species correspond to gene births that are Primate-specific (in humans) or Rodent-specific (in mice).

TAD maps

We used the TAD genomic coordinates as defined by the original papers: hESCs (Ma et al., 2015), mESCs (Bonev et al., 2017), human GM12878 B-cell derived lymphoblastoid cells (original TADs (Rao et al., 2014) and TADs from Gong and co-workers (Gong et al., 2018)), monocytes, neutrophils and naive CD4 T-cells (Javierre et al., 2016). We only modified the original data from Rao *et al* to remove the nested TADs, as they would have given duplicated colocalisations.

Gene age co-localisation analysis

We generated the p-value matrices by performing fifteen randomisations to obtain the variance of the distribution of the probabilities to find each pair of gene ages within the same TAD. Then, using the normal distribution, the p-value were calculated as the probability of obtaining by chance alone the observed probability in the random distribution. We took as average the value of the null hypothesis, which was considered to be the probability of matching every age-pair without taking any TAD structure at all (considering the ages as being taken from a global pool rather than from any particular TAD). To facilitate the visualisation of these p-values, we calculated the natural cologarithm using a maximum value of 30. In order to differentiate between probabilities higher or lower than expected, the natural cologarithm (always positive, since it is from a value lower than 1) was left unchanged (ie, positive) if the probability of finding genes of an age pair together was higher than expected, or its sign was changed (becoming negative) for probabilities of finding a pair of ages lower than expected. The code to implement this is in function getPairData, of file https://github.com/ricolab/TAD_Evolution/blob/master/colocalisation_pValues/colocalisations_functions.R

Gene essentiality

Gene essentiality data from humans and mice was retrieved from OGEE database (Chen et al., 2017) and the code to reproduce he analysis is available at https://github.com/ricolab/TAD_Evolution/blob/master/TAD_Age_and_Essentiality_Analysis/FullScript

PCHi-C networks

Networks were constructed from PCHiC data as previously described (Pancaldi et al., 2016). We annotated each node with the genes and their ages. We created a subnetwork of young genes, keeping interactions where at least one node contains a lineage-specific gene. Then, we grouped the nodes by gene and removed any duplicated edges so that each gene-gene interaction would only be represented once.

Data availability

All the data analysed was publicly available. The data and code to reproduce our analyses is available at: https://github.com/ricolab/TAD_Evolution

Acknowledgements

We are grateful to Maninder Heer and Maria Rigau for their critical comments. This work was supported by a Wellcome Trust (https://wellcome.ac.uk) Seed Award in Science (206103/Z/17/Z) awarded to DR.

Conflicts of interest

The authors declare that they have no conflict of interest.

Author contributions

Conceptualization: CJ, MT, DJ and DR; Analysis: CJ, MT and DR; Writing: CJ, MT, DJ and DR; Supervision, DJ and DR; Funding Acquisition, DR.

References

Álvarez-González, L., Burden, F., Doddamani, D., Malinverni, R., Leach, E., Marín-García, C., Marín-Gual, L., Gubern, A., Vara, C., Paytuví-Gallart, A., et al. (2022). 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. Nat. Commun. *13*, 2608. .

Bartha, I., di Iulio, J., Venter, J.C., and Telenti, A. (2018). Human gene essentiality. Nat. Rev. Genet. *19*, 51–62. .

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557–572.e24. .

Capra, J.A., Stolzer, M., Durand, D., and Pollard, K.S. (2013). How old is my gene? Trends Genet. *29*, 659–668. .

Carelli, F.N., Liechti, A., Halbert, J., Warnefors, M., and Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. Nat. Commun. *9*, 4066. .

Chang, L.-H., Ghosh, S., and Noordermeer, D. (2020). TADs and Their Borders: Free Movement or Building a Wall? J. Mol. Biol. *432*, 643–652. .

Chen, W.-H., Trachana, K., Lercher, M.J., and Bork, P. (2012). Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. Mol. Biol. Evol. *29*, 1703–1706. .

Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2017). OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. Nucleic Acids Res. *45*, D940–D944. .

Cubeñas-Potts, C., and Corces, V.G. (2015). Topologically Associating Domains: An invariant framework or a dynamic scaffold? Nucleus *6*, 430–434. .

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. .

Farré, M., Robinson, T.J., and Ruiz-Herrera, A. (2015). An Integrative Breakage Model of genome architecture, reshuffling and evolution: The Integrative Breakage Model of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity. Bioessays

37, 479–488. .

Fudenberg, G., and Pollard, K.S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proc. Natl. Acad. Sci. U. S. A. *116*, 2175–2180. .

Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I., and Tsirigos, A. (2018). Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nat. Commun. *9*, 542. .

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat. Commun. *8*, 441. .

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database *2016*. https://doi.org/10.1093/database/bav096.

Huynh, L., and Hormozdiari, F. (2019). TAD fusion score: discovery and ranking the contribution of deletions to genome structure. Genome Biol. *20*, 60. .

Ibn-Salem, J., Muro, E.M., and Andrade-Navarro, M.A. (2017). Co-regulation of paralog genes in the three-dimensional chromatin architecture. Nucleic Acids Res. *45*, 81–91. .

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell *167*, 1369–1384.e19. .

Juan, D., Rico, D., Marques-Bonet, T., Fernández-Capetillo, O., and Valencia, A. (2013). Late-replicating CNVs as a source of new genes. Biol. Open *2*, 1402–1411. .

