The impact of whole genome duplications on the human gene regulatory networks

Francesco Mottes, Chiara Villa, Matteo Osella, Michele Caselle

The impact of whole genome duplications on the human gene regulatory networks

Francesco Mottes [1], **Chiara Villa** [2], **Matteo Osella** [1] **and Michele Caselle** [1*]

[1]Department of Physics, University of Turin & INFN, Via Pietro Giuria 1, I-10125 Turin, Italy
[2]School of Mathematics and Statistics, University of St Andrews, Mathematical Institute, North Haugh, St Andrews KY16 9SS, UK

ABSTRACT

This work studies the effects of the two rounds of Whole Genome Duplication (WGD) at the origin of the vertebrate lineage on the architecture of the human gene regulatory networks. We integrate information on transcriptional regulation, miRNA regulation, and protein-protein interactions to comparatively analyse the role of WGD and Small Scale Duplications (SSD) in the structural properties of the resulting multilayer network. We show that complex network motifs, such as combinations of feed-forward loops and bifan arrays, deriving from WGD events are specifically enriched in the network. Pairs of WGD-derived proteins display a strong tendency to interact both with each other and with common partners and WGD-derived transcription factors play a prominent role in the retention of a strong regulatory redundancy. Combinatorial regulation and synergy between different regulatory layers are in general enhanced by duplication events, but the two types of duplications contribute in different ways. Overall, our findings suggest that the two WGD events played a substantial role in increasing the multi-layer complexity of the vertebrate regulatory network by enhancing its combinatorial organization, with potential consequences on its overall robustness and ability to perform high-level functions like signal integration and noise control.

INTRODUCTION

Gene duplication is one of the main drivers of evolutionary genomic innovation (1, 2, 3). Small Scale Duplications (SSDs) typically involve a single gene or a small set of genes within a well defined genomic locus. More rarely, a large-scale genomic duplication may occur, which involves a macroscopic portion of the genome. Such events are called Whole Genome Duplications (WGDs), and it is by now clear that they played a major role in evolution (4, 5). SSD events can induce a local exploration of the phenotypic landscape, introducing small and incremental changes to the genome and consequently to the cell functions. WGD events, on the other hand, typically entail more sudden and dramatic phenotypic changes. They also most likely produce immediate dire consequences on the fertility and fitness of the organism that comprimise its short-term survival (6). As a result, most WGD events are not fixated in the population. In some peculiar circumstances, though, they can constitute an immediate evolutionary advantage. Increasing evidence points towards a central role of WGD in the successful response to sudden environmental changes and stress (5). Furthermore, fixated WGD events can boost the biological complexity of the organism in the long term (5).

This paper focuses specifically on the human genome, and thus on the two rounds of WGDs that occurred about $500-550$ Millions of years ago. More than 50 years ago Susumu Ohno (1) proposed in a seminal paper that two rounds of WGD were at the origin of the vertebrate lineage. The hypothesis was met with both interest and skepticism, and it was only the advent of high-throughput sequencing that provided reliable evidence supporting ancient WGD events. In 1997, a WGD event was unambiguously detected for the first time in Saccharomyces cerevisiae (7, 8) and a few years later in Arabidopsis thaliana (9). Finally, in 2005 Ohno's original intuition regarding the two WGD events at the origin of the vertebrate lineage was also confirmed (10), and WGD duplicates are now also called "ohnologs" in his honour. These events are conjectured to have played a central role in the evolution of complex traits associated with vertebrates. For example, a multi-omics analysis of the Amphioxus genome has shown that the two rounds of vertebrate WGD significantly increased the complexity of the vertebrate regulatory landscape, and possibly boosted the evolution of morphological specializations (11). It was also shown that an important class of human highly interacting proteins, involved in processes that are crucial for the organization of multicellularity, was mainly created by vertebrate WGD (12).

The identification of WGD pairs or quartets in vertebrates is a highly non trivial task because of their ancient origin (13). In fact, a stable and reliable list of human WGD gene pairs

*To whom correspondence should be addressed. Email: caselle@to.infn.it

was only recently proposed (14, 15, 16). This advance made it finally possible to analyze the evolutionary role of WGD and SSD also in human. As a consequence, few interesting features have been identified to be uniquely associated to WGD pairs. For example, WGD genes are subject to more stringent dosage balance constraints and are more frequently associated with disease with respect to other genes (17). Moreover, WGD genes are threefold more likely than non WGD ones to be involved in cancers and autosomal dominant diseases (14). This observation led to the suggestion that WGD genes are intrinsically "dangerous", in the sense that they are more susceptible to dominant deleterious mutations than other genes (18). From a functional point of view, WGD genes are more frequently involved in signalling, development and transcriptional regulation and they are enriched in Gene Ontology categories generically associated to organismal complexity (14, 15, 19, 20, 21). From the gene expression point of view, both the gene expression profile and the subcellular localization seem to be more divergent between the two partners of a WGD-derived pair than for gene pairs derived from SSD (15). In the same work, the authors also note that WGD-derived genes contain a larger proportion of essential genes than the SSD ones and that they are more evolutionary conserved than SSD. Remarkably, several of these recent observations on vertebrates WGD genes agree with what was found years ago both in yeast (22) and in A. thaliana (20, 23). This "universality" supports the hypothesis of general principles or mechanisms behind the unexpected retention of WGD genes and their interactions.

The goal of the present work is to pinpoint the different roles played by the two types of gene duplications - SSD and WGD - in shaping the architecture of the human gene regulatory network. In particular, we investigate the local structure - mainly by analysing the network motif enrichments - of the transcriptional regulatory network, the protein-protein interaction network and the miRNA-gene interaction network, which are partially represented in fig. 1A, B, and C respectively. Network motifs are statistically enriched subgraphs that can be found in many complex networks (24) and they assume particular significance in biology and for gene regulatory networks in particular. In fact, in this context network motifs identified gene circuits that can perform relatively simple computations with specific biological functions. These simple modules can then assemble into a larger network to implement complex and robust regulatory strategies (25). As shown in fig. 1E, gene duplications - and WGD in particular - can create motifs in a very straightforward way by duplicating the genes involved in a simple regulatory interaction. Even though this is certainly not the only way in which motifs may be created, we expect duplication events to have a major impact on the creation and, most importantly, the subsequent retention of these local structures.

We therefore analyzed the statistical enrichment of a selection of motifs - represented in fig. 1D - whose functional importance is widely recognized (25). We observe that SSD and WGD gene pairs are statistically over-represented in different types of motifs. This result is in general agreement with previous observations on the yeast transcriptional network (26). We will show that also the structure of additional layers of regulation present in the human genome, such as miRNA regulation, has also been influenced by duplication events. In conclusion, this work shows that SSD and WGD events shaped the multiple layers of regulation in the human genome in different ways and jointly contributed to their current structure. The specific consequences of WGD events on the regulatory network seem to be associated to an increased redundancy and complexity that would be hard to attain through a sequence of small-scale events.

MATERIALS AND METHODS

Small-scale and Whole-genome duplicates

WGD paralogues. The WGD gene pairs were obtained by merging the results of *Makino and McLysaght* (17) with the latest available OHNOLOGS database (16). In order to have a high-confidence list of paralogies, we considered only WGD couples corresponding to the *strict* criterion in the OHNOLOGS database. Moreover, all the couples that were not recognized as paralogues in the current version of the Ensembl database were excluded. To ensure full compatibility among all of the datasets employed, we updated the gene names to the latest officially accepted version - data about the status of gene names were obtained from the HGNC online service (27). Finally, only protein-coding genes (according to the Ensembl database) were considered in our list of paralogues. With these restrictions, we ended up with a list of 8070 WGD-derived paralogue couples, comprising 7324 different genes.

