

**The influence of ancient Whole-Genome Duplications
on the evolution of human gene regulatory networks**

PhD Program in

Complex Systems for Life Sciences

XXXIV Cycle

Candidate:

Francesco Mottes

Supervisor:

Michele Caselle



UNIVERSITÀ DEGLI STUDI DI TORINO

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Prof. Michele Caselle. His outstanding scientific acumen and deep humanity have guided and inspired generations of students and scientists, I am deeply grateful to have been both. I am also particularly thankful to Dr. Matteo Osella, ultimately as much an additional supervisor as a friend. Looking back, I absorbed during our discussions much of his critical approach to problems, which I regard as one of the most important legacies of my PhD journey.

I am deeply grateful to my parents, Susanna and Gianni. Their discrete presence, not to mention their limitless support in any situation, is what gives me the strength to reach always a bit further into an uncertain future. And to my sister Lorenza, a noisy, cheerful and supportive presence which I could really not do without.

I am incredibly grateful to Maddalena who, for unknown reasons, still bears with me. We supported each other throughout this sometimes very challenging period of our lives, it is difficult to express with words the importance of your presence.

This thesis is also dedicated to a number of people whose presence helped me a great deal. Many others are not written explicitly, as it is often written, for brevity. I am sure they know who they are, though.

To Carla, Giorgio and Walter, no need for words.

To Fabio, my dearest friend from the beginning of my years in Torino, and an amazing flatmate.

To my *Amichetti* in Trento. Seeing you, even for a short time, always make me feel at home.

And to my former colleagues: Marta, Filippo, Silvia, David, Mattia, Elisa and Alberto (aka Il Nonno) for sharing with me part of this journey.

ABSTRACT

This thesis 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 multi-layer 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. We discuss more in detail the RAR/RXR pathway as an illustrative example of the evolutionary impact of WGD duplications in human and present extensive evidence of the robustness of our analyses.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
1 Introduction	1
1.1 Whole-Genome Duplications	1
1.2 Gene Regulatory Networks	3
1.2.1 Network Motifs	5
1.3 Motivation and outline of the work	6
2 Materials and Methods	9
2.1 Gene Nomenclature	9
2.2 Transcriptional Regulatory Network	9
2.3 Protein-Protein Interaction Networks	10
2.4 miRNA-gene Interaction Networks	11
2.5 Small-scale and Whole-genome duplicates	11
2.5.1 WGD paralogues	11
2.5.2 SSD paralogues	12
2.5.3 Age distributions of paralogues	12
2.5.4 Non-duplicated gene couples	14
2.6 Network Motif enrichment	14
2.7 Regulatory redundancy and Similarity coefficient	14
2.7.1 Similarity score vs. Z-scores	16
2.8 Null models	16
2.9 Data and code availability	18
3 Results	19
3.1 Degree distributions	19
3.2 Duplicated genes often interact at the protein level	21
3.2.1 Comparison between the PrePPI and STRING networks	23
3.3 V motifs are enriched of WGD Transcription Factors	23
3.4 Λ motifs are enriched in duplicated targets	24

3.4.1	Comparison between the TarBase and the mirDIP networks . . .	25
3.5	More complex motifs are enriched in duplicated genes.	26
3.5.1	FBLs involving pairs of WGD TFs are predominant.	27
3.5.2	FFLs involving pairs of WGD genes are strongly enriched in the regulatory network.	29
3.5.3	Gene duplications shaped Bifan and FFL arrays.	30
3.6	Synergy between different layers of regulation is facilitated by dupli- cation events.	31
4	Discussion	33
4.1	Target redundancy and dosage balance	33
4.2	Regulatory redundancy	34
4.3	FFL and Bifan arrays	36
4.4	Synergy of different layers of regulation	37
4.5	An example of WGD importance: The RAR/RXR pathway	37
4.6	Robustness of the results	39
4.6.1	Interaction similarity is influenced by duplication age	40
4.6.2	Mixed-type motifs enrichments with different null models	41
5	Conclusions	44
	REFERENCES	45

Introduction

In this chapter we give a brief overview of the main concepts that are needed to understand our work and correctly frame it in the context of the existing literature. In particular, we discuss whole-genome duplications, gene regulatory networks and network motifs. In the last section we show how these concepts are interconnected in the context of our work and why this interplay is important and interesting to analyze.

1.1 Whole-Genome Duplications

Gene duplication is one of the main drivers of evolutionary genomic innovation [1, 2, 3]. Most of the gene duplication events belong to one of two main categories: Small-Scale Duplications (SSDs) or Whole-Genome Duplications (WGDs). Small Scale Duplications typically involve a single gene, or a small set of genes within a well defined genomic locus. Whole-Genome Duplications, on the other hand, happen at a much larger scale and involve a macroscopic portion of the genome. Although even the mere existence of such events was initially met with high skepticism, it is by now clear that they played a major role in evolution instead [4, 5].

From the evolutionary standpoint, the two types of duplications produce very different outcomes. SSD events have a relatively low chance of having a huge impact on the organism in which they happen. The most likely outcome is that they introduce small and incremental (positive or negative) changes to the genome and consequently to cell functions, thus promoting a local exploration of the phenotypic landscape. WGD events in normal conditions, on the other hand, typically produce immediate dire consequences on the fertility and fitness of the organism, compromising its short-term survival [6]. As a result, most WGD events are not fixated in the population. When they do, though, they typically produce sudden and dramatic phenotypic changes, which could hardly be achieved by SSDs alone. In this sense, WGD events allow species to take big jumps in the phenotypic landscape, promoting a non-local exploration of the latter. In some peculiar circumstances though, like in harshly adverse environmental conditions, WGDs can provide an immediate evolutionary advantage to the organism, and help reducing

the risk of extinction of the affected lineage [7]. Also, increasing evidence points towards a central role of WGD in the successful response to the stress induced by sudden environmental changes [5]. On a much longer timescale, instead, fixated WGD events can contribute to the creation of biological complexity in the organism [5].

This thesis aims at quantifying the different impacts of such two different gene duplication mechanisms on the evolution of the human gene regulatory networks. For this reason, we will focus on the two rounds of WGD that occurred about 500–550 Millions of years ago, at the origin of the vertebrate lineage. These two are the only WGD events known today that were retained along the evolutionary trajectory leading to *Homo Sapiens*. More than 50 years ago, the geneticist Susumu Ohno [1] hypothesized for the first time, based on observations on the clustering of HOX genes, that two rounds of WGD were at the origin of the vertebrate lineage. This theory was met with skepticism for a long time, and it was only with the advent of high-throughput sequencing that reliable evidence supporting the existence of ancient WGD events became available. In 1997, a WGD event was unambiguously detected for the first time in *Saccharomyces cerevisiae* [8, 9] and a few years later in *Arabidopsis thaliana* [10]. In 2005 Ohno’s original intuition regarding the two WGD events at the origin of the vertebrate lineage was also confirmed [11]. Parologue genes that descend from a whole-genome duplication event are nowadays also called *ohnologs*, to honor Ohno’s early intuition. Vertebrate-specific double round of ancient WGD events is also often abbreviated with the acronym ”2R-WGD”. Since we will be dealing mainly with such events, we will stick to the simpler ”WGD” acronym.

Whole-genome duplication events have evolutionary consequences which substantially differ from the ones introduced by small-scale duplication events. In fact, WGDs are thought to have played a central role in the evolution of complex traits associated with vertebrates. This conjecture has not found final confirmations yet, but is supported by many different and converging observations. 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 [12]. It was also shown that the emergence of an important class of human highly interacting proteins, involved in processes that are crucial for the organization of multicellularity, was mainly due to vertebrate WGDs [13]. More generally, WGD events are recognized to have influenced many important evolutionary processes, such as gene retention and selection, dosage balance and subgenome dominance effects among others. Whole-genome duplication events are much more commonly found in plants, where these phenomena are especially well studied, in particular in *A. thaliana* [14, 15, 16]. Acting through such mechanisms, WGDs are acknowledged to have played a major role in the introduction of evolutionary novelties in many species. The findings presented in this thesis provide further support

to this theory.

Because of their ancient origin, the identification of WGD pairs or quartets in vertebrates has proven to be a highly non trivial task [17]. In fact, a stable and reliable list of human WGD gene pairs was only recently proposed [18, 19, 20]. 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 [21]. Moreover, WGD genes are threefold more likely than non WGD ones to be involved in cancers and autosomal dominant diseases [18]. 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 [22]. 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 [18, 19, 23, 24, 25]. 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 [19]. 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 [26] and in *A. thaliana* [24, 27]. This “universality” supports the hypothesis of general principles or mechanisms behind the unexpected retention of WGD genes and their interactions.

1.2 Gene Regulatory Networks

Cells can be regarded as very complex information-processing machines acting at the micron scale. They evolved the ability of sensing different signals both from their internal and external environment, successively integrating and elaborating them with the ultimate goal of triggering the best available response to a specific situation. For example, a cell sensing the presence of a specific nutrient in the environment might decide, depending on the type of nutrient and on the current state of the cell itself, to start producing the enzymes needed for the digestion of that nutrient. Cells also respond to signals sent from other cells around them. In multicellular organisms this ability plays a particularly important role, as cell-to-cell communications orchestrate, among others, organismal development, homeostasis and single-cell functions [28]. This wealth of signals need to be integrated, in order for cell to take decisions about its fate and the kind of machinery that it needs to produce. There are many ways in which cells can internally

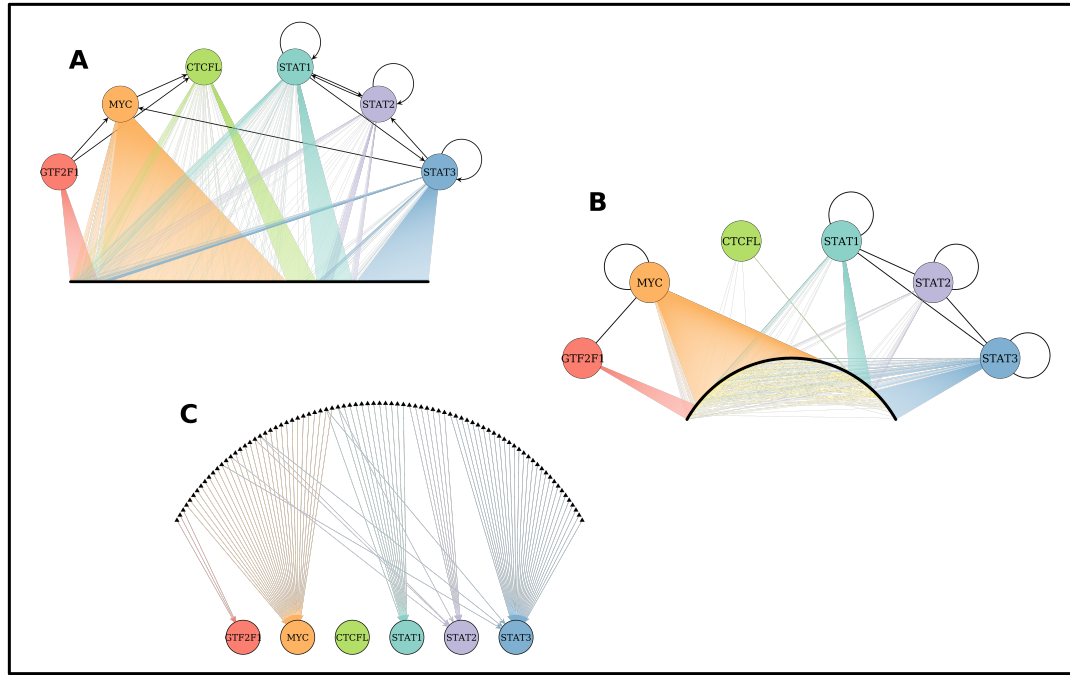


Fig. 1.1 Gene Regulatory Networks. Interactions involving an illustrative subset of TFs are shown for each of the regulatory mechanisms studied in the present work, i.e. for (A) transcriptional regulations from the ENCODE network, (B) protein-protein interactions in the PrePPI network, and (C) miRNA-gene 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.

represent and process the information about their internal and external environments, many of which are yet to be fully understood and, most probably, discovered. To the best of our current knowledge, though, the main mechanism seems to be the regulation of gene expression. Gene products can interact with other gene products or with DNA itself to modulate the expression of other genes in the cells, orchestrating a systemic response to the environmental cues.