Kalsoom, U.-E.-, Klopocki, E., Wasif, N., Tariq, M., Khan, S., Hecht, J., Krawitz, P., Mundlos, S., and Ahmad, W. (2013). Whole exome sequencing identified a novel zinc-finger gene ZNF141 associated with autosomal recessive postaxial polydactyly type A. J. Med. Genet. *50*, 47–53. .

Krefting, J., Andrade-Navarro, M.A., and Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. BMC Biol. *16*, 87. .

Lazar, N.H., Nevonen, K.A., O'Connell, B., McCann, C., O'Neill, R.J., Green, R.E., Meyer, T.J., Okhovat, M., and Carbone, L. (2018). Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. Genome Res. *28*, 983–997. .

Li, D., He, M., Tang, Q., Tian, S., Zhang, J., Li, Y., Wang, D., Jin, L., Ning, C., Zhu, W., et al. (2022). Comparative 3D genome architecture in vertebrates. BMC Biol. *20*, 99. .

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J., and Barker, M.S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc. Natl. Acad. Sci. U. S. A. *115*, 4713–4718. .

Liao, Y., Zhang, X., Chakraborty, M., and Emerson, J.J. (2021). Topologically associating domains and their role in the evolution of genome structure and function in. Genome Res. *31*, 397–410. .

Long, H.S., Greenaway, S., Powell, G., Mallon, A.-M., Lindgren, C.M., and Simon, M.M. (2022). Making sense of the linear genome, gene function and TADs. Epigenetics Chromatin

15, 4. .

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat. Methods *12*, 71–78. .

Majic, P., and Payne, J.L. (2020). Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. Mol. Biol. Evol. *37*, 1165–1178. .

Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. U. S. A. *107*, 9270–9274. .

Marlétaz, F., Firbas, P.N., Maeso, I., Tena, J.J., Bogdanovic, O., Perry, M., Wyatt, C.D.R., de la Calle-Mustienes, E., Bertrand, S., Burguera, D., et al. (2018). Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature *564*, 64–70. .

McArthur, E., and Capra, J.A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. Am. J. Hum. Genet. *108*, 269–283. .

Ohno, S. (1970). Evolution by Gene Duplication. https://doi.org/10.1007/978-3-642-86659-3.

Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A., and Rico, D. (2016). Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol. *17*, 152. .

Rada-Iglesias, A., Grosveld, F.G., and Papantonis, A. (2018). Forces driving the three-dimensional folding of eukaryotic genomes. Mol. Syst. Biol. *14*, e8214. .

Rancati, G., Moffat, J., Typas, A., and Pavelka, N. (2018). Emerging and evolving concepts in gene essentiality. Nat. Rev. Genet. *19*, 34–49. .

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. .

Rigau, M., Juan, D., Valencia, A., and Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. PLoS Genet. *15*, e1007902. .

Sacerdot, C., Louis, A., Bon, C., Berthelot, C., and Roest Crollius, H. (2018). Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. *19*, 166. .

Santiago-Algarra, D., Souaid, C., Singh, H., Dao, L.T.M., Hussain, S., Medina-Rivera, A., Ramirez-Navarro, L., Castro-Mondragon, J.A., Sadouni, N., Charbonnier, G., et al. (2021). Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. Nat. Commun. *12*, 6660. .

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. *25*, 582–597. .

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354. .

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. *10*, 725–732. .

Whalen, S., and Pollard, K.S. (2019). Most chromatin interactions are not in linkage disequilibrium. Genome Res. *29*, 334–343. .

Xie, T., Yang, Q.-Y., Wang, X.-T., McLysaght, A., and Zhang, H.-Y. (2016). Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance. Mol. Biol. Evol. *33*, 2368–2375. .

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. Genome Biol. *19*, 217. .

Figure legends

Figure 1. **Dating gene ages and TAD co-localisation analysis.**

A. Toy gene trees illustrating how the gene age corresponds to the most recent duplication node for duplicated genes (left) or the last common ancestor for singleton genes (right). In both cases, the gene age indicates the last time a gene was involved in a gene birth event.
B. Cartoon representing three TADs and their genes. Genes are kept in their original locations and gene age labels are randomly relocated.

Figure 2. **Genes co-localise by age in TADs.**

Representation of the matrices displaying the evidence that genes with any combination of two ages are found together in the same TAD by using the natural cologarithm of the p-value, using a positive (negative) sign when this is higher (lower) than expected. Increasing intensities of red (blue) show increasing evidence that the combination is found together more often (less often) than expected by chance, white representing that the match found is exactly as expected by chance. See Table 1 for the correspondence between abbreviated and full gene ages.
Left: observed situation for human and mouse. See Figure S1 for the same matrices displaying the values obtained.
Right: representative results of the same operation after randomising the age labels of each gene (so each age group keeps the same number of genes, but just assigned randomly). See similar results with additional randomisations in Figure S3.

Figure 3. **Comparison of fixed-size windows and real TADs.**

A. Each heatmap represent the gene age-colocalisation p-values (see colour legend and more details in Figure 2) obtained using the fixed sized windows of the indicated sizes - from 20 Kb to 50 Mb. The heatmap with p-values obtained with the hESC TADs (same as the one in Figure 2) is included for comparison.
B. Quantitative comparison of the p-value matrices shown in panel A. The reference matrix represents an extreme two-group co-localisation pattern, and the y-axis indicates the Euclidean distance of each real p-value matrix to this reference. The x-axis represent the size of the fixed-size window used (and the ranges of sizes for TADs).