SSD paralogues. SSD-derived paralogues were obtained from the list of all human paralogues involving protein-coding genes in the Ensembl database (28), and subtracting from this list all of the couples that were previously identified as derived from a WGD. One additional factor that must be taken into account when dealing with the distinction between WGD and SSD couples is the huge spread of duplication ages of the SSD paralogues. The two rounds of WGD happened relatively close in time, approximately around the appearance of the Vertebrate lineage ~ 500 Mya. Given this timescale, it is reasonable to assume that the currently retained WGD gene couples have experienced similar evolutionary forces (neutral or selective). On the other hand, SSD couples are continuously generated throughout the history of the human genome evolution. Therefore, there can be SSD events that are significantly more recent than the two rounds of WGD. Following this recent events, sequence evolution had relatively less chances to modify and rewire the gene interactions involving the resulting paralogues.

Therefore, in order to make a sensible comparison between SSD and WGD couples, it is necessary to rule out possible confounding effects due to the different ages of paralogues. Such effects are indeed present, as we show in detail in fig. S2 in the *Supplementary Information*. The duplication age of paralog gene couples was estimated by considering their most recent common ancestor. Specifically, we considered SSD couples whose most recent common ancestor is older than *Sarcopterygii* as roughly contemporary to WGD couples. This approach is in line with a previous suggestion (14) and indeed the estimated ages are compatible, as shown in the *Supplementary Information* (fig. S1). With these criteria,

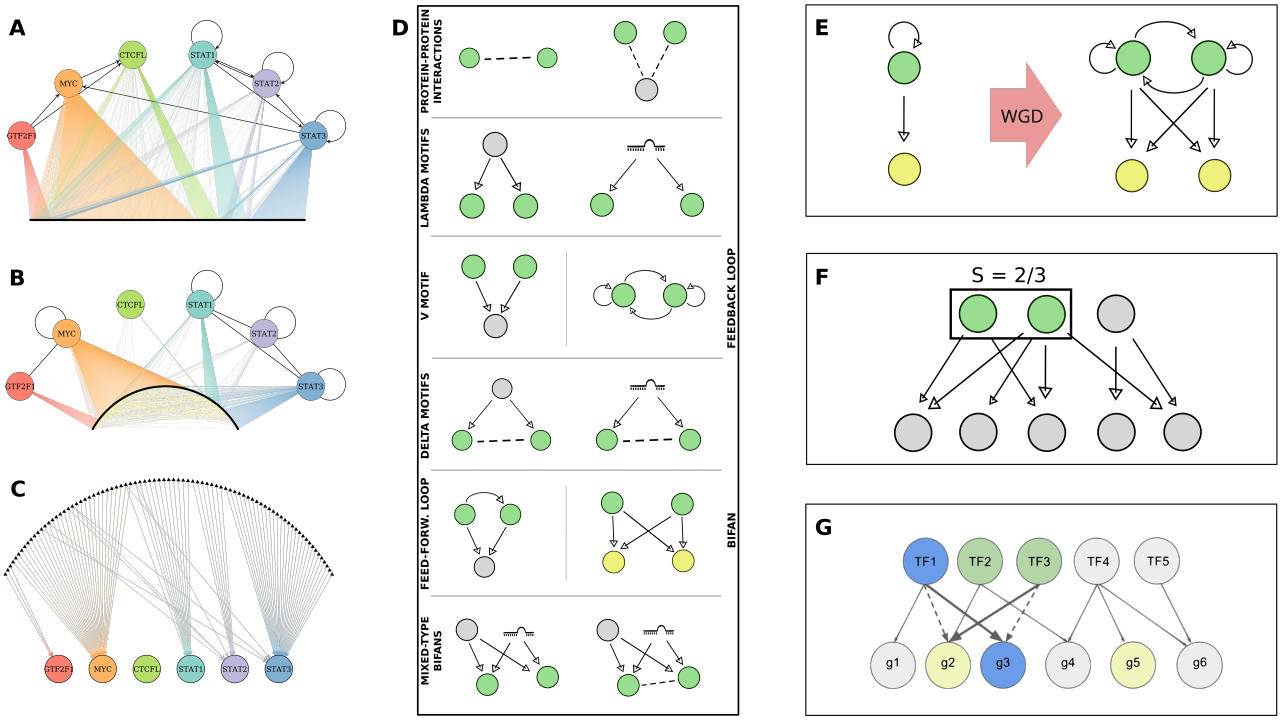


Figure 1. Regulatory Networks and Network Motifs. Interactions involving an illustrative subset of TFs are shown on the left for each of the regulatory mechanisms studied in the present work, i.e. for **(A)** transcriptional regulations from the ENCODE network, **(B)** protein-protein interactions in the PrePPI network, and **(C)** miRNA-gene regulatory interactions in the TarBase network. TFs are represented as colored circles, target genes as small black dots (here appearing as a thick black lines due to large number of genes), and miRNAs as black triangles. Black lines indicate interactions between TFs, while other interactions have the color of the involved TF. Yellow lines are interactions between non-TFs. **(D)** An overview of the network motifs that will be considered in the following. Gray circles represent generic genes, same-color circles (green or yellow) are paralogues, and miRNAs are represented in a stylized form. Solid arrows represent regulatory interactions, while dashed lines represent protein-protein interactions. **(E)** Illustration of how a WGD event can easily create FFLs and Bifans by duplicating the components of a simple regulatory interaction in which the regulator also has self-regulation. Many of the created interactions will then be lost during the evolutionary process, leaving only those that are not negatively selected . **(F)** Example of the structure of a Dense Overlapping Regulon (DOR) embedded in a gene regulatory network, with the target similarity S calculated for an illustrative couple. **(G)** Graphical representation of the degree-preserving procedure used to generate the null models: the dashed links are randomly chosen and their ends swapped, thus generating the new bold links. Note that all of the involved genes maintain their in and out degree in the process.

we identified 8663 young SSD duplicates (comprising 3442 genes) which we excluded from the comparison, and a final list of 13,618 SSD genes organized in 122,889 gene couples that we can safely use for a comparison with WGD genes couples.

Transcriptional Regulatory Network

We used the human transcriptional regulatory network presented in (29), a portion of which is displayed in fig. 1A. The network was obtained by the curation of data from ChIP-seq experiments by the ENCODE project, so we will be referring to it in the following as the "ENCODE network". We combined the information regarding proximal and distal regulation into a single regulatory network, with 122 transcription factors (TFs) and 9986 target genes. ChIP-seq based transcriptional networks should present the least amount of biases for the kind of analysis we are interested in, which essentially focuses on duplicated genes and network motifs. In fact, there are essentially three other methods to construct transcriptional regulatory networks besides Chip-seq derived networks (see for instance (30) for a recent review). Literature-based collections (such as TRRUST (31) or HTRIdb (32)) are by definition biased towards genes that received more attention from the scientific community. As pointed out in the *Introduction*, WGD-derived genes were shown to be often associated with diseases and organismal complexity, which are preferential subjects of published papers. Another possible approach is based on *in silico* predictions of the interactions from TF binding sequences. However, many of the duplicated TFs (especially the recent ones) can still present very similar binding sequences. Therefore, a network constructed in this way would lead to an artificially strong enrichment of some motifs (e.g., V motifs, shown in fig. 1D). Finally, methods based on reverse engineering gene expression data, such as the popular ARACNE (33), involve a pruning step that leads to an artificial decrease of the network clustering coefficient, and thus to an alteration in the statistics of three-node motifs.

The Protein-Protein Interaction Networks

We extracted the protein-protein interaction (PPI) network from the PrePPI database (34) and the STRING database (35). We downloaded the high-confidence predictions from the PrePPI database, selecting only the experimentally validated interactions, and updated the gene identifiers. The result is

a network of 15,762 genes and 237,272 PPIs. From the STRING database, we selected interactions that were both experimentally validated and with high confidence score (interaction score $>.700$, a parameter pre-set by the authors), in order to enforce stringency and to have a network size comparable with the size of the PrePPI network. We ended up with a STRING PPI network with 10,725 genes and 108,129 PPIs. There is a large overlap in the nodes present in the two networks (10,087 genes are in common) but a much lower overlap in the interactions (only 36,863 interactions are present in both networks). We will present in the main text the results obtained with the PrePPI network (a portion of which is shown in fig. 1B). However, all of the results are confirmed by analysis of the STRING network (see *Supplementary Information*, fig. S4 and S5), thus proving the robustness of our results.