Interactions between components in a complex system have a very natural abstract representation as networks. In a network, each of the system's component is represented as a node and a connection is established between two nodes whenever the two corresponding components interact with each other in the real system. Connections can be *undirected* when the relation is symmetric (e.g. the two components are in contact with each other) or *directed*, when the relation is asymmetric (e.g. one component activates or switches off the other one). This representation proved very powerful, in that it was used to describe and extract useful information from a wide variety of complex systems and opened new perspectives in the analysis of biological systems [29]. Reg-

ulatory interactions among genes in a cell can also be represented in this way. In our case, genes are represented as nodes in the network and their interactions are described as links between nodes. Different networks can be built with the same set of genes, by considering different interaction mechanisms. This specific kind of networks, representing interactions between genes, are commonly known as *Gene Regulatory Networks* (GRNs).

Gene regulation can take place at many different levels. At the transcriptional level, when proteins produced by specific genes bind to the promoter or the enhancer region of another gene, activating or repressing its activity. Genes that regulate other genes' expression in this way are called Transcription Factors (TFs). In Fig. 1.1A a portion of the human transcriptional regulatory network: we can see that a small number of transcription factors regulate a host of target genes, while also regulating each other's activity at the same time. At the post-transcriptional level, the activity of some genes can be modulated by micro-RNAs (miRNAs), very short RNA sequences (around 22 nucleotides long) that target the RNA molecules transcribed from the target gene, leading to their degradation. Fig. 1.1C shows the miRNAs that target the same TFs shown for the transcription network. Gene products can also physically interact at the protein level. Proteins generated by different genes might participate in the same protein complex, or one of the two could act as an activator or repressor by inducing chemical or structural changes in the other one. Fig. 1.1B portrays the protein-protein interaction occurring between six transcription factors and their target, and among the targets themselves.

Gene duplication is widely recognized as one of the main mechanisms underlying the evolution of gene regulatory networks [30, 31, 32]. Gene duplications, both at small and large scales, generate a lot of redundant interactions, that provide raw material for the subsequent evolution of the network. It is indeed easier to rewire existing duplicated interactions, than creating *de novo* interactions between genes from scratch. Moreover, the presence of duplicated interactions confers additional robustness to the system, that can keep one of the old functional interactions as backup and let the other one evolve freely to implement new functions. All of these processes deeply influence the topology of regulatory networks, by adding and deleting nodes and links and rewiring existing links. Such changes evidently also shape network functions and, ultimately, the ability of cells to elaborate complex signals and decision-making strategies.

1.2.1 Network Motifs

If we look at the local structure of complex networks, we find that there are some combinations of nodes and regulatory interactions that are statistically over-represented in networks [33], commonly called *network motifs*. In this case, "over-represented"

means that their occurrence in the network is much higher than what we would expect to see if the links in the network were arranged in a random manner. The presence of networks motifs was assessed in many complex networks of very different kinds, social, technological and biological. They assumed a special importance in the analysis of gene regulatory networks in particular, both in bacteria [34] and higher organisms, notably humans [35].

In the context of transcriptional networks, network motifs were shown to perform elementary regulatory functions [36, 37, 38] 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 network [31].

This thesis focuses on network motifs involving pairs of duplicated genes, as illustrated in Fig. 1.2. 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. We will also quantify the effects of the different types of duplications on the structure of the PPI network.

1.3 Motivation and outline of the work

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. 1.1 A, B, and C respectively. As exposed in the previous section, network motifs assume particular significance in biology and for gene regulatory networks in particular. In this context, network motifs identified gene circuits that can perform relatively simple computations with specific biological functions. These

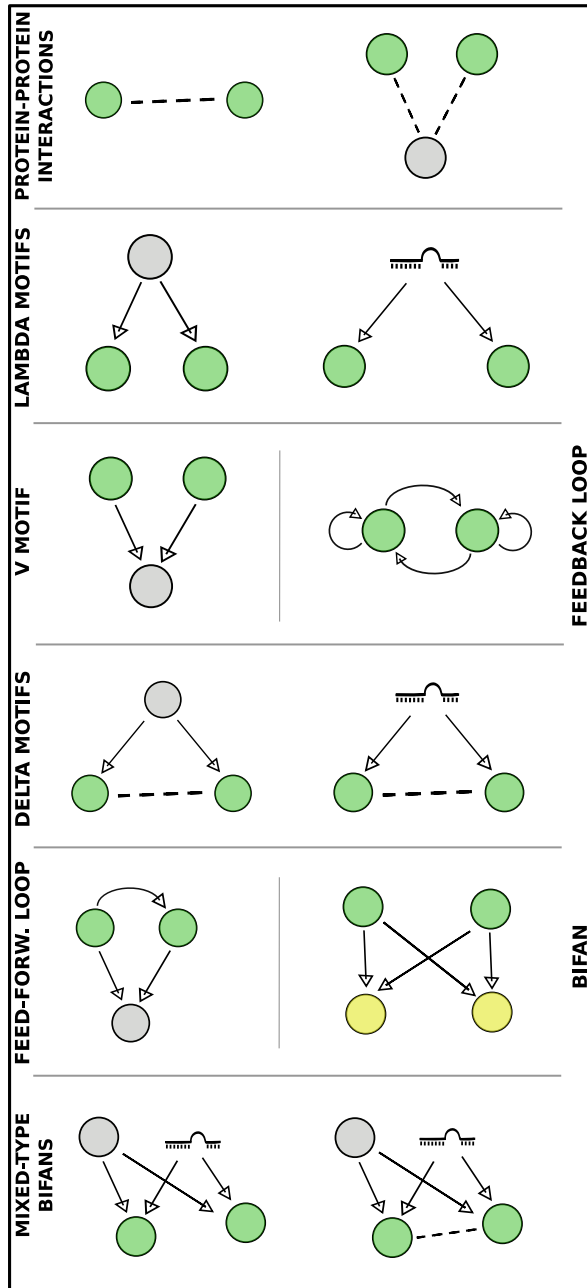


Fig. 1.2 Network Motifs considered in this thesis. An overview of the network motifs that will be considered in this work. Gray circles represent generic genes, same-color circles (green or yellow) are paralogues, and miRNAs are represented in a stylized form. Solid arrows represent regulatory interactions, while dashed lines represent protein-protein interactions.

simple modules can then assemble into a larger network to implement complex and robust regulatory strategies [38]. As shown in Fig. 1.3, 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. 1.2 - whose functional importance is widely recognized [38]. 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 [39]. We will show that also the structure of additional layers of regulation present in the human genome, such as miRNA regulation, has been influenced by duplication events. In conclusion, this work will show 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. We will argue that 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 obtain (and retain) through a sequence of small-scale duplication events.

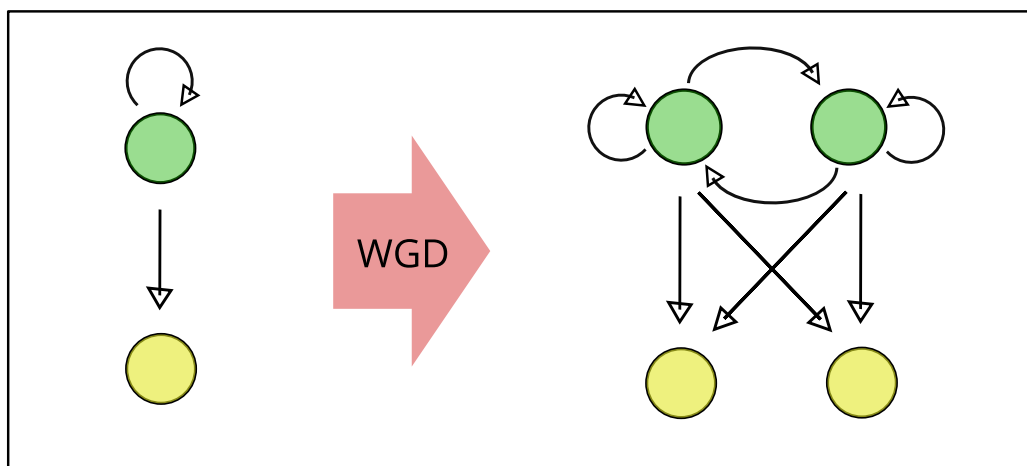


Fig. 1.3 Gene duplications generate network motifs. Illustration of how a WGD event can easily create FFLs and Bifans, by duplicating the components of a simple regulatory interaction in which the regulator also has self-regulation. Many of the created interactions will then be lost during the evolutionary process, leaving only those that are not negatively selected.

Materials and Methods

2.1 Gene Nomenclature

This thesis uses data from many different sources, some of them produced in times pretty distant from one another. It was therefore necessary to unify consistently the notation for the unique identifiers of the genes, in order to make them comparable in a sensible way. The Gene Symbol format was chosen, and all the genes appearing in the various datasets were translated to their official Gene Symbol, as indicated by the HGNC (HUGO Gene Nomenclature Committee - <https://www.genenames.org>).

Gene Symbols can come in different possible statuses, and different actions were taken depending on it:

- APPROVED: symbol left untouched.
- PREVIOUS: obsolete gene symbol, updated to the official one.
- WITHDRAWN: the gene symbol does not exist anymore, deleted.

For data available only in Ensembl ID format, we used the official mapping from one nomenclature to the other obtained from the Ensembl BioMart interface.

2.2 Transcriptional Regulatory Network

We used the human transcriptional regulatory network presented in [40], a portion of which is displayed in Fig. 1.1A. 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”. Our aim in this work is to analyze the effects of gene duplication events on the regulatory interactions in general, so we combined the information regarding proximal and distal regulation into a single regulatory network, with 122 transcription factors (TFs) and 9986 target genes. Since this choice of the network is at the foundation of all of the results that will follow, it is worth discussing it in some more detail.

The fundamental requirement of the transcriptional regulatory network to be used in our work is that it must not introduce - ideally - any kind of bias that could mislead our statistical analyses. In particular, we require that the chosen network present the least possible amount of bias with respect to specific kinds of network motifs and with respect to the duplication status of the considered genes. Such requirements are all met by the network in [40], which was indeed already used by the same authors to carry out motif enrichment analyses in a spirit similar to the one of this work.

In this paragraph we instead briefly outline the reasons that led us to discard other possible network candidates. Besides the Chip-seq derived networks, there are essentially three other methods to construct transcriptional regulatory networks, see for instance [41] for a recent review. Literature-based collections (such as TRRUST [42] or HTRIdb [43]) 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. 1.2). In the end, methods based on reverse engineering gene expression data, such as the popular ARACNE algorithm [44], 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.

2.3 Protein-Protein Interaction Networks

We extracted two protein-protein interaction (PPI) networks from the PrePPI database [45] and the STRING database [46]. 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). In order to limit possible confusions, we will talk mainly about the results obtained with the PrePPI network, which was deemed more robust due to the presence of experimentally validated interactions. However, all of the results are confirmed by analysis of the STRING network, which

will also be shown and commented more briefly.

2.4 miRNA-gene Interaction Networks

The miRNA-target interaction networks we considered come from the TarBase database [47] and the mirDIP database [48]. 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). Also in this case, the TarBase network was deemed more dependable due to the fact that the interactions are not mere predictions and to absence of an arbitrary limitation in the amount of interactions that were retained. The *Results* and *Discussion* chapters will therefore be mostly concerned with commenting the behavior of the TarBase network, while results for the mirDIP network will be briefly exposed. Again, the trends we find in the two cases show some degree of robustness, despite the low overlap in interactions between the two networks.

2.5 Small-scale and Whole-genome duplicates

2.5.1 WGD paralogues

The Whole-Genome Duplicate (WGD) gene pairs were obtained by merging the results of *Makino and McLysaght* [21] with the latest available OHNOLOGS database [20]. 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 [49]. In this work we only consider paralogue couples composed of protein-coding genes, so we restricted all the obtained paralogues' lists accordingly. Data about the protein coding nature of the genes were also obtained from the Ensembl database. After such pre-

liminary data manipulations, we ended up with a list of 8070 WGD-derived paralogue couples, comprising 7324 different genes.

2.5.2 SSD analogues

SSD-derived analogues were obtained from the list of all human analogues involving protein-coding genes in the Ensembl database [50], 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 analogues. 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). This assumption is not valid for SSD couples, since they do not have an overall precise time location, but are generated continuously throughout the history of human evolution. It is therefore reasonable to expect that a portion of the SSD couples are more or less contemporary to the WGD ones, while some others are younger or considerably younger. Following this recent duplication events, sequence evolution had relatively less chances to modify and rewire the gene interactions involving the resulting analogues. 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 analogues. Such effects are indeed present, even if small, as we show in detail in Fig. 4.2 of the *Discussion* chapter.

The duplication age of paralogue gene couples was estimated by considering their most recent common ancestor, as reported in the Ensembl database. Data obtained in this way are not suitable to give a precise date of duplication for each couple of genes, but can be used in an heuristic fashion to discern SSD couples that roughly date back to the same time period when WGD couples were generated from more recent duplicates. More 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 previous approaches [18] and indeed the estimated ages are compatible, as shown below in Fig. 2.1. Following these criteria, 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.