Figure 4. **TAD boundary insulation.**

A. Co-localisation p-values obtained for the five groups of TADs of GM12878 lymphoblastoid cells, based on their boundary insulation strength. For details see Figure 2 and legend.

B. Average TAD size for the five groups of TADs in panel A.

C. Average number of genes per TAD (blue bars) and mean gene density (number of genes divided by TAD size, orange bars).

D. Mean number of Old (yellow bars) and Young (purple bars) genes per TAD.

E. Average ratio of Old vs Young genes per TAD.

Figure 5. Different proportions of young genes, old genes and essential genes across TADs in mice (left) and humans (right).

Top: variation of the percentage of essential genes in comparison to the percentage of old genes.

Middle: percentage of essential genes across four gene 'groups' - young genes in Young TADs, young genes in Old TADs, old genes in Young TADs and old genes in Old TADs. The asterisks indicate the p-values when carrying out chi-squared tests between the different groups. Significance of chi-squared tests: * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, (ns: non significant).

Bottom: comparison of interactions with old, essential, and old and essential interactors for essential Rodent-specific and non-essential Rodent-Specific genes (left) and essential Primate-specific and non-essential Primate-specific genes (right). Asterisks indicate that p-values for the chi squared tests were all below 0.001.

Figure 6. HiC interaction network for the TAD containing ZNF141.

Nodes represent genes within the TAD, whilst the edges represent the interactions between the genes. Essential genes are represented with a large node, whilst non-essential genes are represented with a small node. Black edges represent interactions involving ZNF141. The age of a gene is represented by the colour of the node.

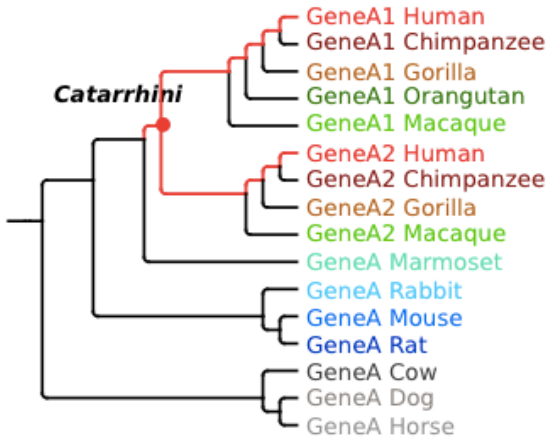
Tables

Gene age	Abbreviation	Evolutionary Time Periods (Mya)
Fungi-Metazoa	FungM	-1270
Bilateria	Bilat	1270-680
Chordata	Chord	680-586
Euteleostomi	Eutel	586-431
Sarcopterygii	Sarcop	431-414
Tetrapoda	Tetra	414-353
Amniota	Amnta	353-319
Mammalia	Mamml	319-180
Theria	Thria	180-160
Eutheria	Euthr	160-94
Primates	PrmtSp	94-0
Rodents	RdnSp	94-0

Table 1. Gene ages used in this study. It includes the abbreviations used in some figures and the evolutionary time periods covered (in million of years). All ages are shared between human and mouse, except Primates (Primates specific, or PrmtSp, in human) and Rodents (Rodent specific, or RdnSp, in mouse).