The miRNA-gene Interaction Networks

The miRNA-target interaction networks we considered come from the TarBase database (36) and the mirDIP database (37). The TarBase network was constructed by selecting all the interactions coming from normal (non-cancer) cell lines or tissues, with positive evidence for a direct interaction between the miRNA and the target gene. This leaves us with 913 miRNAs regulating 10,497 genes, with 89,736 interactions. The mirDIP database integrates instead miRNA-target predictions coming from different databases and prediction methods, combining the different database-specific scores into a unified integrative score. Since no specific method is provided in order to choose an integrative score threshold, we chose to keep the 90,000 top-scoring interactions. Such a stringent threshold allows us to make a sensible comparisons with the TarBase network. The resulting mirDIP network has 513 miRNAs and 7965 genes with 89,991 interactions. As for the PPI networks, the overlap between the nodes is very high (406 miRNAs and 6241 genes are in common), but the overlap in edges is pretty low (only 9320 interactions are found in both networks). In the rest of the paper, results obtained with the TarBase network will be shown (represented in fig. 1C). The analogous results obtained with mirDIP network are available in the *Supplementary Information* (fig. S6). Again, the trends we find are robust despite the low overlap between the two networks.

Network Motifs

Networks motifs are combinations of nodes and regulatory interactions which are statistically over-represented in the regulatory network, with respect to an ensemble of null network models. They were shown to perform elementary regulatory functions (25) and the common lore is that some motifs were positively selected for by evolution precisely because of their ability to perform elementary computations. Such elementary modules can then be composed together to implement more complex regulatory functions in the regulatory network (38). This paper focuses on network motifs involving pairs of duplicated genes, as illustrated in fig. 1D.

Two duplicated transcription factors may regulate the same target (or set of targets) without interactions between the two duplicated genes, in a configuration we refer to as V motif. On the contrary, a couple of genes may be regulated by the same TF or by a common miRNA, giving rise to a Λ motif. We will explicitly distinguish between transcriptional and miRNA-mediated Λ motifs. If the duplicated genes involved in a Λ motif also interact at the protein level, we have a Δ motif, which again can be transcriptional or miRNA-mediated. The duplicated genes may be simultaneously involved in transcriptional and miRNA-mediated Λ or Δ motifs, hence resulting in mixed-type network motifs. More complex transcriptional motifs will also be analyzed, such as feed-forward loops (FFL) and feedback loops (FBL), including self-regulations and toggle-switch-like architectures. We will also consider Bifan motifs, where a couple of duplicates regulates another one but there are no interactions between the two regulators, and FFL+Bifan motifs, which have the additional regulatory interaction between one regulator and the other. Finally, we will also quantify the effects of the different types of duplications on the structure of the PPI network.

Motif enrichment and Z-score. The standard way to measure network motif enrichment is by reporting the Z-score associated with the motif counts. The Z-score is calculated in the following way:

$$Z = \frac{n - \bar{n}_{null}}{\sigma_{null}}$$

where n is the motif count in the real data, $\bar{n}_{null}$ and σ_{null} are the mean value and the standard deviation of the motif count distribution in the null model. Z-scores are considered to be significant when their absolute value is larger than ~ 5. We generated 100 realizations per each of the random models that are defined in a following section.

Regulatory redundancy and Similarity coefficient

As a measure of the interaction similarity between two duplicated genes, we used the Sorensen-Dice Similarity coefficient, defined in the following way for two sets A and B:

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|}.$$

In our case, A and B are the sets of interactions (regulators, targets or PPI depending on the task at hand) of two different genes a and b. This measure ranges from 0, when the two genes have no common interactions, to 1, when two genes share all of their interactions. Note that this similarity score is more general than motif enrichment, since we only take into account interactions common to both genes in a couple of paralogues and do not restrict in any way the connectivity between them. In some cases, for example for mixed-type motifs, the definition and interpretation of a similarity score is not straightforward and we resort to the Z-scores to gain more clear insights on the contribution of gene duplication. A more in-depth discussion on the differences between the similarity scores and the motif Z-scores can be found in the corresponding *Results* section, while a simple graphical example of the similarity between two regulators is shown in fig. 1F.

The statistical significance of the comparison between the similarity distributions of two different categories of gene couples is assessed by means of a two-tail Mann-Whitney U Test, with its associated P-value. The P-values of the comparisons between the real distributions and the null models are reported directly in the figures. If a comparison between two distributions is statistically significant ($P < 0.01$) we show in the figures the following symbols: $*$ for SSD-WGD comparison, ■ for WGD-NOT DUPLICATED comparison and ▲ for SSD-NOT DUPLICATED comparison. Note that when the symbol is reported, the P-value is typically much lower than the 0.01 threshold, and usually we have at least $P < 1e-5$.

Similarity score vs. Z-scores

It is worth noting that the motif enrichment Z-score and the similarity score distributions do not convey the same information. The Z-score counts the overall number of times we encounter a motif in the network, thus generically measuring the contribution of a type of duplicate to the non-random local structure of the whole network and the tendency to retain a specific motif when it is created in the network, either by chance or by other mechanisms (such as gene duplications). It does not, however, convey any information regarding the way in which motifs are distributed among different couples of duplicates, which is instead captured by our similarity measure. This is a very important statistic for our purposes, since we can interpret the similarity score of a duplicate couple as a proxy of the evolutionary constraints that act on it. In fact, higher similarity implies that a stronger evolutionary pressure is preventing the duplicated genes from changing their interactions, and thus their role in the regulatory network. Note that, in principle, the same kind of effect can derive from the duplication age of the paralogues - younger paralogues did not have enough time to lose or rewire connections and thus share more interactions than older ones. This effect is indeed present and shown in fig S2 of the *Supplementary information*. We mitigated this kind of bias by considering only SSD couples that were duplicated approximately in the same distant time when also the two rounds of WGD took place, as explained above.

Null models

We evaluated the motif enrichment by suitably rewiring the regulatory and protein interaction networks. More precisely we constructed randomized versions of the networks using the *degree-preserving* procedure proposed in (39) and illustrated in fig. 1G. This randomization algorithm destroys the local topology of the network but leaves the node degree intact, so that each gene retains the same number of interactions as in the real network, only with different neighbors. In this way we can rule out the possibility that the enrichment patterns we observe are only due to degree-degree correlations in the paralogues, since these correlations are kept unaltered in the ensemble of randomized networks. This is a standard procedure and has also been implemented in widely used motif counting software packages (24, 40).

If the motif under study involves interactions of different types, e.g. transcriptional and protein-protein interactions, we constructed several null models, each one with a randomized version of a different network while keeping the others fixed. Since this work is mainly focused on the effects of duplications at the transcriptional level, we report in the main text only the Z-scores referred to the randomizations of the ENCODE transcriptional regulatory network for mixed-type motifs. The complete results can be found in fig. S7 in the *Supplementary Information*.

We also compare the results about interaction similarities of the paralogues with interaction similarities of random non-duplicated gene couples (labelled as "not DUP" in the figures), in order to highlight the role of duplication mechanisms in shaping the network structure.

RESULTS

The following sections present the results of our motif enrichment analyses in order of increasing topological and functional complexity of the circuits considered.

Degree distributions

In network theory, the *degree* of a node, which in our case represents a gene, is the number of interactions it has with other nodes in the network. For directed networks, such as transcriptional networks, one can further distinguish between the *in-degree* of a node, i.e., the number of incoming links, and the *out-degree*, i.e., number of outgoing links. The degree distributions of the different networks considered are shown in fig. 2. The degree distributions and the average degree of genes duplicated by SSD and WGD do not display any striking difference with respect to the global degree distributions. Therefore, duplications do not display specific biases in terms of gene degree in the different networks considered. This is a relevant preliminary observation, since in the following we will focus on regulatory circuits whose statistics could be dependent on the degree of the nodes.