2.5.3 Age distributions of analogues

Duplication age was shown to bear sizable effects on the evolution of gene networks, in protein-protein interactions in particular [51]. It is thus important to account for this

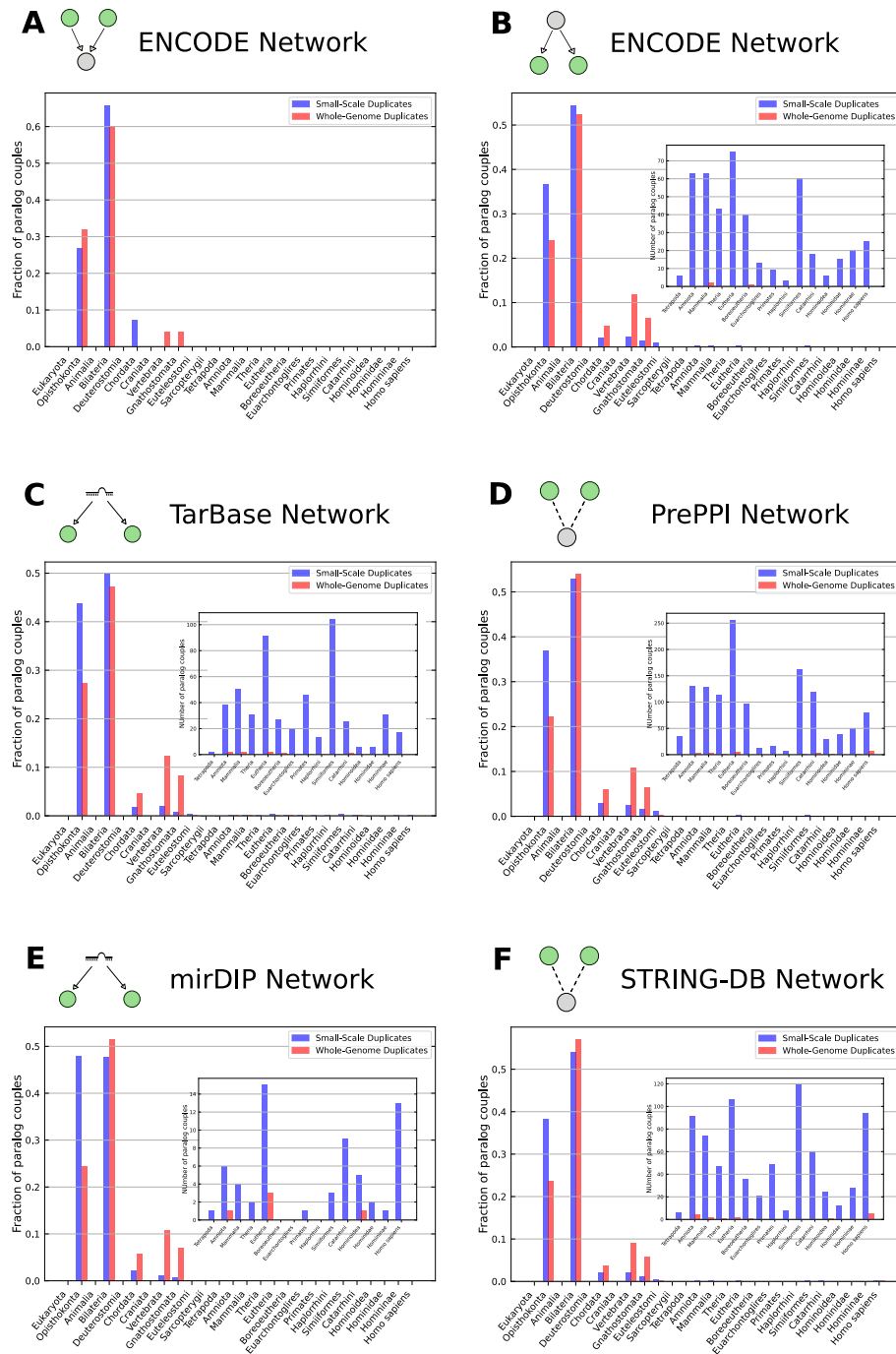


Fig. 2.1 Age distribution of WGD and SSD paralogues contained in the interaction networks. The insets zoom in on the portion of the distribution after *Sarcopterygii*. The distribution for the ENCODE network is subdivided into TF couples (A) and target couples (B) for clarity.

additional factor when comparing duplicates of different ages, as it happens for SSD and WGD couples. We show in Fig. 2.1 that the dating of the paralogues with their most recent common ancestor is consistent with the independently found distinction between SSD and WGD couples. The shown distributions also justify the heuristic

choice of considering SSD couples duplicated before *Sarcopterygii* as evolutionarily comparable to WGD couples. The figure also highlights the presence of SSD couples that were duplicated in relatively recent times, but which are a very small percentage of the total number of SSD couples. Such recent couples were not taken into consideration in all of the subsequent analyses.

2.5.4 Non-duplicated gene couples

In the analyses that follow we will sometimes compare the results obtained for SSD and WGD couples to those obtained for non-duplicated gene couples. We consider as non-duplicated couples all of the couples that one can construct with the genes in a specific network that are neither SSD couples nor WGD couples.

2.6 Network Motif enrichment

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 for each of the random models that are defined in a following section.

2.7 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

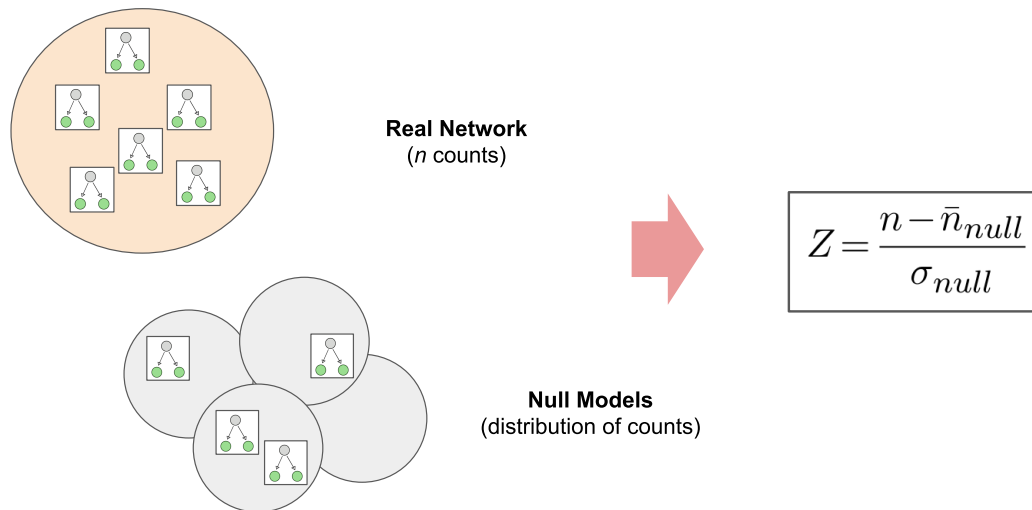


Fig. 2.2 Calculation of network motif enrichment. First the number of occurrences of each motif is calculated in the real network (top). Then an ensemble of randomized networks is created, in such a way that the degree sequence of the original network is conserved. The occurrence of each motif is calculated in each of the null networks and the obtained distribution is compared to the real count to obtain a Z-score.

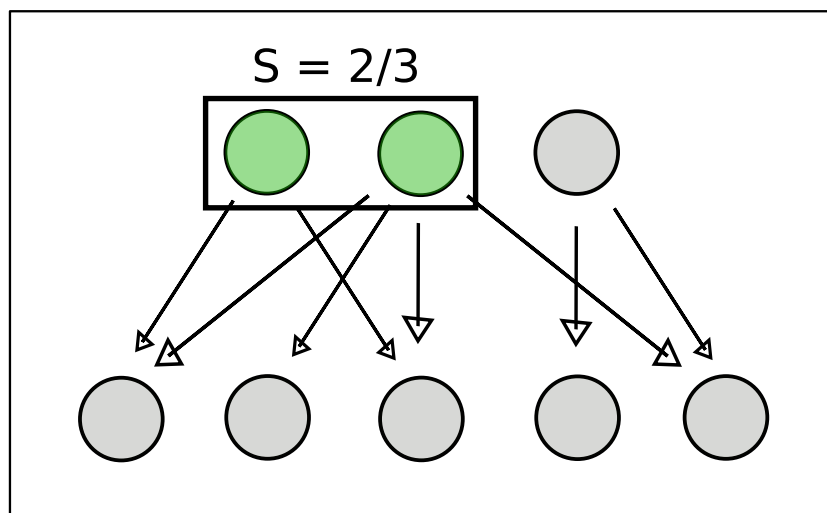


Fig. 2.3 An example of similarity calculation. Example of the structure of a Dense Overlapping Regulon (DOR) embedded in a gene regulatory network, with the target similarity S calculated for an illustrative couple. In this case the gene on the left has a total of 2 targets and the one on the right a total of 4, while they have 2 common targets. Plugging these numbers into the similarity formula gives the shown result of $S = 2/3$ for the selected gene couple.

similarity scores and the motif Z-scores can be found in the next section. A clarifying graphic example of how the similarity between two regulators is calculated is shown in Fig. 2.3.

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. Should one or more of such symbols not be reported, it would mean that the corresponding comparison is not statistically significant. 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 $P < 1e - 5$.

2.7.1 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. Since these two different measures are at the core of our results, in this section we discuss how they differ and why it is important to show them both, with the aim of clarifying their role and interpretation in the context of the present work. 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. 4.2 of the *Discussion* chapter. 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.

2.8 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 [52] and illustrated in Fig. 2.4. This

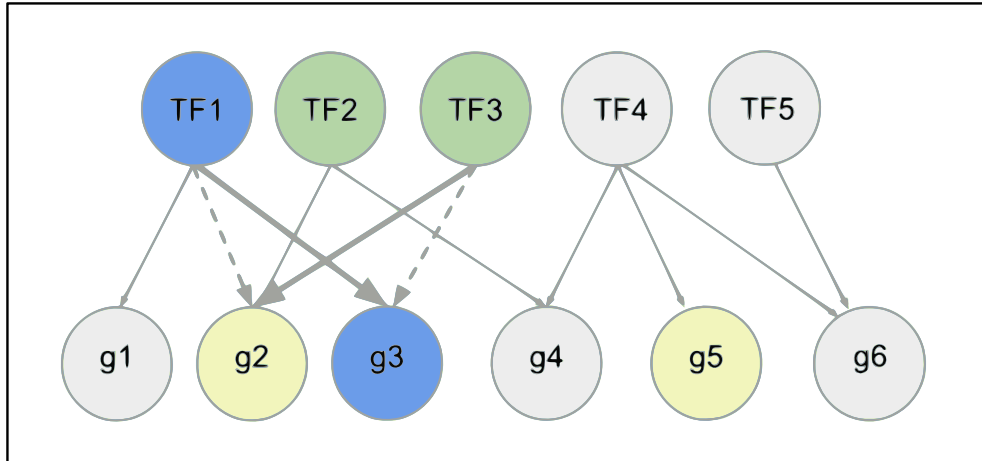


Fig. 2.4 Null models creation with degree-preserving randomization. Graphical representation of the degree-preserving procedure used to generate the null models. The dashed links are randomly chosen and their ends swapped, thus generating the new bold links. Note that all of the involved genes maintain their in and out degree in the process.

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 [33, 53].

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. 4.3 of the *Discussion* chapter.

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.

2.9 Data and code 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](https://doi.org/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>.

Results

This chapter presents the results of our motif enrichment analyses, in order of increasing topological and functional complexity of the circuits considered.