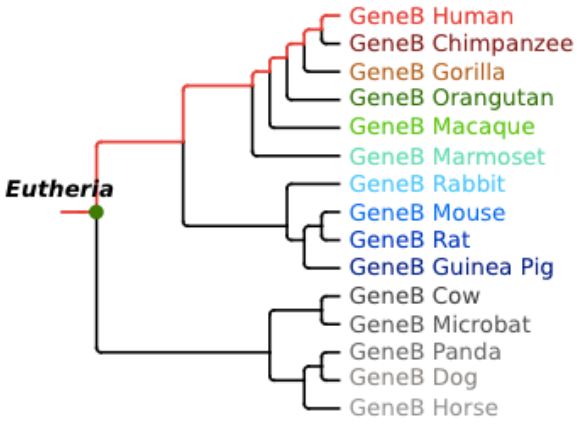
A

Age of a duplicated gene



Gene duplication event

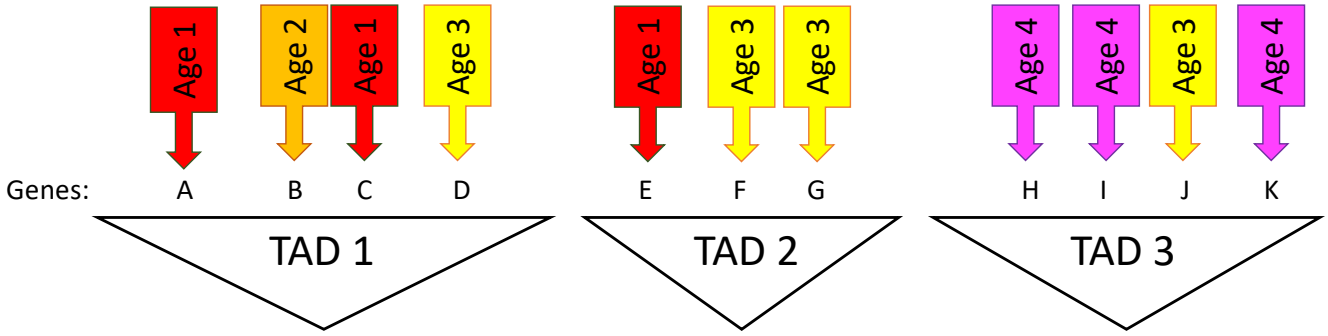
Age of a singleton gene



Gene family birth event