Duplicated genes often interact at the protein level

The first question we address is about the tendency of duplicated genes to interact at the protein level. The PPI network (see the *Methods* section) is very sparse, with 15,762 nodes and only 237,272 links. In this network, we identified 65,057 SSD pairs and 6,182 WGD pairs. Among these duplicated genes, approximately 4% of SSD pairs and (17% of WGD pairs show evidence of a protein-protein interaction in the PPI database. Such percentages, shown in fig. 3, are remarkably high. In the null models used for comparison the proportion of duplicates with an interaction never exceeds 1% and it is usually much lower. This leads to the impressive Z-scores reported in the figure. This behavior is also in stark contrast with the $\sim 0.2\%$ of couples of non-duplicated genes with a protein-protein interaction. Overall, we observe a strong correlation between presence of links in the PPI network and the pairing organization of duplicated genes. In other words, duplicated genes have a high probability of interaction in the PPI network. This effect is more pronounced for WGD duplications with almost 1 in 5 couples presenting a protein-protein interaction, compared to just 1 in 25 in the SSD case.

We also analyzed the tendency of couples of duplicated genes to form protein complexes with a third common protein,

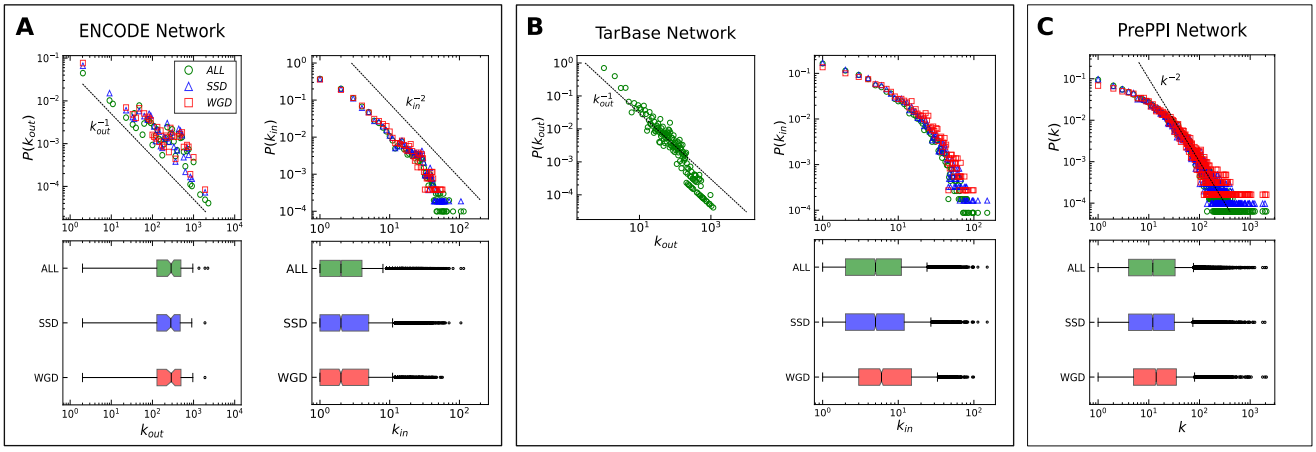


Figure 2. Degree distributions. Indegree (k_{in}) and outdegree (k_{out}) distributions of **(A)** the ENCODE transcriptional regulatory network and of **(B)** the TarBase miRNA-gene regulatory interactions network, and the degree (k) distribution of **(C)** the PrePPI protein-protein interactions network. Each degree distribution is shown both as a probability distribution (upper figure) and as a boxplot (lower figure). The global degree distribution of each network is represented in green, while the degree distributions of genes involved in a SSD couple and in a WGD couple are represented in blue and red, respectively. Dotted lines, corresponding to the reported scaling of the degree, are not the result of a fit and are shown as a reference only.

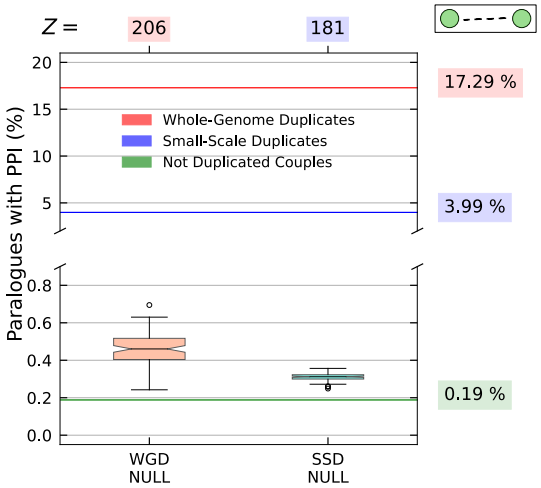


Figure 3. Interactions of duplicated genes at the protein level. The percentages of gene pairs that present an interaction in the PrePPI database are indicated by the bold horizontal lines and explicitly stated in the labels on the right. The null model distributions are reported in the boxplots and the corresponding Z-scores are shown at the top.

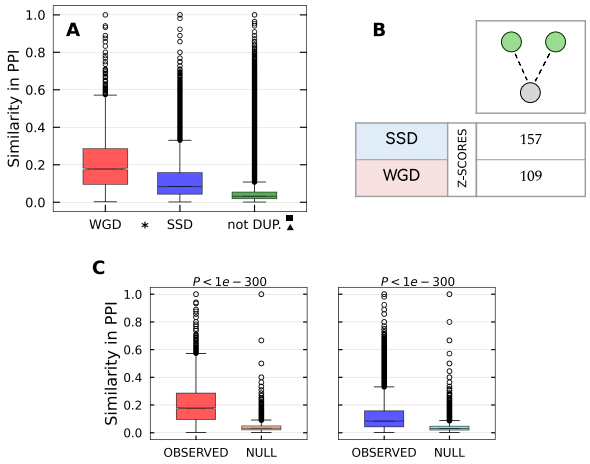


Figure 4. Pairs of duplicated genes interacting with a third protein. (A) Similarity distributions for WGD, SSD and not duplicated gene couples in the PrePPI network. All of the pairwise comparisons between distributions are statistically significant, as indicated by the presence of the symbols explained in the *Methods* section. **(B)** Z-scores measuring the enrichment of the co-interaction motif with respect to the null model. **(C)** Pairwise comparison between each real similarity distribution and the null distribution for the respective duplication type.

which is captured by the statistics of co-interaction motifs presented in fig. 4. In particular, fig. 4A shows that WGD couples have a higher interaction similarity with respect to SSD couples and, generally, duplicates have a significantly larger proportion of common interactions than non-duplicated couples. This is confirmed by the comparison with the null model obtained by rewiring the PPI network, as discussed in the *Methods* section (fig. 4C). This tendency explains the enrichment of co-interaction motifs shown by the Z-scores in fig. 4B.

The evolutionary tendency to retain WGD couples that participate in common protein complexes agrees with previous observations in yeast (22, 41), where the observed tendency was less significant but exactly in the same direction. This result also agrees with a previous observation that proteins belonging to protein complexes were retained more frequently after WGD events than SSD events (42). The same trend was reported for the human genome using a database of transient protein complexes (18).