3.1 Degree distributions

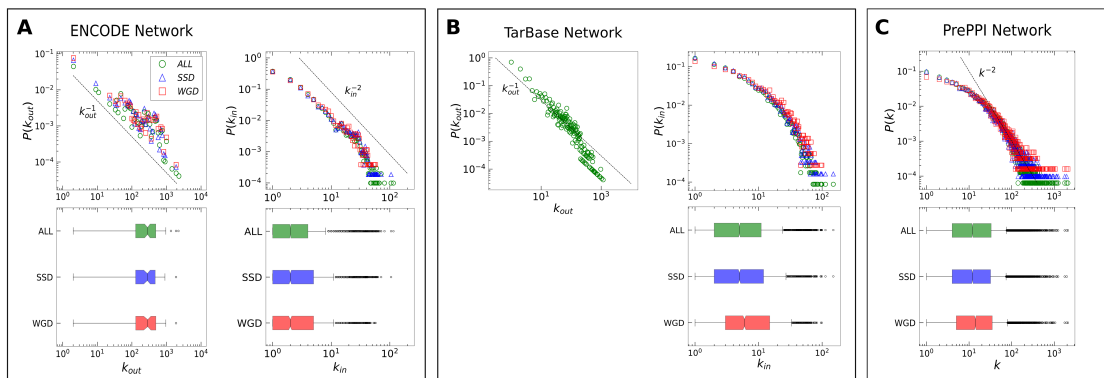


Fig. 3.1 Degree distributions (ENCODE, PrePPI, TarBase networks). Indegree (k_{in}) and outdegree (k_{out}) distributions of (A) the ENCODE transcriptional regulatory network and of (B) the TarBase miRNA-gene regulatory interactions network, and the degree (k) distribution of (C) the PrePPI protein-protein 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. The boxplot representation, although less commonly used for degree distributions, clearly shows the similarity of the distributions in all of the three networks.

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

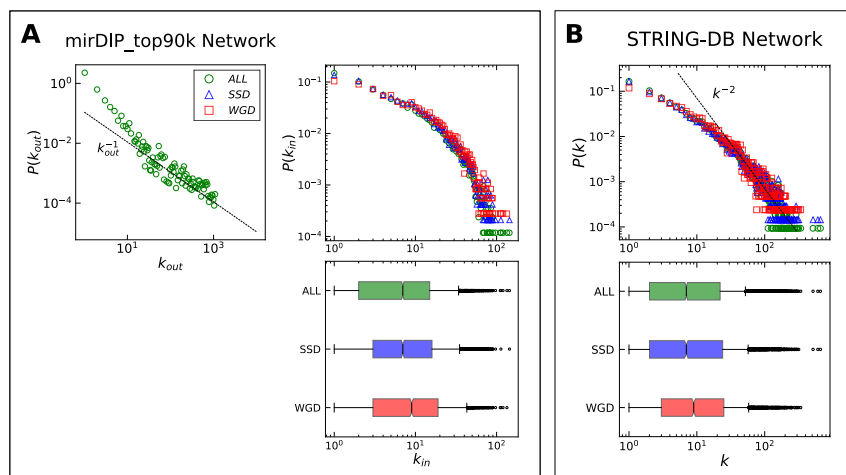


Fig. 3.2 Degree distributions (STRING and mirDIP networks). Indegree (k_{in}) and outdegree (k_{out}) distributions of **(A)** the mirDIP miRNA-gene regulatory network and the degree (k) distribution of **(B)** the STRING protein-protein interactions network. Each degree distribution is shown both as 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.

Fig. 3.1 for the ENCODE, TarBase and PrePPI networks and in Fig. 3.2 for the mirDIP and STRING networks. 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. As it is often observed in this type of networks, the distributions are power-like. However, as far as the present work is concerned, it does not really matter which is the exact shape of the various distributions, as long as any macroscopic difference in degree distribution for the two types of duplicates is ruled out. Therefore, we conclude that duplications do not display specific biases in terms of gene degree in the different networks considered. This is an important preliminary observation, since in the following we will focus on regulatory circuits whose statistics could be dependent on the degree of the nodes. The absence of relevant differences suggests that the peculiar role of duplicated genes in the regulatory network is not associated to the sheer number of their interactions, but it must be due to more complex topological properties (i.e. network motifs). As we discuss in the next sections, this is exactly what we observe.

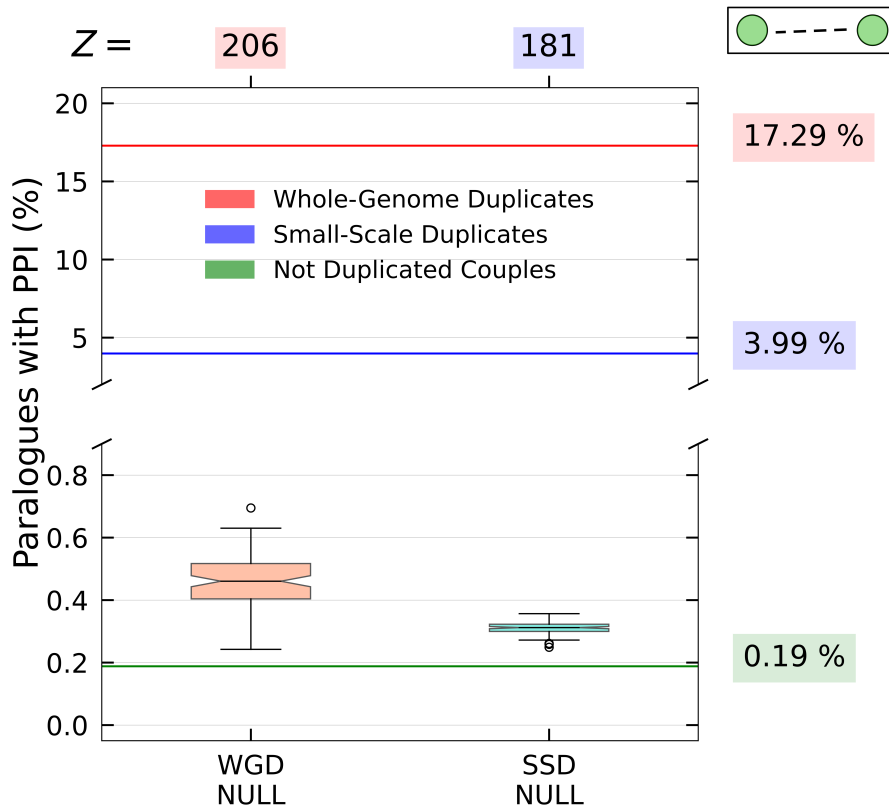


Fig. 3.3 Interactions of duplicated genes in the PrePPI network. The percentages of gene pairs that present an interaction at the protein level according to the PrePPI database are indicated by the bold horizontal lines and explicitly stated in the labels on the right. The null model distributions are reported in the boxplots and the corresponding Z-scores are shown at the top.

3.2 Duplicated genes often interact at the protein level

In this section we analyze 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. 3.3, 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

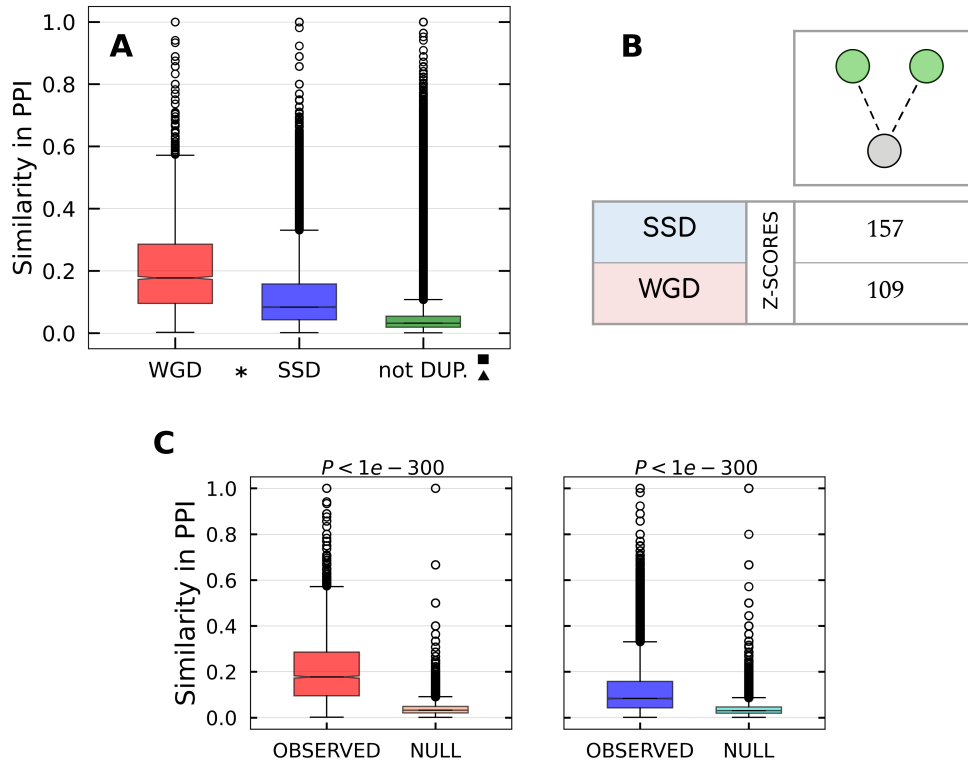


Fig. 3.4 Pairs of duplicated genes interacting with a third protein in the PrePPI network. (A) Similarity distributions for WGD, SSD and not duplicated gene couples in the PrePPI network. All of the pairwise comparisons between distributions are statistically significant, as indicated by the presence of the symbols explained in the *Methods* chapter. (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.

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, which is captured by the statistics of co-interaction motifs presented in Fig. 3.4. In particular, Fig. 3.4A 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. 3.4C). This tendency explains the enrichment of co-interaction motifs shown by the Z-scores in Fig. 3.4B.

The evolutionary tendency to retain WGD couples that participate in common protein complexes agrees with previous observations in yeast [26, 54], 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 [55]. The same trend was reported

for the human genome using a database of transient protein complexes [22]. We shall see in the *Discussion* chapter a nice example (the RAR/RXR pathway) of how the retention of protein-protein interactions among WGD pairs and the tendency to maintain their interactions with common partners may increase the variety and complexity of the functions performed by the genes involved in the WGD event.

3.2.1 Comparison between the PrePPI and STRING networks

The results referred to the PrePPI network, presented in the section above, are independently confirmed by analyzing the STRING protein-protein interaction network. By comparing Fig. 3.3 and Fig. 3.5, we can easily see that the proportion of paralogue couples which also have a protein-protein interaction are almost identical in the PrePPI and in the STRING case. The Z-scores associated to the STRING results in this case are also of the same order of magnitude of the ones found in the PrePPI network, confirming the overall statistical significance of the observed signal. As observed in Fig. 3.6, the effects of paralogy relations on the tendency of couples of duplicated genes to form protein complexes with a third common protein are even more pronounced in the STRING network than in the PrePPI network. The similarity in common contacts is much higher in the STRING network for both SSD and WGD couples, while the similarity distribution of non-duplicated is comparable in the two different networks. All of the pairwise comparisons between the three different types of couples are statistically significant (see Fig. 3.6A), as well as the comparisons with the null model (see Fig. 3.6C). The Z-scores associated to the co-interaction motif are well beyond statistical significance and an order of magnitude higher than the ones observed for the PrePPI network (Fig. 3.6B), strongly supporting all of the findings presented in the previous section.

3.3 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 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. 3.7A). 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. 3.7 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

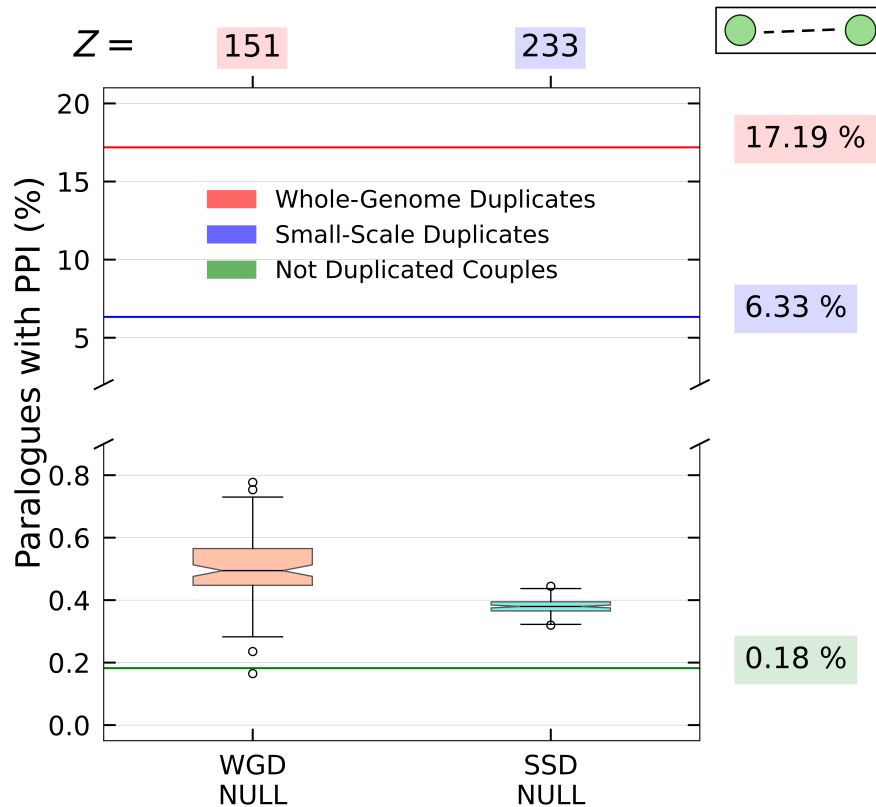


Fig. 3.5 Interactions of duplicated genes in the STRING network. The percentages of gene pairs that present an interaction at the protein level according to the STRING database are indicated by the bold horizontal lines and explicitly stated in the labels on the right. The null model distributions are reported in the boxplots and the corresponding Z-scores are shown at the top.

lost or rewired during evolution. A similar phenomenon was observed in yeast [39], 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 [56]), 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 [57], thus suggesting that it could be a general phenomenon.