B

Real age labels



Randomized age labels

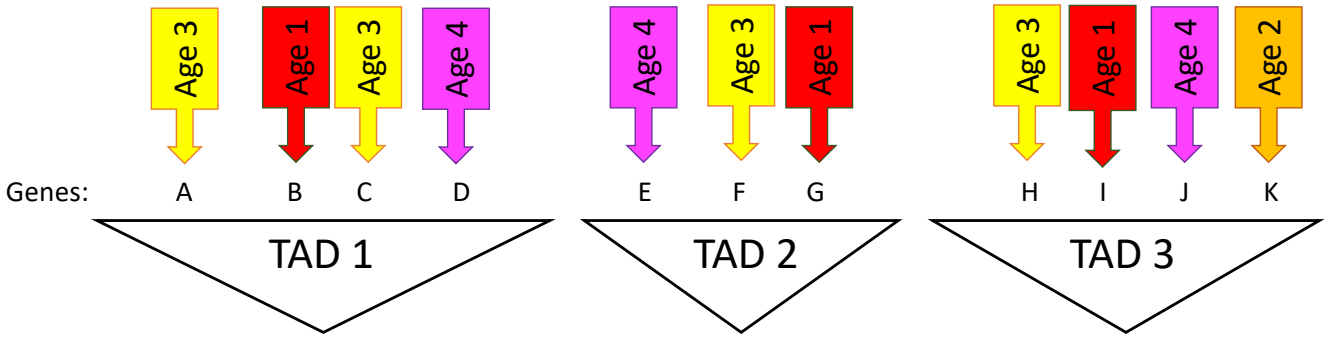


Figure 1

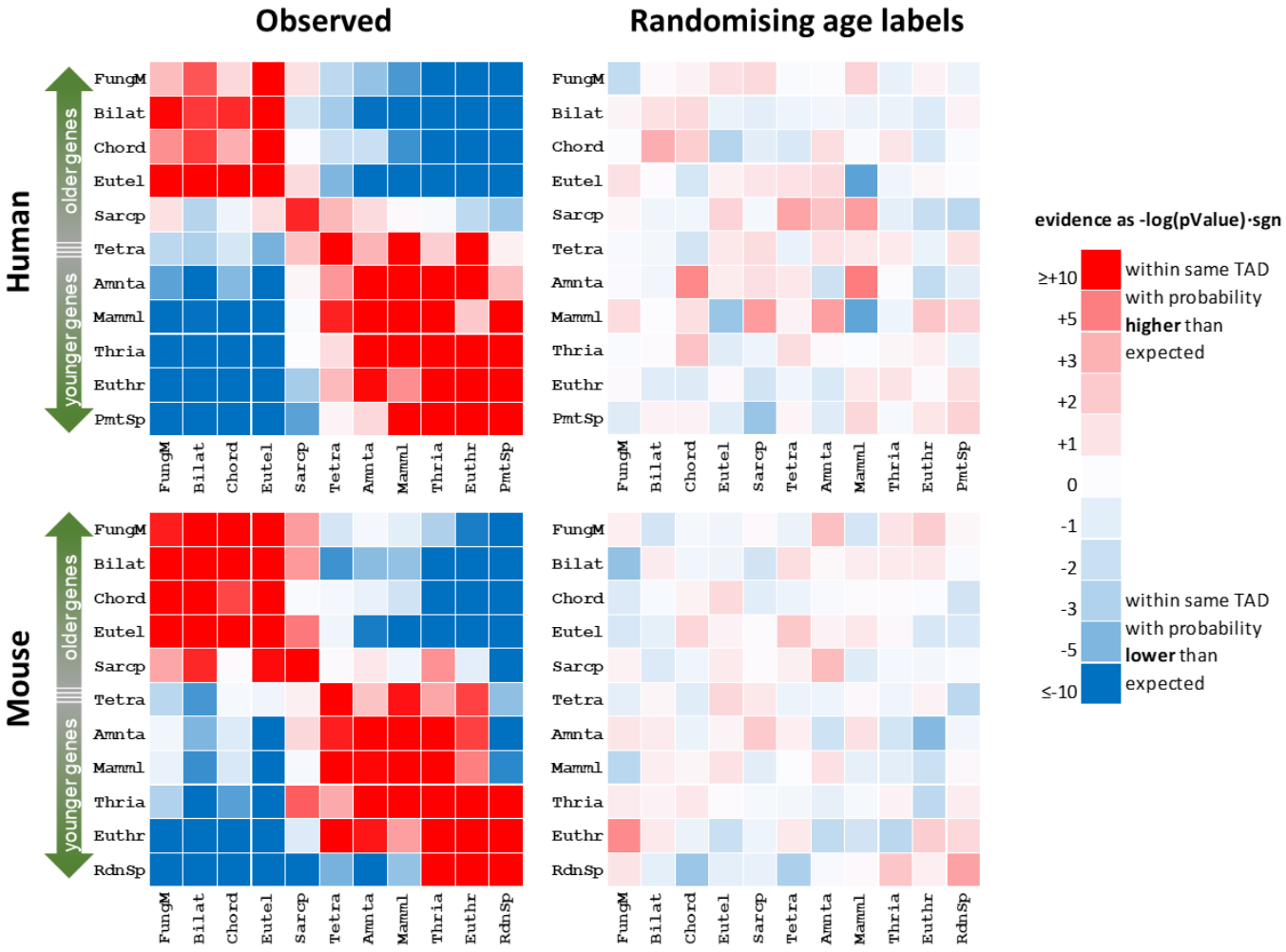
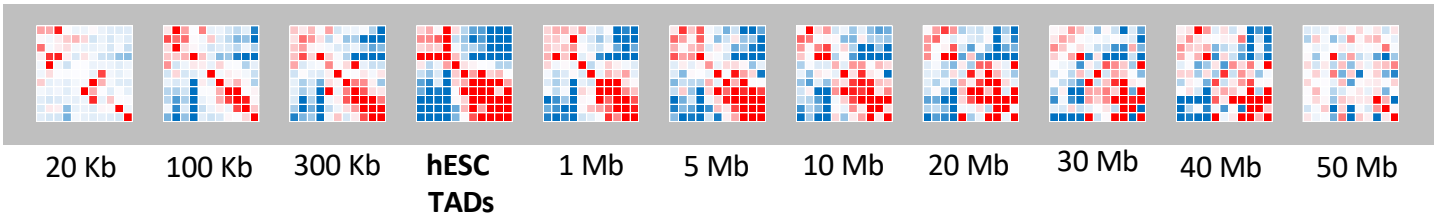