V motifs are enriched of WGD Transcription Factors

Transcriptional V motifs are genetic circuits in which a couple of duplicated transcription factors regulate a common

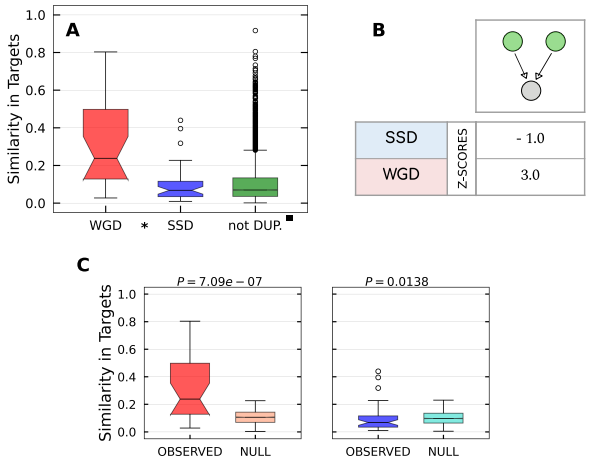


Figure 5. Transcriptional V motifs. **(A)** Similarity distributions for WGD, SSD and not duplicated TF couples in the ENCODE network. As indicated by the presence of the symbols explained in the *Methods* section, the difference between SSD and not-duplicated distributions is not statistically significant while the comparisons involving the WGD distribution are instead significant. **(B)** Z-scores measuring the enrichment of the V motif with respect to the null model. **(C)** Pairwise comparison between each real similarity distribution and the null distribution for the respective duplication type.

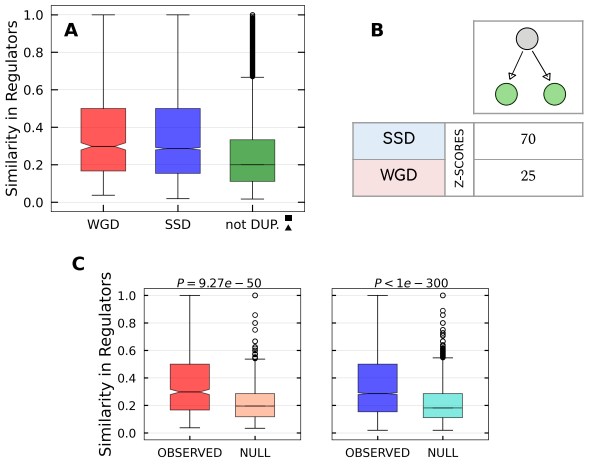


Figure 6. Transcriptional Λ motifs. **(A)** Similarity distributions for WGD, SSD and not duplicated target genes couples in the ENCODE network. As indicated by the presence of the symbols explained in the *Methods* section, the difference between SSD and WGD distributions is not statistically significant, while both of them are significantly greater that the similarity distribution of non duplicated genes. **(B)** Z-scores measuring the enrichment of the Λ motif with respect to the null model. **(C)** Pairwise comparison between each real similarity distribution and the null distribution for the respective duplication type.

target gene. The motif enrichment analysis and the similarity distributions indicate that WGD pairs of TFs tend to co-regulate the same target genes more than SSD pairs, whose behaviour is instead comparable with that of non duplicated TF couples (fig. 5A). Since the number of duplicated TFs (both through WGD and SSD events) is rather small, motif enrichment analysis and similarity scores are expected to show larger fluctuations and smaller Z values. However, fig. 5 shows that the result are still consistent. These findings indicate that WGD had a crucial role in shaping the transcriptional regulatory mechanisms, by introducing regulatory redundancies that were retained by evolution over millions of years. On the other hand, regulatory redundancies created by SSD duplications have been generally lost or rewired during evolution. A similar phenomenon was observed in yeast (26), and thus seems to be an universal trend characterizing WGD-derived genes.

The different behavior of WGD and SSD derived couples is corroborated by the observation that WGD pairs of TFs tend to maintain the same DNA Binding Sequence (DBS) much more than SSD pairs. In fact, out of the 25 pairs of WGD TFs, 20 (i.e 80%) kept the same DBS (more precisely they belong to the same motif family, as defined in (43)), while in the SSD case this happens only for 7 out of 41 TFs pairs. The specific conservation of DBS in WGD pairs was observed also in yeast (44), thus suggesting that it could be a general phenomenon.

Λ motifs are enriched in duplicated targets

Λ motifs are simple circuits in which a regulator acts on a couple of targets. We considered transcriptional and miRNA-mediated Λ motifs as reported in fig. 6 and fig. 7 respectively. The similarity distributions of WGD and SSD genes are both larger than the non-duplicate one for both types of Λ motifs.

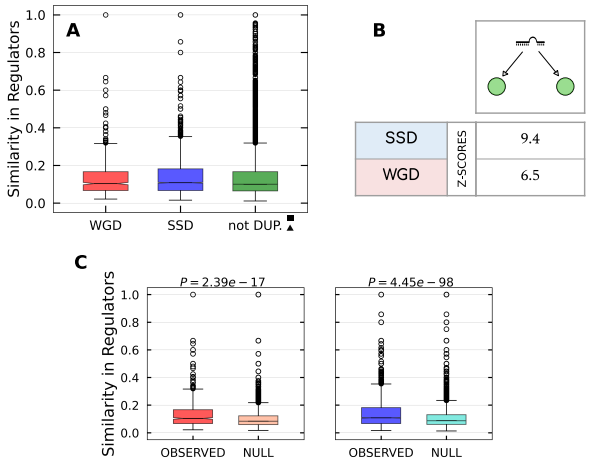


Figure 7. miRNA Λ motifs. **(A)** Similarity distributions for WGD, SSD and not duplicated target genes couples in the TarBase network. As indicated by the presence of the symbols explained in the *Methods* section, the difference between SSD and WGD distributions is not statistically significant, while both of them are significantly greater that the similarity distribution of non duplicated genes. **(B)** Z-scores measuring the enrichment of the Λ motif with respect to the null model. **(C)** Pairwise comparison between each real similarity distribution and the null distribution for the respective duplication type.

Coherently, the Z-scores indicate enrichment for both SSD and WGD motifs. The Z values suggest that motifs derived from SSD have been retained with higher significance with respect to WGD ones. The same trend is present in miRNA-mediated motifs, but with lower enrichment scores. Overall we observe a tendency of duplicated couples to share the

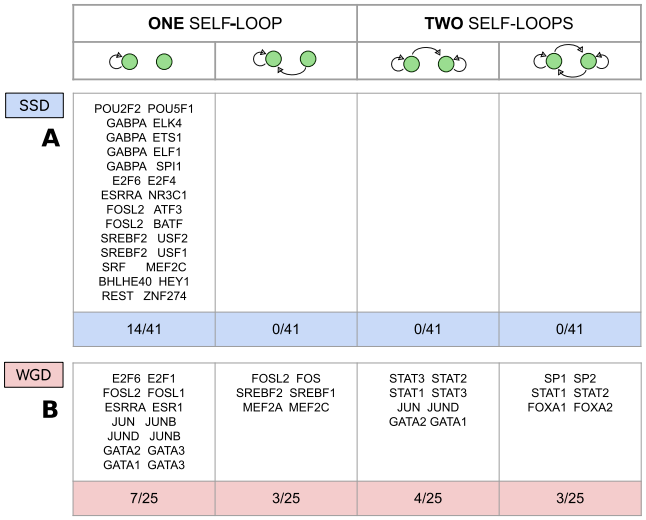


Figure 8. Feedback Loops and Self-Loops in couples of duplicated Transcription Factors. (A) SSD and **(B)** WGD duplicate TF couples that contain at least one gene with a self-loop or that display a mutual regulatory interaction in the ENCODE regulatory network, subdivided by equal topological arrangements.

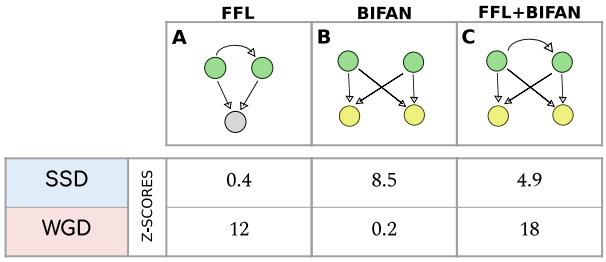


Figure 9. Transcriptional FFL, Bifan and FFL+Bifan motifs. (A) Transcriptional Feed-Forward Loops (FFLs). **(B)** Transcriptional Bifan motifs (in which no regulatory is present between the two TFs). **(C)** FLL+Bifan motif. In both **(B)** and **(C)** the two regulators and the two targets are duplicated couples of the same type (i.e. both WGD or both SSD pairs).

same regulatory interactions. The pattern is more evident at the trascriptional level, and it is stronger for SSD than for WGD pairs.