3.4 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. 3.8 and Fig. 3.9 respectively. The similarity distributions of WGD and SSD genes are both larger than the non-duplicate one for both types of Λ motifs. Coherently, the Z-scores indicate en-

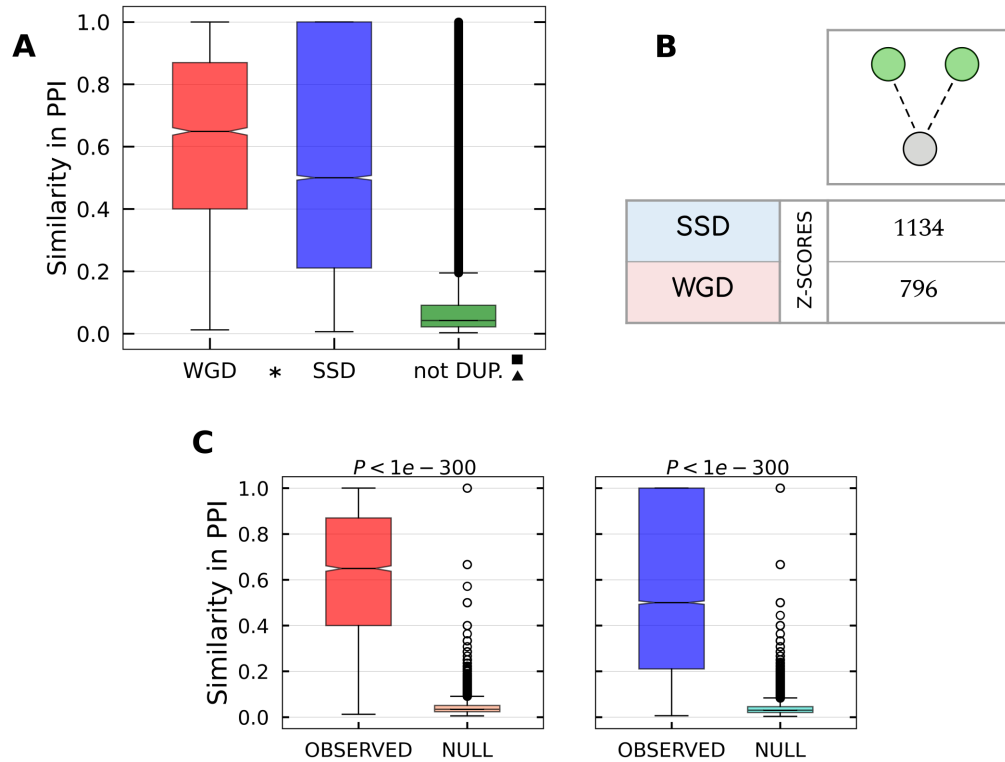


Fig. 3.6 Pairs of duplicated genes interacting with a third protein in the STRING network. (A) Similarity distributions for WGD, SSD and not duplicated gene couples in the PrePPI network. All of the pairwise comparisons between distributions are statistically significant, as indicated by the presence of the symbols explained in the *Methods* chapter. (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.

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

3.4.1 Comparison between the TarBase and the mirDIP networks

We conducted the same analyses for the enrichment of miRNA-mediated Λ motifs on the mirDIP network. The results are reported in Fig. 3.10 and completely confirm all of the observations made in the previous section, which were based on the TarBase network instead. In particular, we see in Fig. 3.10A that the similarity distributions of SSD and WGD couples differ from the one for non-duplicated couples in a statistically significant manner, but are compatible with each other. Fig. 3.10C also confirms that

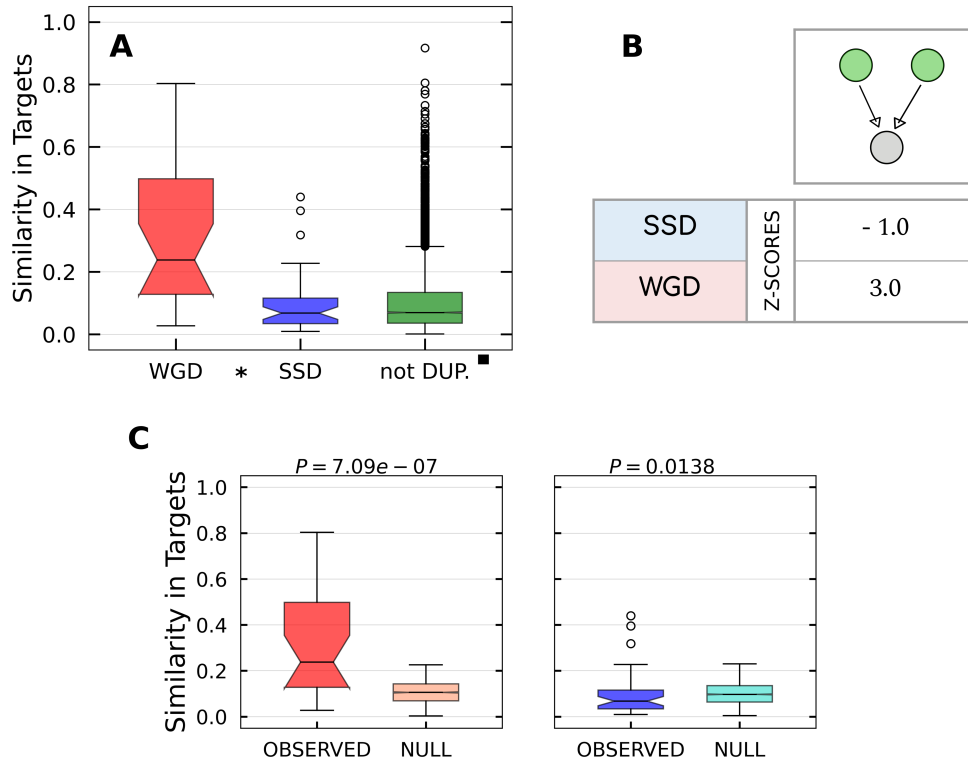


Fig. 3.7 Transcriptional V motifs. (A) Similarity distributions for WGD, SSD and not duplicated TF couples in the ENCODE network. As indicated by the presence of the symbols explained in the *Methods* chapter, the difference between SSD and not-duplicated distributions is not statistically significant 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.

both SSD and WGD similarity distributions are significantly higher than the null model. In the end, Fig. 3.10B shows that motifs enrichments for SSD and WGD couples are even stronger than in the TarBase network, but go in the exact same directions.

3.5 More complex motifs are enriched in duplicated genes.

The role played by WGD-derived genes in shaping the regulatory network emerges more clearly if we look at more complex network motifs, such as Feed-Back Loops (FBLs), Feed-Forward Loops (FFLs) and BiFan-type motifs (Fig. 3.11, and 3.12). These motifs were all shown to be associated to specific and relevant functions, that will be discussed in the corresponding sections.

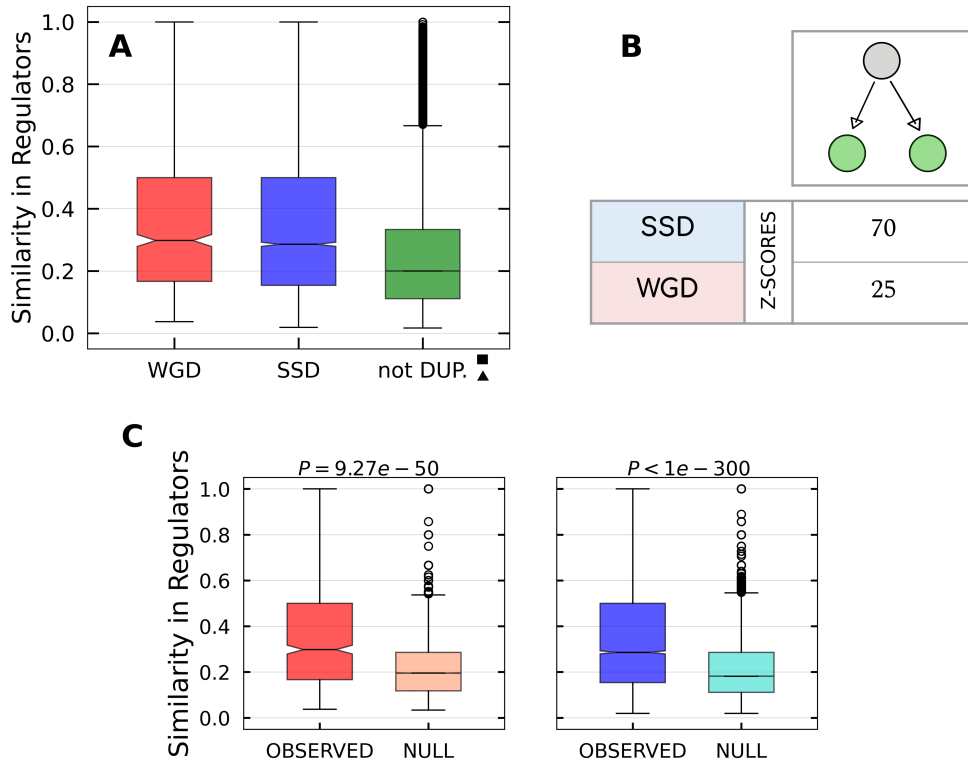


Fig. 3.8 Transcriptional Λ motifs. (A) Similarity distributions for WGD, SSD and not duplicated target genes couples in the ENCODE network. As indicated by the presence of the symbols explained in the *Methods* chapter, the difference between SSD and WGD distributions is not statistically significant, while both of them are significantly greater than 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.

3.5.1 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 [38] 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 [38]. This simple motif is associated to several important functions, such as the modulation of the expression response time, robustness to stochastic noise, and bimodality in the protein levels [38]. In our analysis, the number of observed FBLs is so small that statistical tests are not meaningful, therefore we simply categorised the 25 pairs of WGD TFs and the 41 pairs of SSD TFs according to their topological configuration. Fig. 3.11 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

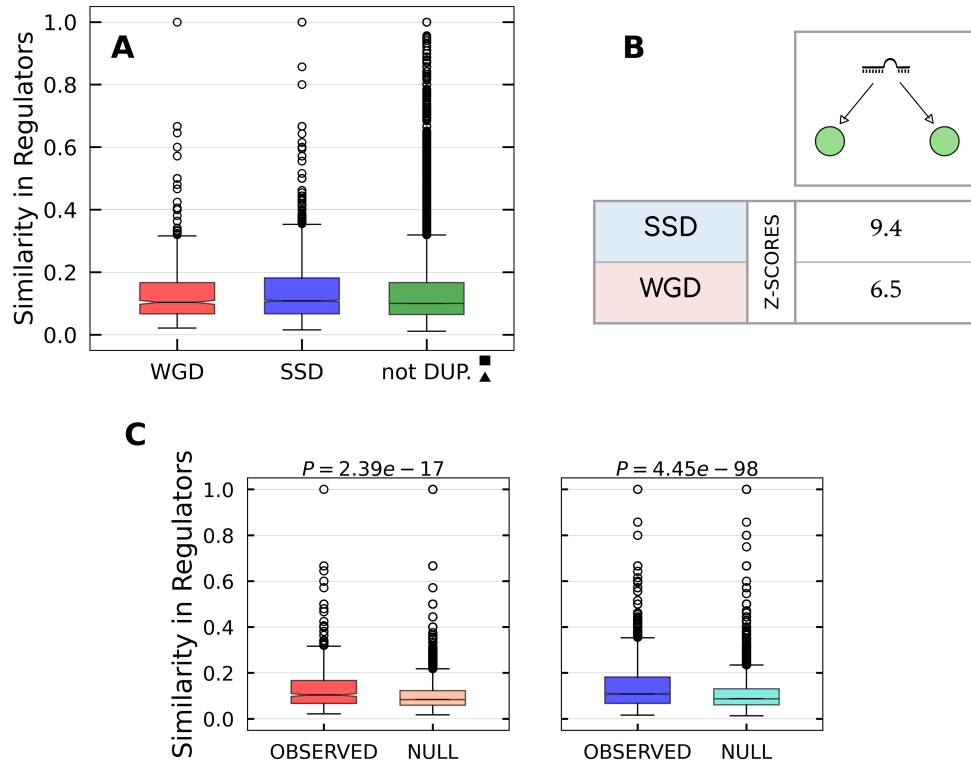


Fig. 3.9 miRNA Λ motifs in the TarBase network. (A) Similarity distributions for WGD, SSD and not duplicated target genes couples. The difference between SSD and WGD distributions is not statistically significant, while both of them are significantly greater than the similarity distribution of non duplicated 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.