Observed
Randomising age labels
Human
Mouse
older genes
younger genes
FungM
Bilat
Chord
Eutel
Sarcp
Tetra
Amnta
Mamml
Thria
Euthr
PmtSp
RdnSp
evidence as -log(pValue)·sgn
≥+10
+5
+3
+2
+1
0
-1
-2
-3
-5
≤-10
within same TAD
with probability
higher than
expected
within same TAD
with probability
lower than
expected

Figure 2

A

Age co-localisation at different window sizes



B

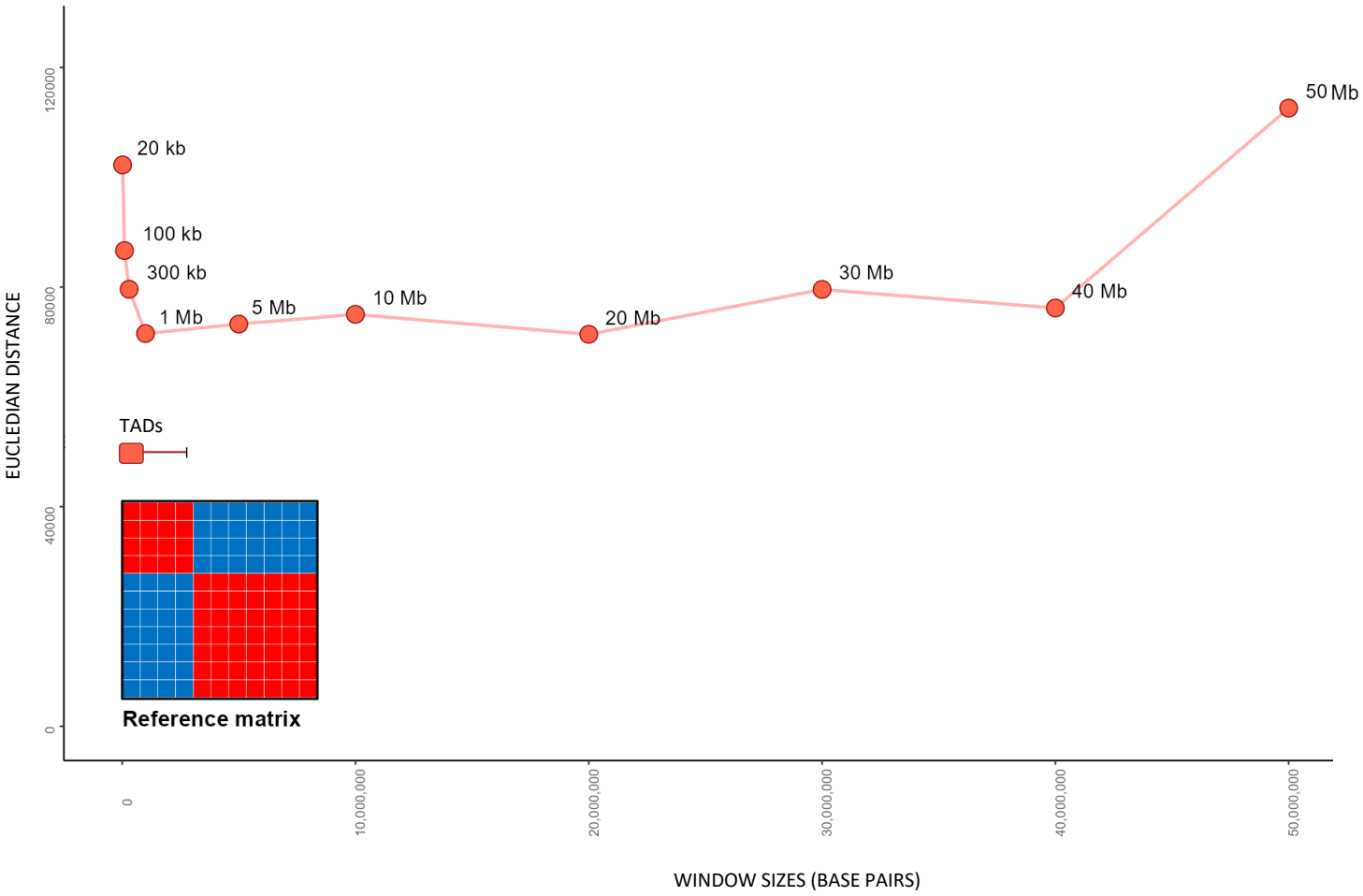
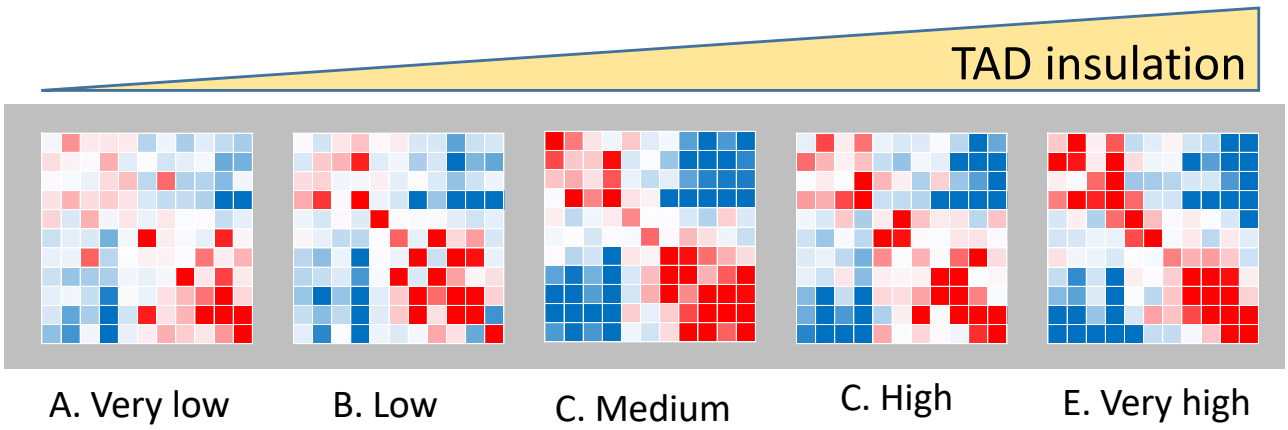
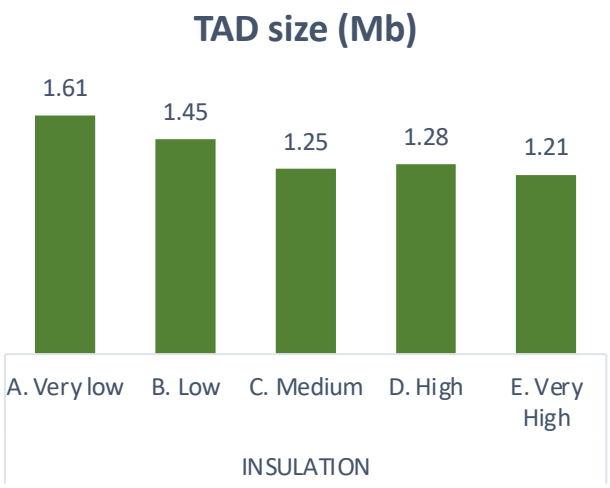