More complex motifs are enriched in duplicated genes.

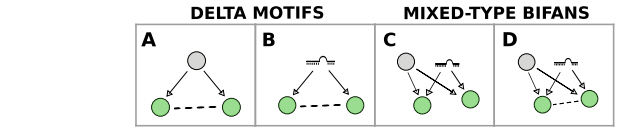
The role played by WGD-derived genes in shaping the regulatory network emerges more clearly looking at more complex network motifs such as Feed-Back Loops (FBLs), Feed-Forward Loops (FFLs) and BiFan-type motifs (fig. 8, and 9). These motifs were all shown to be associated to relevant specific functions that will be discussed in the corresponding sections.

FBLs involving pairs of WGD TFs are predominant. Feedback Loops (FBLs) are a key component of regulatory networks, since they can implement bi-stable switches (25) that represent an excellent decision-making circuit. FBLs can be easily created by duplicating a TF with a self-regulating loop and self regulation is a widespread network motif, from bacteria to humans (25). This simple motif is associated to several important functions, such as the the modulation of the expression response time, robustness to stochastic noise, and bimodality in the protein levels (25). In our analysis, the number of observed FBLs is so small that statistical tests are not meaningful, thus we simply categorised the 25 pairs of WGD TFs and the 41 pairs of SSD TFs according to their topological configuration. Fig. 8 reports the duplicated TF couples that contain at least one gene with a self-loop or that display a mutual regulatory interaction. We immediately see that FBLs involving SSD pairs are completely absent in the network, while 3 out of the 25 pairs of WGD TFs present in the network display a FBL topology and, interestingly, all 3 pairs involved in a FBL motif also present two self-loops. In general, the data presented in fig. 8 show that it is more likely for a pair of WGD-derived TFs to retain a self-regulatory mechanism, together with some kind of mutual regulatory interaction. These observations suggest that the evolutionary pressure favoured the retention of new FBLs created during the two WGD rounds while disfavouring the retention of those created by a SSD event.

FFLs involving pairs of WGD genes are strongly enriched in the regulatory network. Feed-Forward Loops (FFLs) are another fundamental component of gene regulatory networks and are often strongly enriched in regulatory networks (25). Depending on the exact nature and strength of the interactions, they can implement complex functions such as detection of signal persistence, pulse generation, noise buffering and fold-change detection (25).

Fig. 9A shows that FFL motifs generated by WGD events are strongly conserved, while the statistics of FFLs involving SSD TFs is compatible with the null model. Once again this clearly shows that evolutionary constraints applied to WGD genes are very different from the ones that affect SSD couples.

Gene duplications shaped Bifan and FFL arrays. Bifan and FFL+Bifan motifs (also called “Multi-output Feed-Forward Loops” in the literature) are shown in fig. 9B and 9C respectively. The main function of these motifs is to integrate different input signals, in order to organize the transcription of downstream target genes. They can both be seen as combinatorial decision-making devices, but with an important difference: the additional presence of a regulatory interaction between the two TFs in the second case transforms a simple Bifan into a double FFL, which allows to combine the input signals in a nonlinear fashion, leading to more complex regulatory programs. Another peculiarity of Bifan motifs is their tendency to cluster together, forming extensive superstructures named “Bifan arrays” (44) or “Dense Overlapping Regulons” (DORs) (25), that were identified for the first time in E. Coli (45). In such superstructures, regulators and targets are arranged on two different layers, with a very large number of regulatory interactions between them. The situation is similar to the one depicted in fig. 1F and 1G, but in real regulatory networks Bifan arrays can involve dozens of genes. The additional presence of regulatory interactions among regulators further



	Z-SCORES	A	B	C	D
SSD		21	4.6	17	3.6
WGD		12	4.6	1.0	6.9

Figure 10. Motifs with mixed-type regulatory interactions. (A) Transcriptional Δ motifs. **(B)** miRNA-mediated Δ motifs. **(C)** Mixed Bifan motifs, in which a pair of target genes are regulated both by a common TF and a common miRNA. **(D)** Mixed Bifan motif in which the two target genes interact at the protein level. The reported Z-scores are referred to the null model obtained by randomizing the transcriptional regulatory network (apart from the miRNA Δ motif for which the miRNA-gene network is randomized).

increases the complexity of the functions that can be implemented.

We consider the special case where both the regulators and the targets are two - different - duplicated couples, along with motifs that do not contain any duplicated couple. Their levels of enrichment in the ENCODE transcriptional network are shown in fig. 9B for simple Bifans and in 9C for the FFL+Bifan configuration.

The relevance of these two motifs in the structure of the regulatory network is confirmed by their statistical enrichment. In particular, simple Bifans are retained with higher probability when they are created by SSD duplications, while WGD pairs are preferentially involved in FFL+Bifan motifs. This result again confirms that WGD-derived genes are subjected to different evolutionary constraints with respect to SSD-derived genes, and that WGD has driven the formation of motif that are associated to more complex functions.

Synergy between different layers of regulation is facilitated by duplication events.

By analysing different layers of regulation combined together, we can quantify the role of duplication events in fostering the synergy between different regulation layers. For example, considering Δ motifs we can assess the tendency of a particular type of regulators to act on a couple of duplicated genes that also interact at the protein level (fig. 10A). We observe a strong enrichment of both SSD and WGD motifs, with a slight preference for the former type, which is in line with the results reported in the section on Λ motifs. In the case of miRNA-mediated Δ motifs (fig. 10B), we again observe a clear role of duplicated genes in their retention but there are no clear preferences for SSD or WGD genes.

The enrichment analysis for the mixed-type Bifans in absence of protein-protein interactions, i.e., the motif observed when a duplicated pair is simultaneously involved in a transcriptional and miRNA-mediated Λ motif, are reported in fig. 10C. The enrichment of mixed-type Bifans with additional protein-protein interactions between the duplicated genes, is instead shown in fig. 10D. Different types of duplicates appear to promote different integration strategies between layers of regulation. SSD couples are strongly associated with integration between miRNA and transcriptional regulators, when there is no direct PPI interaction between the targets. On the other hand, WGD couples promote the retention also of a direct PPI link between them. This clearly shows that gene duplications facilitate the creation of a significant three-way synergy among the three layers of regulation. This effect can in principle lead to more complex and robust regulatory mechanisms. In fact, the combination of miRNA-mediated and transcriptional regulatory interactions has been shown to ensure optimal noise control, together with a set of interesting complex properties like adaptation and fold-change detection, depending on the parameters of the regulatory interactions (46, 47).

DISCUSSION

Target redundancy and dosage balance

The exact mechanisms involved in the retention of duplicated genes are still debated, but most proposed explanations focus on dosage balance constraints (48, 49, 50). For example, a recent analysis of genetic interactions involving WGD couples in yeast proposed that evolutionary trajectories of duplicated genes are dictated by the combination of dosage balance constraints with functional and structural entanglement factors (51). The dosage balance explanation relies on the importance of keeping the correct stoichiometric ratios of protein products within the cell. If the balance is preserved by the duplication event, the duplicated genes will be conserved by evolution with higher probability. This scenario was first proposed to explain the retention of WGD duplicates, since the duplication of the whole genome facilitates an overall balancing of gene expression (50).

However, the same principle was recently invoked to explain SSD retention as well (52). In this case, dosage balance (and thus duplicate retention) is granted by a substantial decrease in gene expression of the duplicated pair, which allows to re-balance gene dosage after duplication. Examples of this behaviour have been found both in yeast and in mammals (52). The decrease in expression levels needed for dosage balance could be achieved more easily if both duplicated genes were regulated by the same set of TFs, possibly the same TFs which regulated the ancestral gene (52). The presence of an evolutionary pressure to keep co-regulation of duplicated targets is also supported by recent observations: duplicated gene pairs are enriched for co-localization in the same Topologically Active Domain (TAD), share more enhancer elements than expected, and have increased contact frequencies in Hi-C experiments (53). From a regulatory network perspective, this evolutionary pressure would imply the selective enrichment we observe of the transcriptional Λ motifs stemming from duplicated targets.