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. 3.11 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.

The tendency for SSD events not to maintain the cross-regulations generated after the duplication of a self-regulating gene was also noted in *E. Coli* [58], strengthening our confidence that the effect we see might not only be due to the small number of couples present in this case.

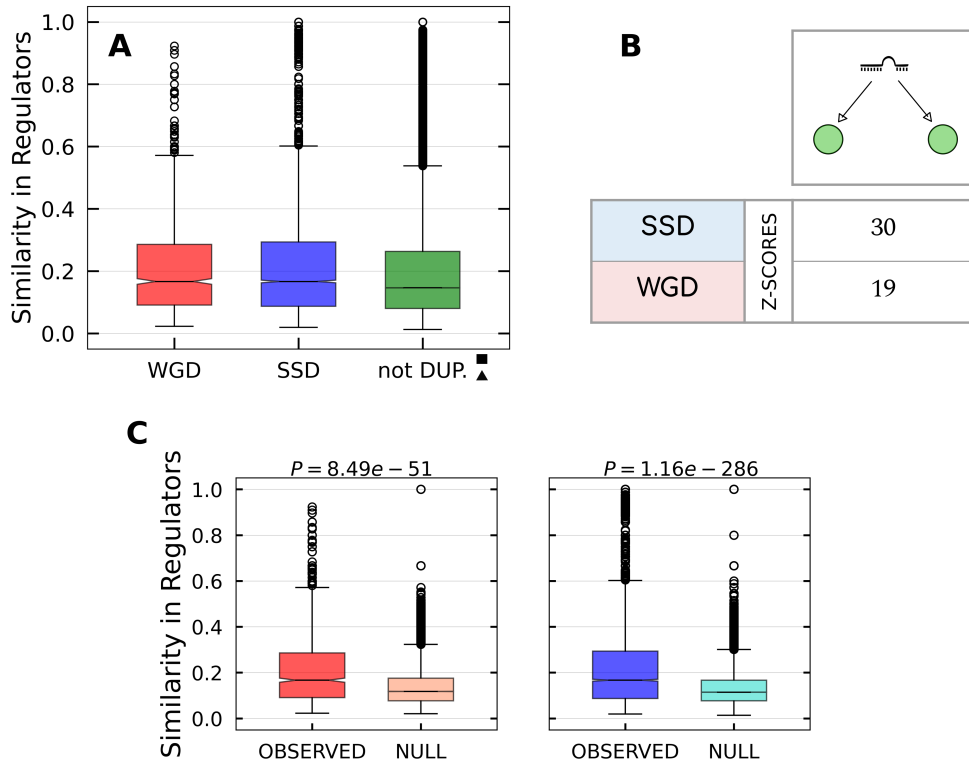


Fig. 3.10 miRNA Λ motifs in the mirDIP network. (A) Similarity distributions for WGD, SSD and not duplicated target genes couples. Also in this case the difference between SSD and WGD distributions is not statistically significant, while both of them are significantly greater that the similarity distribution 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.

3.5.2 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 [38]. 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 [38].

Fig. 3.12A 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.





		ONE SELF-LOOP		TWO SELF-LOOPS	
					
SSD	A	POU2F2 POU5F1 GABPA ELK4 GABPA ETS1 GABPA ELF1 GABPA SPI1 E2F6 E2F4 ESRRA NR3C1 FOSL2 ATF3 FOSL2 BATF SREBF2 USF2 SREBF2 USF1 SRF MEF2C BHLHE40 HEY1 REST ZNF274			
		14/41	0/41	0/41	0/41
WGD	B	E2F6 E2F1 FOSL2 FOSL1 ESRRA ESR1 JUN JUNB JUND JUNB GATA2 GATA3 GATA1 GATA3	FOSL2 FOS SREBF2 SREBF1 MEF2A MEF2C	STAT3 STAT2 STAT1 STAT3 JUN JUND GATA2 GATA1	SP1 SP2 STAT1 STAT2 FOXA1 FOXA2
		7/25	3/25	4/25	3/25

Fig. 3.11 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 topological arrangements.

3.5.3 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. 3.12B and 3.12C 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” [57] or “Dense Overlapping Regulons” (DORs) [38], that were identified for the first time in *E. Coli* [34]. 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. 2.3 and 2.4, but in real regulatory net-

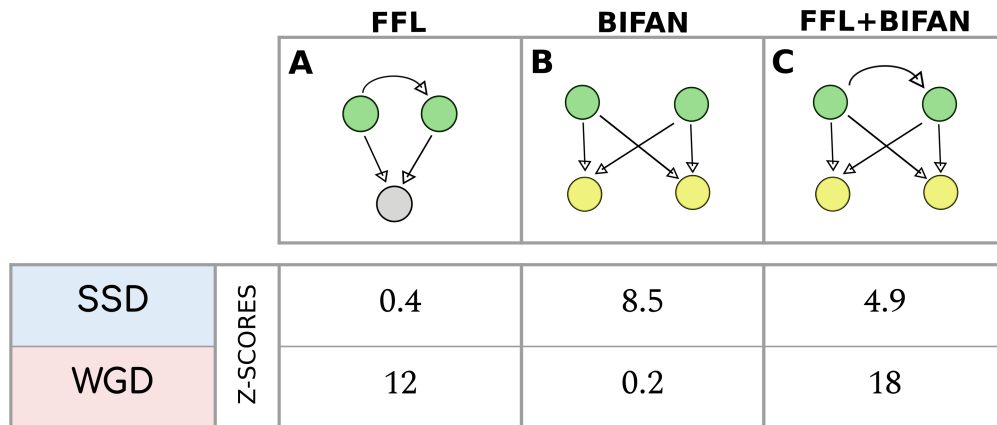


Fig. 3.12 Transcriptional FFL, Bifan and FFL+Bifan motifs. (A) Transcriptional Feed-Forward Loops (FFLs). (B) Transcriptional Bifan motifs (in which no regulatory is present between the two TFs). (C) FFL+Bifan motif. The extra connection between the two regulators transforms the array of simple regulations into an array of FFL motifs. 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).

works Bifan arrays can involve dozens of genes. The additional presence of regulatory interactions among regulators further 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. 3.12B for simple Bifans and in 3.12C 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.

3.6 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

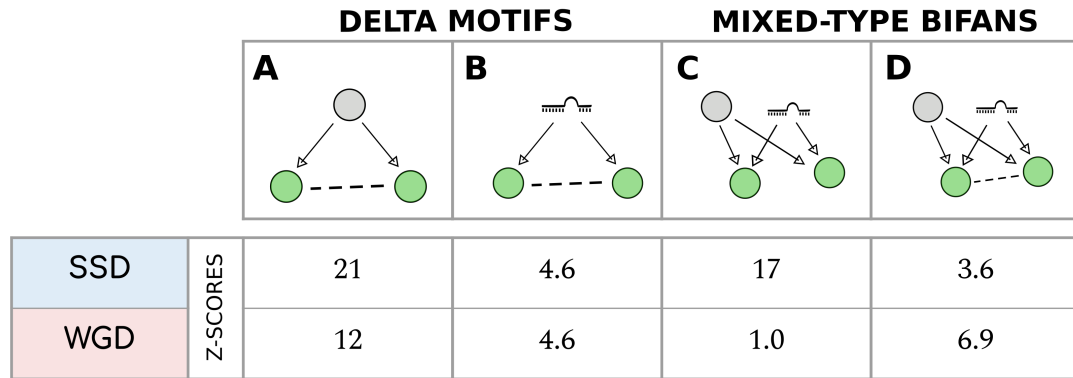


Fig. 3.13 Motifs with mixed-type regulatory interactions. (A) Transcriptional Δ motifs. (B) miRNA-mediated Δ motifs. (C) Mixed Bifan motifs, in which a pair of target genes are regulated both by a common TF and a common miRNA. (D) Mixed Bifan motif in which the two target genes interact at the protein 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).

(Fig. 3.13A). 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 Δ motifs. In the case of miRNA-mediated Δ motifs (Fig. 3.13B), 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 Δ motif, are reported in Fig. 3.13C. The enrichment of mixed-type Bifans with additional protein-protein interactions between the duplicated genes, is instead shown in Fig. 3.13D. 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 [59, 60].

Discussion

In this chapter we discuss the results presented in the previous chapter. We will draw connections between our observations and known facts about gene regulatory strategies and genome evolution, and analyze the evolutionary implications of our findings.

4.1 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 [61, 62, 63]. 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 [64]. Another recent study on *A. thaliana* similarly concluded that dosage balance constraints operate immediately after WGD and that duplicate gene retention patterns are shaped by selection to preserve dosage balance [65].

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 [63]. This is especially important for classes of genes which show a high level of dosage sensitivity: such genes are preferentially retained in double copy over long evolutionary timescales [66]. Studies conducted on the metabolic network of *A. thaliana* also suggested that different types of dosage constraints - relative and absolute - influence the retention of duplicates at different timescales after WGD events [67].

The dosage-balance principle was also recently invoked to explain SSD retention [68]. 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 last type of behaviour have been found both in yeast and in mammals [68].

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 [68]. 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 [69]. 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 [60] 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 [60, 59], i.e., in presence of a “mixed”-type network motifs. All of these considerations are indeed confirmed by the findings presented in Fig. 3.8, 3.9 and 3.13C.

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 [70]. 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. 3.3, 3.4 and 3.13).

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 [21].

4.2 Regulatory redundancy

It is widely recognized that gene duplications played a central role in the evolution of gene regulatory networks [31, 71] and in setting the TF repertoire [56].

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) [72]. The typical completion time for these processes is of a few millions of years [73], 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 [74].

Our study suggests that the retention of regulatory redundancy is strongly dependent on the duplication mechanism. The topological enrichment of V motifs and the distribution of target similarity (Fig. 3.7) suggest a significant 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 [56]. Most of the duplicated TFs kept almost unchanged their DNA binding sequence leading to an organization of TFs in “motif families”. The TFs within one of these families have similar DNA binding sequences and are thus expected to perform redundant regulatory functions. The size and distribution of these motif families is well fitted by a simple Birth-Death-Innovation evolutionary model [56] which is controlled by a single parameter θ . This parameter encodes the level of regulatory redundancy of the TFs repertoire of the organism, i.e. the tendency of TFs to keep the same binding preferences [56]. It was shown in [56] that the parameter θ increases with the complexity of the organism. It takes rather small values for yeast ($\theta \sim 0.2$) and *C. elegans* ($\theta \sim 0.3$), intermediate values for *D. melanogaster* ($\theta \sim 0.5$) and much higher values for mouse and human ($\theta \sim 0.75$). The fact that the tendency to maintain high redundancy in couples of duplicated TFs is much stronger for WGD paralogues once again associates WGD events to a higher complexity.

There are several possible explanations connecting increased genetic regulatory redundancy with increased complexity. First of all, regulatory redundancy can increase the robustness against mutations [75], 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 evolved to respond to different cellular signals or to interact with different upstream proteins [76, 2]. We shall see a nice example of this pattern in the RAR/RXR pathway which we discuss more in depth in one of the next sections.

In principle, combinatorial regulation - and the associated benefit of an increased environmental responsiveness - could arise by combining the regulatory interactions of any 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 responsiveness and noise control in gene regulation, and it has been suggested that it could be resolved by gene duplications [77, 78]. This hypothesis was tested in yeast for the specific Msn2-Msn4 pair of WGD-derived Transcription Factors [77], 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 model organisms like *S. cerevisiae* and *A. thaliana*. The newly available data on WGD genes give us the unique opportunity to extend previous studies, also encompassing the vertebrate lineage. We observed that several trends are conserved across different species and overall it seems that ancient WGD events had a relevant role in shaping current regulatory redundancy.

4.3 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 is graphically represented in Fig. 1.3, 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 common 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, Bifan and FFL+Bifan circuits can also 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. The presence of complex structures retained for more than 500 millions of years is highly non-trivial, since they must have survived a lot of selective pressure. Interestingly, Fig. 3.12 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 retention of the less complex Bifan motif, while WGD duplications are associated to more complex FFL arrays.