Figure 3

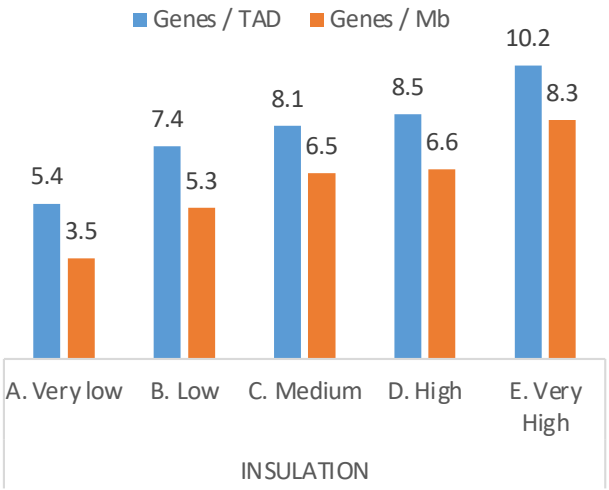
A



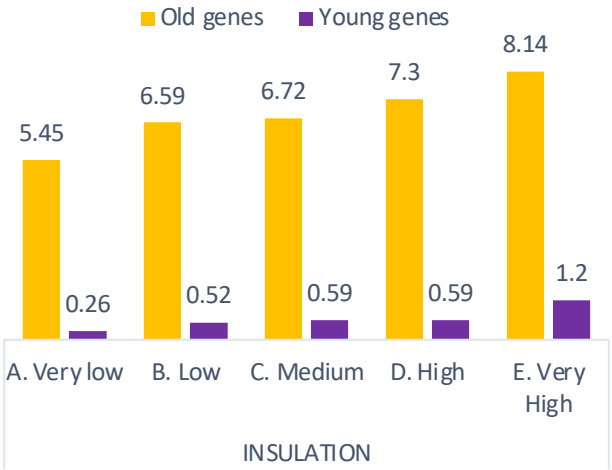
B



C



D



E

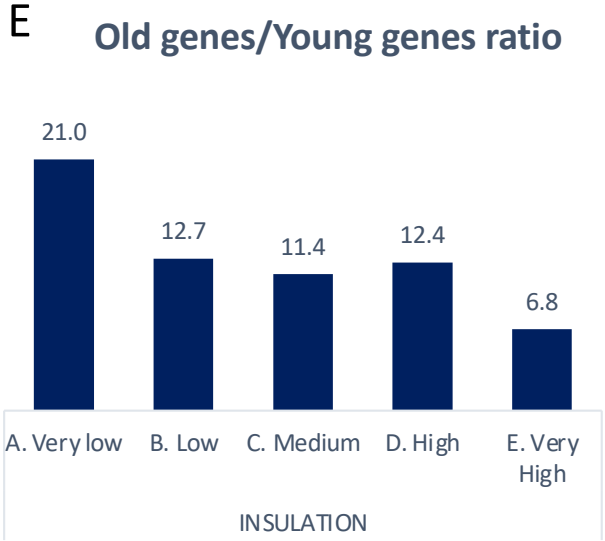
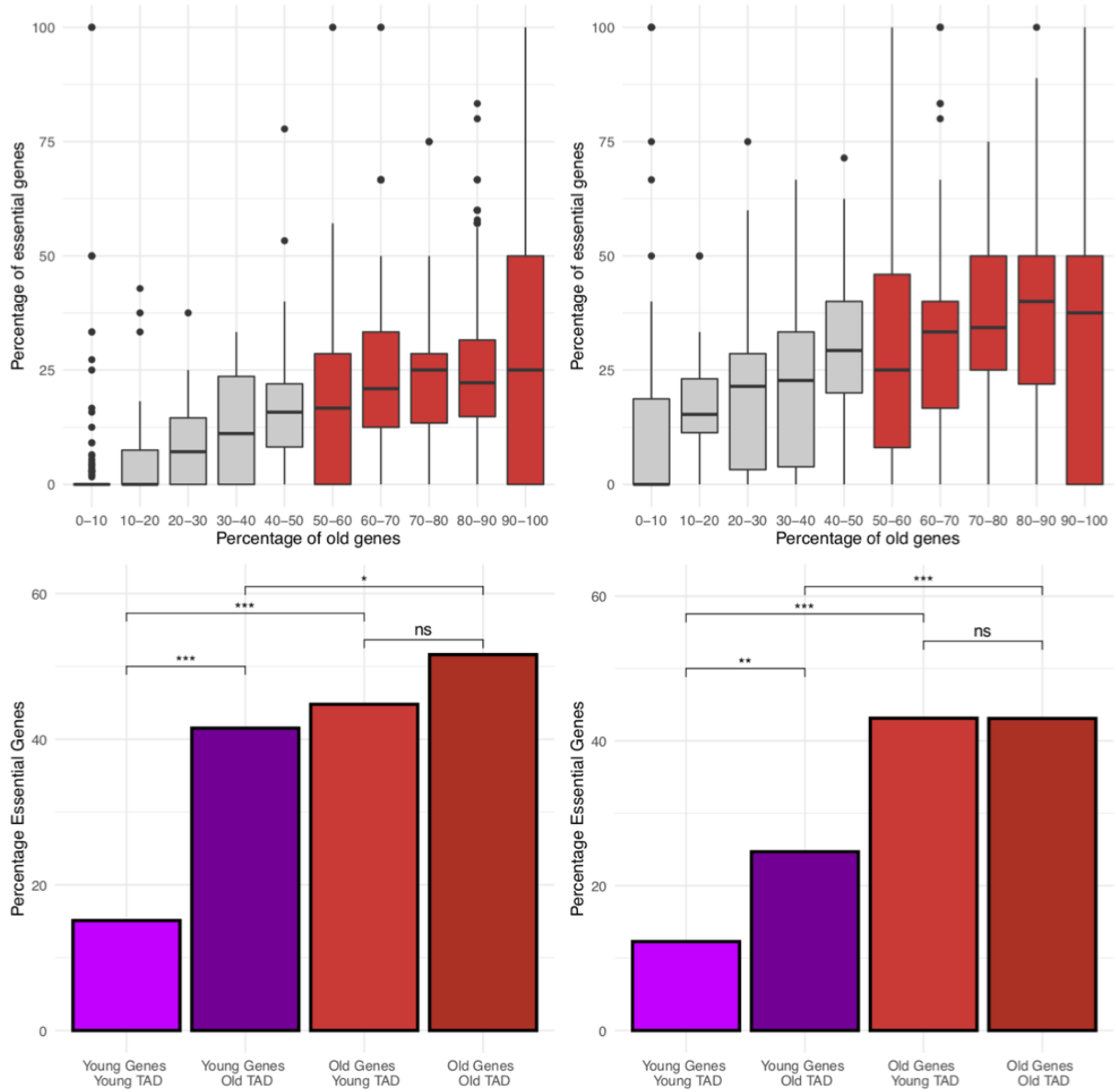


Figure 4


Percentage of essential genes
Percentage of old genes
0–10
10–20
20–30
30–40
40–50
50–60
60–70
70–80
80–90
90–100
Percentage Essential Genes
Young Genes Young TAD
Young Genes Old TAD
Old Genes Young TAD
Old Genes Old TAD
ns