However, this is not the only reason for which one could expect an over-representation of the Λ motif. Motifs of this type ensure a reduction of the relative fluctuations of the two targets (47) and improve the stochastic stability of the duplicated genes. This noise buffering action is particularly effective in presence of a combined and coordinated action of transcription factors and miRNAs (46, 47), i.e., in presence of a “mixed”-type network motifs. All of these considerations are indeed confirmed by the findings presented in fig. 6, 7 and 10C.

Dosage balance constraints and stochastic stability are particularly important if the two duplicated proteins are in interaction between them or are involved in a complex (54). If this is the case, we should expect a specific enrichment of protein-protein interactions between the two duplicated genes and of Δ motifs. These effects are indeed observed in our analysis (fig. 3, 4 and 10).

The tendency to interact and to share interacting proteins is even more evident for WGD-derived gene couples. This could be again a consequence of how the two different mechanisms of duplication alter the dosage balance (17).

Regulatory redundancy

It is widely recognized that gene duplications played a central role in the evolution of gene regulatory networks (38, 55) and in setting the TF repertoire (43).

An immediate consequence of TF duplication is the creation of a regulatory redundancy, meaning that after the duplication event the two TFs regulate the same set of target genes. However, this potential functional redundancy is expected to be transient. In fact, during evolution one gene copy may be lost or become a pseudogene, it may acquire a new function (neofunctionalization) (1), or it may share the ancestral functions of the original gene with the other copy (subfunctionalization) (56). The typical completion time for these processes is of a few millions of years (57), thus for most of the SSD and for all the WGD gene pairs, we should expect no functional redundancy at all. On the contrary, there are strong indications that this is not the case and that for several pairs of both SSD and WGD redundancy is preserved, in some cases, for billions of years (58).

Our study suggests that the retention of regulatory redundancy is strongly dependent on the duplication mechanism. The topological enrichment of V motifs an the distribution of target similarity (fig. 5) suggest a significative preference for WGD TF pairs to retain common targets. SSD couples display instead a weak similarity in targets, compatible with null models. Therefore, WGD events seem to have promoted regulatory redundancy during network evolution. Interestingly, there is a non-trivial relation between redundancy in the interactions of the transcription factor repertoire and organismal complexity (43). This associates once again WGD events to an increased complexity.

There are several possible paths that connect genetic regulatory redundancy with complexity. First of all, regulatory redundancy can increase the robustness against mutations (59), which is a safety mechanism that is more and more necessary as the interplay of regulatory interactions increase in complexity. Moreover, regulatory redundancy facilitates the implementation of articulated combinatorial regulations. In many cases two duplicated TFs could keep the same set of target genes, but evolve to respond to different cellular signals or to interact with different upstream proteins (2, 60).

In principle, combinatorial regulation - and the associated benefit of an increased environmental responsivity - could evolve by combining the regulations of two TFs, with no need for specifically retaining duplicated TFs. However, such a mechanism would unavoidably increase the noise in the regulatory process. There is indeed a tension between environmental responsivity and noise control in gene regulation, and it has been suggested that it could be resolved by gene duplications (61, 62). This hypothesis was tested in yeast for the specific Msn2-Msn4 pair of WGD-derived Transcription Factors (61), and our results suggest that it could be a general evolutionary trend that applies also to gene regulation in vertebrates.

Most of the results mentioned above on duplication mechanisms are based on observations and experiments performed in simple model organisms like S. cerevisiae and A. thaliana. The new data on vertebrate WGD genes give us the unique opportunity to extend previous studies in a more complex setting. We observed that several trends are conserved across different species and overall seems that ancient WGD events had a relevant role in shaping current regulatory redundancy.

FFL and Bifan arrays

The specific combination of FFL+Bifan arrays that, we found, is promoted by WGD-derived genes can have important consequences on the network dynamics. By combining the combinatorics of Bifan with the nonlinear signal integration of FFLs, these circuits can process signals in a highly non-trivial way. As fig. 1E shows, WGD events can create FFL+Bifan motifs in a very easy and natural way. Duplication of a TF with a self-loop interaction generates a couple of TF paralogues with a mutual regulatory interaction and a commmon set of targets. If the original regulator does not have a self-regulatory interaction, the WGD event creates a simple Bifan motif instead. In principle, the same circuits can be generated by a succession of SSD events: the chances of duplicating a TF and its target in two distinct SSD events is reasonably low, but SSD events occur quite frequently. However, there is no guarantee that the created motif will survive. In a relatively short evolutionary timescale many of the created connections could be rewired and duplicated genes could be lost. Therefore, the presence of complex structures retained for more than 500 millions of years is non-trivial and imputable to selective pressure. Interestingly, fig. 9 shows that there are specific retention biases for different circuits depending on the the duplication mechanism at the origin of their formation. Our findings suggest that SSD duplications favoured the formation and retention of the less complex Bifan motif, while WGD duplications are associated to more complex FFL arrays. A similar retention pattern (over-representation of Bifan motifs for duplicated TFs and in particular for WGD versus SSD pairs) was also observed in yeast (44).

These observations again support the conjecture that WGD-derived genes follow a different evolutionary trajectory with respect to SSD ones, and that their emergence favoured the development of complex regulatory strategies.

Synergy of different layers of regulation

Besides the vertebrates' WGDs, there are other well known examples of WGD events in eukaryotes, such as those observed in *S. cerevisiae* (26, 44) and in *A. thaliana* (9). Several of the trends we identified in human are in agreement with previous analysis in those two model organisms, suggesting some universality of the results despite the increase in organism complexity. This increase in complexity is

also linked to the presence of several post-transcriptionl layers of regulation, such as miRNA regulation, that are much less developed in simpler organisms such as yeast. Analyzing the human regulatory network, we could identify an important role of gene duplication events in promoting the interplay between different layers of regulation. Specifically, we identified an emergent statistical enrichment of motifs involving both protein-protein interactions and trascriptional regulation, as well as motifs combining transcriptional and post-transcriptional regulation. This agrees with the general observation that complex regulatory functions like adaptation, fine tuning, fold change detection or noise buffering can be better achieved by suitable combinations of miRNAs and TFs, arranged in well defined network motifs (46, 47, 63). Our analysis indicates that several of these mixed motifs arose with ancient gene duplication events - both SSD and WGD - at the beginning of the vertebrate lineage and were then conserved by evolution for more than 500 million years.

Robustness of the results

The nature of the motifs that we studied and the type of enrichment in which we are interested (WGD versus SSD, or pairs of duplicated genes versus non-duplicated ones) requires a careful control over possible spurious effects. The first necessary control is that the three gene classes do not differ significantly in the number of interactions they have since this could affect the motif statistics. The absence of this possible bias is tested in fig. 2.

To further assess the robustness of our analysis, we considered two alternative protein-protein interaction networks (the PrePPI and STRING-DB network) and two alternative miRNA-gene networks (the TarBase and the mirDIP network). Despite significant differences both in the genes and in the interactions in the different databases, we found consistent enrichment patterns (see the *Supplementary Information*).

CONCLUSIONS

Gene duplications played a crucial role in the evolution of the human genome, and it is by now widely accepted that two rounds of whole genome duplication happened at the origin of the vertebrate lineage (1). How these two global-scale events affected the gene regulatory networks is, however, still to be fully understood. Thanks to the recently published lists of WGD pairs (14, 16, 17), we had the possibility to tackle this problem. This paper quantifies the effects of WGD and SSD events on the structure of regulatory networks in human, and the results support the idea that these networks were significantly shaped by the two rounds of WGD at the beginning of the vertebrate lineage.

Our analysis of network motifs specifically indicates that the two rounds of WGD contributed substantially to the overall regulatory redundancy, promoted synergy between different regulatory layers, and typically generated motifs that can be associated with complex functions.