The fact that WGD events selectively favoured the enrichment of more complex FFL+Bifan configuration only might sound surprising. In general, common sense suggests that mechanisms that are able to produce structures with a certain degree of complexity should be able to produce also the less complex ones. We can make some hypotheses on the reasons why we observe this peculiar pattern. First of all, we observe that WGD events duplicated every single gene by definition and that SSD events target every gene with (approximately) the same probability. Therefore, we have no specific reason to believe that some genes were preferred over the other at duplication time and every difference we see must be due to the subsequent evolution of the gene network.

We can then speculate that the absence of excess Bifan motifs could be then explained if couples of duplicated TFs which interact with each other had a much higher probability of being retained. Note that this evolutionary dynamics must be WGD-specific, since SSD couples present exactly the opposite pattern. This hypothesis is also supported by the results shown in Fig. 3.11.

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 [57]. 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.

4.4 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* [57, 39] and in *A. thaliana* [10]. 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-transcriptional layers of regulation, such as miRNA regulation, that are much less developed in simpler organisms like 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 transcriptional 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 [60, 59, 79]. 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.

4.5 An example of WGD importance: The RAR/RXR pathway

In this section we discuss more in depth the RAR/RXR pathway (schematized in Fig. 4.1A and 4.1B), a tangible example that hints at the importance of WGD events in contributing to the evolution of complex traits in vertebrates. The pathway is composed by four sets of WGD-derived genes: the RAR family (RARA,RARB,RARG), the RXR family (RXRA,RXRB,RXRG), the NCOA family (NCOA1,NCOA2,NCOA3) and the

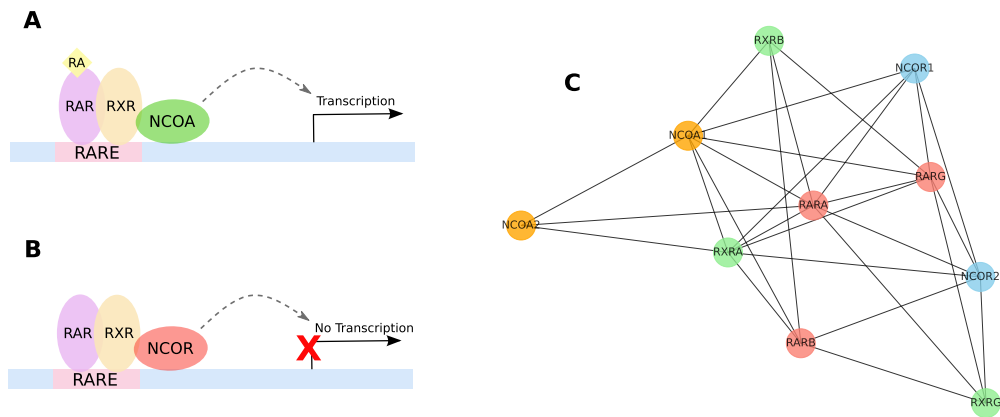


Fig. 4.1 RAR/RXR pathway is an example of WGD importance. (A) In presence of Retinoic Acid (RA) the RAR/RXR complex recruits Nuclear CoActivators (NCOA), activating the transcription of the downstream gene. (B) In absence of RA the RAR/RXR complex recruits Nuclear CoRepressors (NCOR) instead, blocking the transcription of the downstream gene. (C) Protein-protein interactions among the genes involved in the RAR/RXR pathway, common interactions with other genes are not shown for clarity. Genes colored in the same way are WGD copies of a common ancestral gene.

NCOR family (NCOR1, NCOR2). They densely interact among themselves at the protein level (see Fig. 4.1C) and they co-interact with a host of other genes. The choice of this particular example is due both to its central role in the embryonic development of vertebrates and to the high statistical significance of the number of common interactions of the genes involved. An hypergeometric test conducted on the overlap of interactions between any two of the considered genes gives a p-value smaller than $1e - 35$.

The Retinoic Acid Receptors (RAR) genes are nuclear receptors of the Retinoic Acid (RA) which is a metabolite of retinol (vitamin A). They form heterodimeric complexes with the Retinoid X Receptors (RXR), which then target a DNA binding sequence known as Retinoic Acid Response Element (RARE) and act as transcriptional regulators for a host of target genes. Binding of the RAR/RXR complex at the RARE site induces the recruitment of either the Nuclear Receptor CoActivators (NCOA), in presence of the retinoic acid (Fig. 4.1A), or the Nuclear Receptor CoRepressors (NCOR), when RA is absent (Fig. 4.1B), thus directly activating or repressing the transcription of the target genes. The RAR pathway is known to be involved in the formation of the body axis and is essential for the development of several organs including the hindbrain, the spinal cord, the skeleton, the heart, the eye, the pancreas, the lung and the reproductive tract. Importantly, it plays a prominent role in the development of the central nervous system. It mediates the anteroposterior regionalization, by regulating the transcription of Hox genes and subsequently stimulating neurogenesis and promoting neuronal differentiation. For an in-depth review of the RAR/RXR pathway see for instance [80] and

references therein.

The three components of the RAR family, *RAR α* (RARA), *RAR β* (RARB) and *RAR γ* (RARG), happen to be WGD copies of an ancestral RAR gene. Such ancestral gene can still be found in several non-chordate organisms like, for instance, anellids and mollusks [81, 82]. By comparing the ligand affinity and the expression patterns of the three versions of the vertebrates' RARs with that of the ancestral RAR, it has been shown that the ancestral RAR has a much lower affinity with its ligand with respect to the vertebrate RARs[81, 82] and that each of the three WGD copies of RARs evolved to gain a different ligand specificity and expression pattern [83]. At the same time, we observe a large number of common interactions among the three, sign that a significant regulatory redundancy has nonetheless been retained.

The scenario which emerges from these observations (for a thorough discussion see for instance [82]) is that before the WGD event the ancestral RAR was only involved in neuronal differentiation, with no involvement in spatial patterning. After WGD on the other hand, thanks to the higher affinity with the ligand and to the specificity of the binding interactions, the RAR system developed the ability of reading the spatial distribution of the RA. In particular the RAR pathway became, via the regulation of the Hox genes, the controller of the anteroposterior patterning in chordates. Evidently, this important gain of functionality is connected to the two rounds of WGD that created redundant copies of the genes involved in the pathway.

There is one last interesting fact to notice in connection to our discussion on the role of WGD events in the evolution of the RAR/RXR pathway. The anteroposterior patterning must be ultimately due to an increased complexity of the spatial distribution of the RA, otherwise the increased ability of the RAR system to read the RA distribution would have been useless. Such non trivial spatial organization requires an articulated degradation machinery for the RA. This degradation is performed in vertebrates by the CYP26 family, which is also composed by a triplet of WGD-derived paralogues, namely the CYP26A1, CYP26B1, CYP26C1 genes[84]. This fact, once again, strongly points to a fundamental role of WGD duplications in the evolution of some complex vertebrate traits.

4.6 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) require a careful control over possible spurious signals. The first necessary control is that the three gene classes do not differ significantly in the distribution of the number of interactions, since this could affect the motif statistics. In all of the networks we studied, all kinds of genes (duplicated and not) follow the same degree distribution. It

is only when genes are combined into network motifs that we can really appreciate the effects of different gene duplications on the topology of the networks. The observations in Fig. 3.1 and Fig. 3.2 ensure that the observed network motif enrichments are not spuriously produced by anomalies in the degree distributions, but are instead robust signatures of the statistical significance of the interactions among the elements of the motif.

In the next two sections we expose two other very important checks that confer further robustness to our analyses and greater confidence in the soundness of the patterns we observe. In the first section we will show that duplication age indeed has an effect on the interaction similarity between SSD couples, supporting our choice of excluding recent SSD duplicates from our analyses, as explained in the *Methods* chapter. In the other one we will show that our results on mixed-type motifs are robust with respect to the choice of different networks of miRNA-gene and protein-protein interactions and to null models constructed from different networks.

4.6.1 Interaction similarity is influenced by duplication age

In this section we assess the effects of duplication age on the retention of interaction similarity in SSD paralogues. We divided the SSD paralogues into two different age classes, based on the estimated node in the phylogenetic tree where the couple was generated. We considered SSD couples generated before the appearance of *Sarcopterygii* as roughly contemporary to the WGD-generated couples. We then compared their interaction similarity distribution, calculated in different networks for different kinds of interactions, to the one of SSD couples generated in more recent times.

As we can easily observe in Fig. 4.2, younger SSD couples present interaction similarities that are higher or much higher than the ones showed by older duplicates. We can hypothesize that this is due to the fact that the binding sequences of the two young duplicates did not have enough time to diverge significantly. For this reason, including such young couples in our analysis would have introduced an unwanted bias, which is the reason why we excluded them in the first place, as explained in the *Methods* chapter.

As a side note, it is also interesting to notice that the bias introduced by the presence of younger SSD couples would have been negligible, at least when considering similarity distributions. This can be seen in Fig. 4.2, since the similarity distributions of the *pre-Sarcopterygii* SSD couples and the overall distribution of similarities are almost indistinguishable. While the more recent SSD couples present a higher degree of similarity, in fact, they are also in a much smaller number compared to the older duplicates. This numerical imbalance is well represented in Fig. 2.1. For this this reason, their presence would not have hampered our analyses in any case.

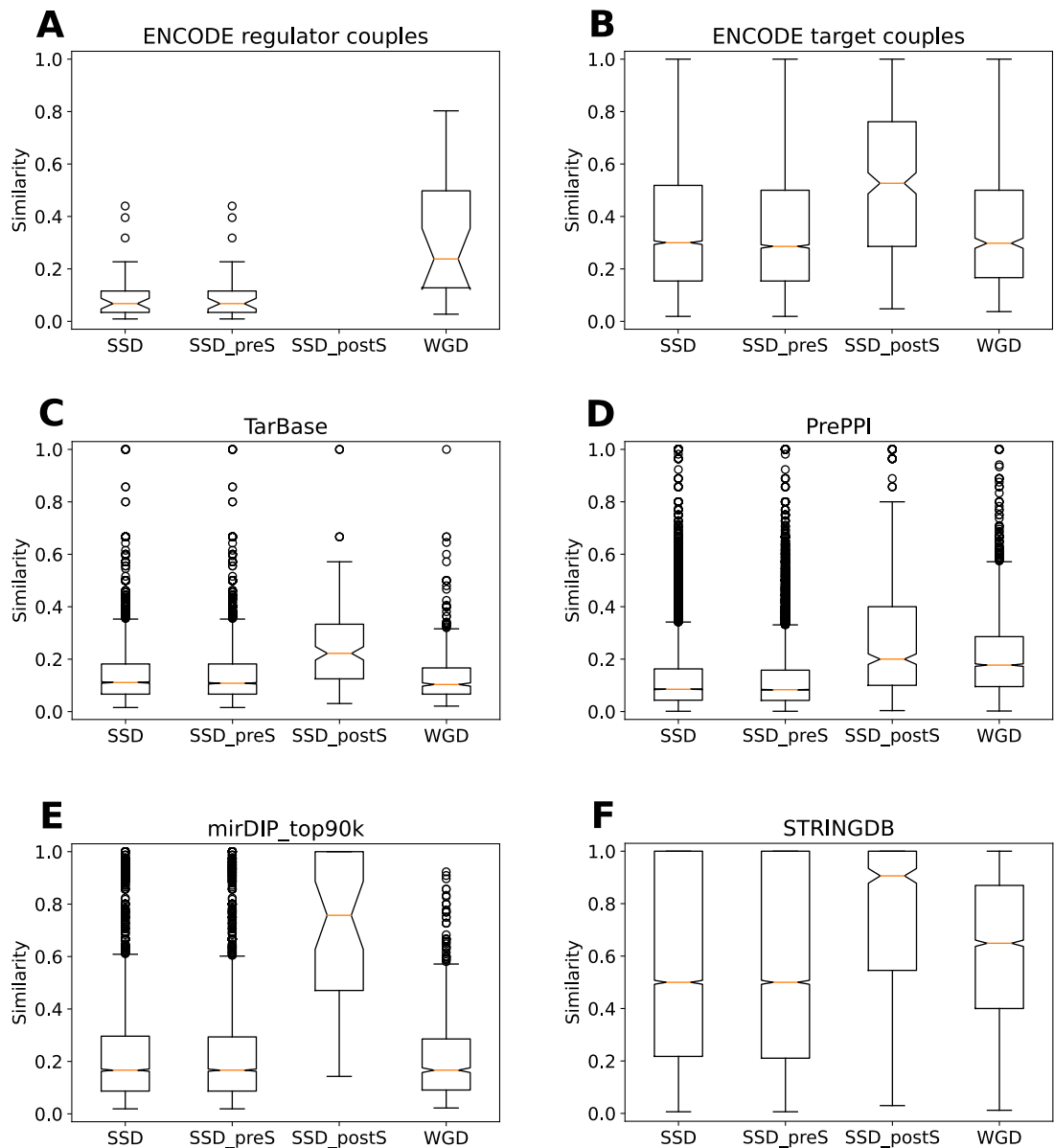


Fig. 4.2 The effect of SSD duplication age on interaction similarity. For each network considered in this work, we compare interaction similarity distributions of, from left to right in the plots: all of the SSD couples, SSD couples roughly contemporary to WGD events, younger SSD couples and WGD couples. Transcriptional interactions are further divided into (A) TF couples and (B) target couples. Notice on the hand end the striking difference between young and old duplicates and, on the other hand, the lack of appreciable differences between the overall SSD distributions and the distributions for the older copies only.