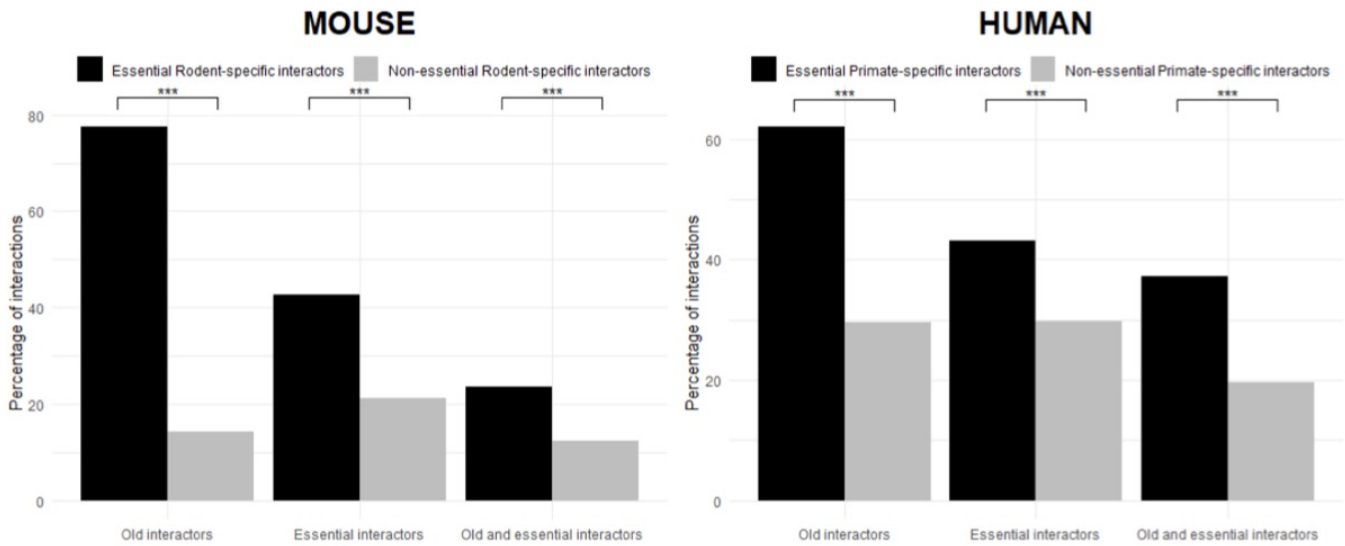

MOUSE
HUMAN
Essential Rodent-specific interactors
Non-essential Rodent-specific interactors
Essential Primate-specific interactors
Non-essential Primate-specific interactors
Percentage of interactions
Old interactors
Essential interactors
Old and essential interactors

Figure 5

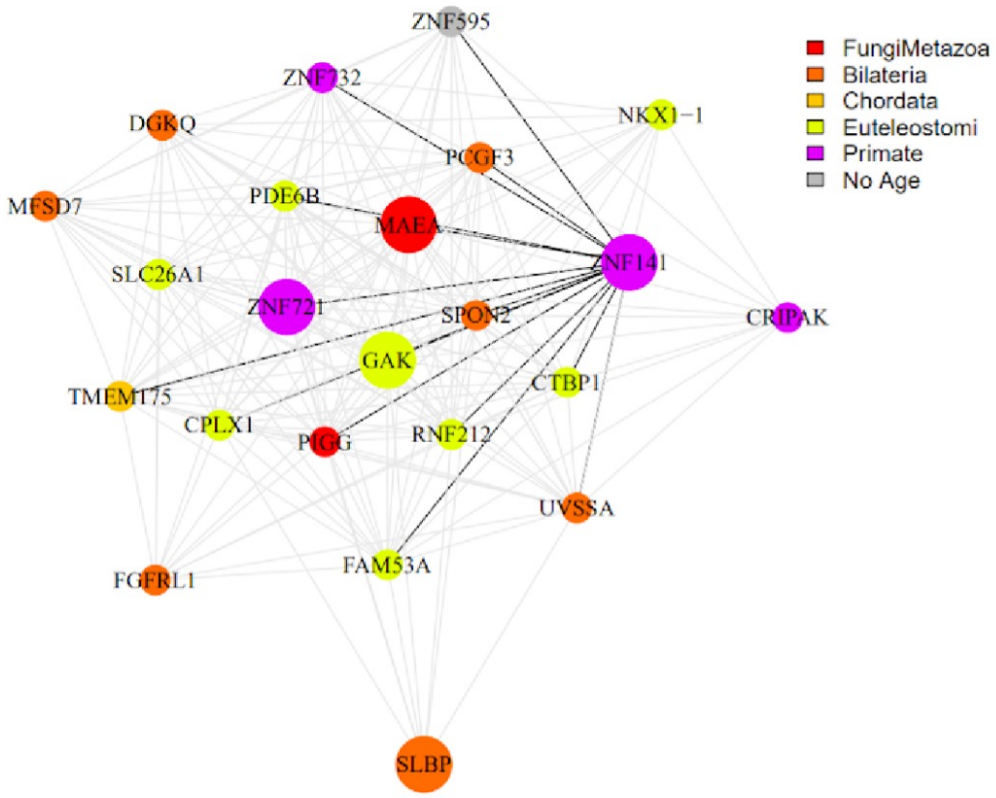

ZNF595
ZNF732
DGKQ
NKX1-1
PCGF3
MFSD7
PDE6B
MAEA
ZNF141
SLC26A1
ZNF721
SPOM2
CRIPAK
GAK
CTBP1
TMEM175
CPLX1
PIGG
RNF212
UVSSA
FGFRL1
FAM53A
SLBP
FungiMetazoa
Bilateria
Chordata
Euteleostomi
Primate
No Age

Figure 6