DATA AVAILABILITY

The raw data used for this study are all publicly available from their respective sources. The data and the code required to replicate the analyses and figures in this work are available on Zenodo with the following DOI: 10.5281/zenodo.5110112 Our processed lists of SSD and WGD paralogues and the processed regulatory networks are also easily downloadable from the following GitHub repository: `https://github.com/fmottes/wgd-network-motifs`.

FUNDING

This work was partially supported by the ”Departments of Excellence 2018–2022” Grant awarded by the Italian Ministry of Education, University and Research (MIUR) (L.232/2016)

ACKNOWLEDGMENTS

Conflict of interest statement. None declared.

REFERENCES

1. Ohno, S. (1970) Evolution by Gene Duplication, Springer-Verlag, Berlin Heidelberg.
2. Zhang, J. (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution,* **18**(6), 292–298.
3. Demuth, J. P. and Hahn, M. W. (2009) The life and death of gene families. *BioEssays,* **31**(1), 29–39.
4. Van de Peer, Y., Maere, S., and Meyer, A. (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.,* **10**(10), 725–732.
5. Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.,* **18**(7), 411–424.
6. Comai, L. (2005) The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.,* **6**(11), 836–846.
7. Wolfe, K. H. and Shields, D. C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature,* **387**(6634), 708–713.
8. Kellis, M., Birren, B. W., and Lander, E. S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature,* **428**(6983), 617–624.
9. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* **408**(6814), 796–815.
10. Dehal, P. and Boore, J. L. (2005) Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology,* **3**(10), e314.
11. 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**(7734), 64–70.
12. D'Antonio, M. and Ciccarelli, F. D. (2011) Modification of Gene Duplicability during the Evolution of Protein Interaction Network. *PLOS Comput. Biol.,* **7**(4), e1002029.
13. Nakatani, Y., Takeda, H., Kohara, Y., and Morishita, S. (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.,* **17**(9), 1254–1265.
14. Singh, P. P., Arora, J., and Isambert, H. (2015) Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLOS Comput. Biol.,* **11**(7), e1004394.
15. Acharya, D. and Ghosh, T. C. (2016) Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics,* **17**(1), 71.
16. Singh, P. P. and Isambert, H. (2020) OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res.,* **48**(D1), D724–D730.
17. 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**(20), 9270–9274.
18. Singh, P. P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., and Isambert, H. (2012) On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. *Cell Rep.,* **2**(5), 1387–1398.
19. Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.,* **7**(5), R43.
20. Freeling, M. and Thomas, B. C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.,* **16**(7), 805–814.
21. Huminiecki, L. and Heldin, C. H. (2010) 2R and remodeling of vertebrate signal transduction engine. *BMC Biology,* **8**(1), 146.
22. Guan, Y., Dunham, M. J., and Troyanskaya, O. G. (2007) Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. *Genetics,* **175**(2), 933–943.
23. Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.,* **102**(15), 5454–5459.
24. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science,* **298**(5594), 824–827.
25. Alon, U. (2019) An Introduction to Systems Biology: Design Principles of Biological Circuits, CRC Press, .
26. Fusco, D., Grassi, L., Bassetti, B., Caselle, M., and Cosentino Lagomarsino, M. (2010) Ordered structure of the transcription network inherited from the yeast whole-genome duplication. *BMC Syst. Biol.,* **4**, 77.
27. Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R., Yates, B., and Bruford, E. A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.,* **49**(D1), D939–D946.
28. Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., et al. (2021) Ensembl 2021. *Nucleic Acids Res.,* **49**(D1), D884–D891.
29. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature,* **489**(7414), 91–100.
30. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., and Saez-Rodriguez, J. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.,* **29**(8), 1363–1375.
31. Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C. Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.-N., Jung, H., Nam, S., Chung, M., Kim, J.-H., and Lee, I. (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.,* **46**(D1), D380–D386.
32. Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics,* **13**, 405.
33. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* **7 Suppl 1**, S7.
34. Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., and Honig, B. (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.,* **41**(Database issue), D828–833.
35. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Mering, C. (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.,* **47**(D1), D607–D613.
36. Karagkouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., Vergoulis, T., Dalamagas, T., and Hatzigeorgiou, A. G. (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.,* **46**(D1), D239–D245.
37. Tokar, T., Pastrello, C., Rossos, A. E. M., Abovsky, M., Hauschild, A.-C., Tsay, M., Lu, R., and Jurisica, I. (2018) mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Res.,* **46**(D1), D360–D370.
38. Teichmann, S. A. and Babu, M. M. (2004) Gene regulatory network growth by duplication. *Nat. Genet.,* **36**(5), 492–496.
39. Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science,* **296**(5569), 910–913.
40. Wernicke, S. and Rasche, F. (2006) FANMOD: a tool for fast network motif detection. *Bioinformatics,* **22**(9), 1152–1153.
41. Conant, G. C. and Wolfe, K. H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.,* **9**(12), 938–950.
42. Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007) All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.,* **8**(10), R209.
43. Rosanova, A., Colliva, A., Osella, M., and Caselle, M. (2017) Modelling the evolution of transcription factor binding preferences in complex eukaryotes. *Sci. Rep.,* **7**(1), 7596.
44. Ward, J. J. and Thornton, J. M. (2007) Evolutionary models for formation of network motifs and modularity in the Saccharomyces transcription factor network. *PLoS Comput. Biol.,* **3**(10), 1993–2002.
45. Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.,* **31**(1), 64–68.
46. Osella, M., Bosia, C., Corá, D., and Caselle, M. (2011) The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comput. Biol.,* **7**(3), e1001101.
47. Riba, A., Bosia, C., El Baroudi, M., Ollino, L., and Caselle, M. (2014) A combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes. *PLoS Comput. Biol.,* **10**(2), e1003490.

48. Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J. Mol. Evol.,* **49**(2), 169–181.
49. Qian, W., Liao, B.-Y., Chang, A. Y.-F., and Zhang, J. (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.,* **26**(10), 425–430.
50. Conant, G. C., Birchler, J. A., and Pires, J. C. (2014) Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.,* **19**, 91–98.
51. Kuzmin, E., VanderSluis, B., Nguyen Ba, A. N., Wang, W., Koch, E. N., Usaj, M., Khmelinskii, A., Usaj, M. M., van Leeuwen, J., Kraus, O., Tresenrider, A., Pryszlak, M., Hu, M.-C., Varriano, B., Costanzo, M., Knop, M., Moses, A., Myers, C. L., Andrews, B. J., and Boone, C. (2020) Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science,* **368**(6498), eaaz5667.
52. Lan, X. and Pritchard, J. K. (2016) Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science,* **352**(6288), 1009–1013.
53. 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**(1), 81–91.
54. Qian, W. and Zhang, J. (2008) Gene dosage and gene duplicability. *Genetics,* **179**(4), 2319–2324.
55. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.,* **14**(3), 283–291.
56. Lynch, M. and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. *Science,* **290**(5494), 1151–1155.
57. Lynch, M. and Conery, J. S. (2003) The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics,* **3**(1-4), 35–44.
58. Vavouri, T., Semple, J. I., and Lehner, B. (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.,* **24**(10), 485–488.
59. Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W.-H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature,* **421**(6918), 63–66.
60. Baker, C. R., Hanson-Smith, V., and Johnson, A. D. (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science,* **342**(6154), 104–108.
61. Chapal, M., Mintzer, S., Brodsky, S., Carmi, M., and Barkai, N. (2019) Resolving noise-control conflict by gene duplication. *PLoS Biol.,* **17**(11), e3000289.
62. Hallin, J. and Landry, C. R. (2019) Regulation plays a multifaceted role in the retention of gene duplicates. *PLoS Biol.,* **17**(11), e3000519.
63. Bosia, C., Osella, M., Baroudi, M. E., Corà, D., and Caselle, M. (2012) Gene autoregulation via intronic microRNAs and its functions. *BMC Syst. Biol.,* **6**, 131.