4.6.2 Mixed-type motifs enrichments with different null models

in the case of mixed-type motifs, there is no obvious way to define a single null model with respect to which one needs to calculate motif enrichments. In our case in particular, is not clear which of the different networks participating in the definition of the

motif we would need to reshuffle. We therefore calculated different Z-scores for each mixed-type motif, each obtained by shuffling one of the networks participating in the motif and keeping the others fixed. Moreover, 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). This generates a wealth of possible combinations of networks that have to be checked for each of the considered motifs. Take for example a miRNA-mediated Δ motif, as shown in Fig. 4.3B. We need to check four possible combinations of networks: TarBase-PrePPI, TarBase-STRING, mirDIP-PrePPI and mirDIP-STRING. For each of these combinations we calculate two different Z-scores, one obtained by shuffling the miRNA-gene network and the other obtained by shuffling the protein-protein interaction network. The results we obtain when carrying out this procedure for every mixed-type motif are reported in figure 4.3.

Despite significant differences both in the genes and in the interactions found in the different databases, we found quite consistent enrichment patterns. This observation is quite reassuring, boosting our confidence in the fact that the effects we see are a real biological signal and are not simply due to experimental artifacts or to some other bias.

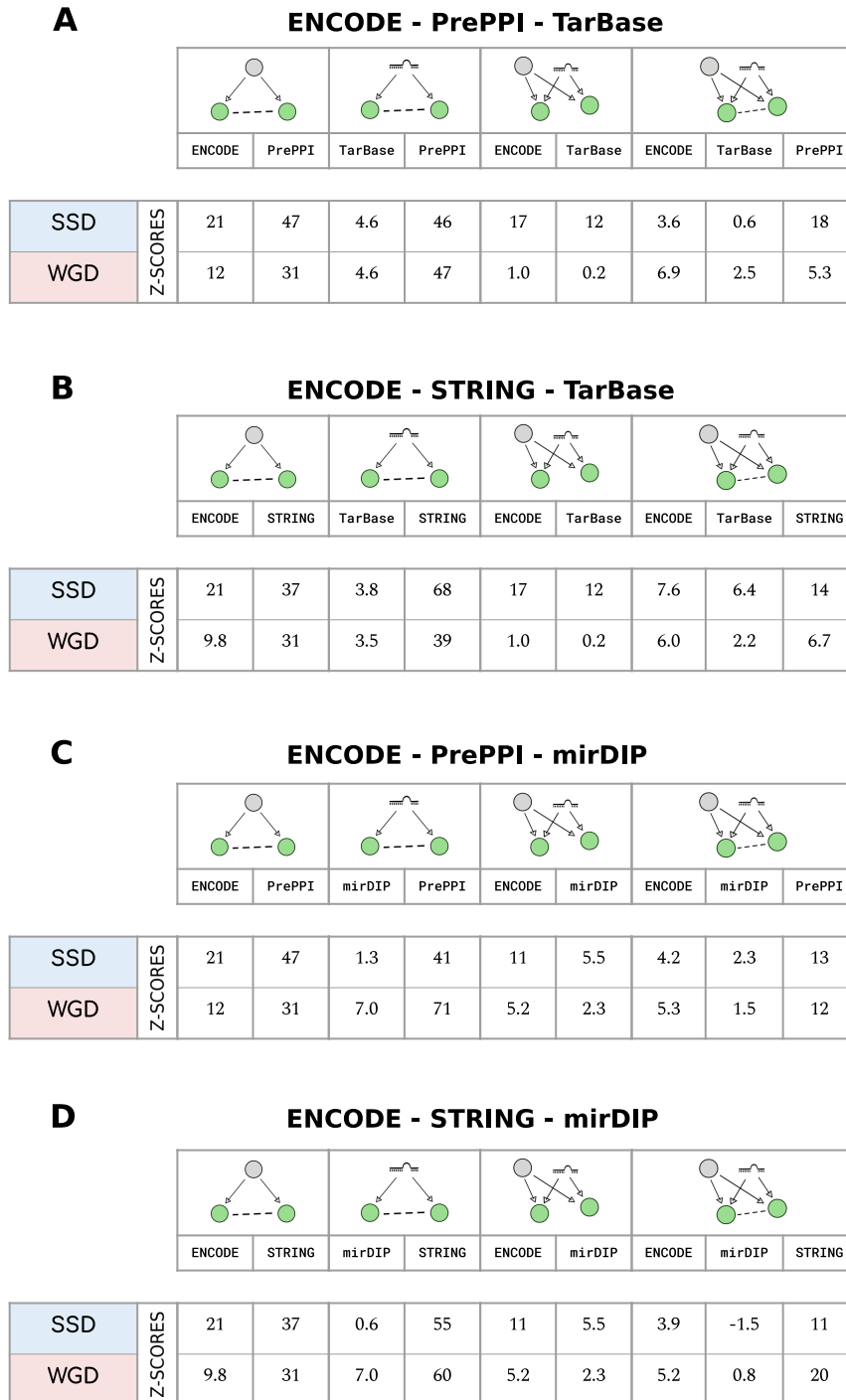


Fig. 4.3 Motifs with mixed-type regulatory interactions (different null models and network combinations). Each table is referred to a different combination of databases, always comprising transcriptional, miRNA-gene and protein-protein interactions. In each table, from left to right: Transcriptional Δ motifs, miRNA-mediated Δ motifs, Mixed Bifan motifs, in which a pair of target genes are regulated both by a common TF and a common miRNA, and Mixed Bifan motif in which the two target genes interact at the protein level. The Z-scores are referred to the randomization of the network indicated in the column header. The table shown in the main text is a subset of (A).

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, however, is still to be fully understood. Thanks to the recently published lists of WGD pairs [21, 18, 20], we had the possibility of tackling this problem. This thesis quantifies for the first time 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 and peculiarly shaped by the two rounds of WGD at the beginning of the vertebrate lineage. These two ancient events seem to have played a central role in promoting the impressive plasticity and complexity of the regulatory networks of vertebrates.

In order to support these conclusions, we studied how small-scale duplications (SSDs) and whole-genome duplications (WGDs) differently affected the statistics of simple building blocks of complex networks, that go under the name of *network motifs*. We considered, in particular, three different layers of gene regulatory interactions in human, namely transcriptional regulation, protein-protein interactions and miRNA-gene interactions.

At the transcriptional level, we showed that both SSD and WGD events played a significant role in promoting target redundancy. We argued that the retention of such redundancy could be connected both to its role in facilitating the satisfaction of dosage balance constraints applied to the targets and to the improved stochastic stability of the resulting gene circuits. We also showed that, on the contrary, WGD events seem to have played a much more important role in the retention of transcription factor redundancy with respect to SSD events. Regulatory redundancy increases robustness against mutations and facilitates the implementation of articulated gene regulatory strategies, therefore playing an important role in increasing the complexity of transcriptional programs.

WGD events also display a very strong preference for the retention of self-interactions and mutual transcriptional interactions between duplicates, although this result could

not be statistically validated. Gene circuits featuring these kind of interactions present a set of desirable properties, such as reduced response times and bi-stable toggle-switch-like behavior, fundamental for the implementation of decision-making strategies. Lastly, we showed how WGD events promoted the retention of feed-forward loop (FFL) configurations and their coupling with Bifan structures. This coupling allows to combine the ability to generate temporal activation programs typical of Bifans with the non-linear response to input signals typical of FFLs, opening up many opportunities for the development of complex regulatory strategies. On the contrary, SSD events favoured the retention of the less complex Bifan motifs.

At the protein-protein interaction level, we showed that WGD events favoured the retention of contacts between two duplicated genes in a much stronger way with respect SSD events, indicating a stronger tendency of WGD duplicates to participate in the same protein complexes. In the same spirit, we observed a similar preference for WGD couples to form complexes with many common co-interactors. At the miRNA-gene interaction level, instead, we observe a generic tendency for duplicated couples to be regulated by many common miRNAs, although there seem to be no big difference connected to the two different duplication processes.

By looking at motifs composed of interactions at different levels, that we refer to as "mixed-type motifs", we conclude that both SSD and WGD duplications acted in the direction of promoting synergy between different regulatory mechanisms, although in different ways. These kind of mixed interactions are known to be prominent in the implementation of fundamental regulatory functions, such as noise buffering and fold-change detection. In the end, we discussed a concrete example of the importance of WGD events in the evolution of regulatory networks. We considered, specifically, the effects of whole-genome duplications on the RAR/RXR pathway, showing how such events allowed to develop a refined control over the spatial distribution of retinoic acid, leading to an increase in the organisms' complexity.

Overall, our results support the hypothesis that whole-genome duplicated paralogues follow a different evolutionary trajectory with respect to small-scale duplicated paralogues. They shaped the topology of regulatory networks at many different interaction levels, favouring the retention of network motifs that are typically associated to complex functions and to a more refined control of gene expression levels.

REFERENCES

- [1] Susumu Ohno. *Evolution by Gene Duplication*. Springer-Verlag, Berlin Heidelberg, 1970.
- [2] Jianzhi Zhang. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18(6):292–298, 2003.
- [3] Jeffery P. Demuth and Matthew W. Hahn. The life and death of gene families. *BioEssays*, 31(1):29–39, 2009.
- [4] Yves Van de Peer, Steven Maere, and Axel Meyer. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, 10(10):725–732, 2009.
- [5] Yves Van de Peer, Eshchar Mizrachi, and Kathleen Marchal. The evolutionary significance of polyploidy. *Nat. Rev. Genet.*, 18(7):411–424, 2017.
- [6] Luca Comai. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, 6(11):836–846, 2005.
- [7] Karen D. Crow and Günter P. Wagner. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, 23(5):887–892, 2005.
- [8] Kenneth H. Wolfe and Denis C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713, 1997.
- [9] Manolis Kellis, Bruce W. Birren, and Eric S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624, 2004.
- [10] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815, 2000.
- [11] Paramvir Dehal and Jeffrey L. Boore. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLOS Biology*, 3(10):e314, 2005.

- [12] Ferdinand Marlétaz, Panos N. Firbas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry, Christopher D. R. Wyatt, Elisa de la Calle-Mustienes, Stephanie Bertrand, Demian Burguera, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature*, 564(7734):64–70, 2018.
- [13] Matteo D’Antonio and Francesca D. Ciccarelli. Modification of Gene Duplicability during the Evolution of Protein Interaction Network. *PLOS Comput. Biol.*, 7(4):e1002029, 2011.
- [14] Lex E. Flagel and Jonathan F. Wendel. Gene duplication and evolutionary novelty in plants. *New Phytol*, 183(3):557–564, 2009.
- [15] Feng Cheng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants*, 4(5):258–268, 2018.
- [16] Hui Guo, Tae-Ho Lee, Xiyin Wang, and Andrew H. Paterson. Function Relaxation Followed by Diversifying Selection after Whole-Genome Duplication in Flowering Plants . *Plant Physiol*, 162(2):769–778, 2013.
- [17] Yoichiro Nakatani, Hiroyuki Takeda, Yuji Kohara, and Shinichi Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, 17(9):1254–1265, 2007.
- [18] Param Priya Singh, Jatin Arora, and Hervé Isambert. 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, 2015.
- [19] Debarun Acharya and Tapash C. Ghosh. 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, 2016.
- [20] Param Priya Singh and Hervé Isambert. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res.*, 48(D1):D724–D730, 2020.
- [21] Takashi Makino and Aoife McLysaght. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. U.S.A.*, 107(20):9270–9274, 2010.
- [22] Param Priya Singh, Séverine Affeldt, Ilaria Cascone, Rasim Selimoglu, Jacques Camonis, and Hervé Isambert. On the expansion of ”dangerous” gene repertoires

- by whole-genome duplications in early vertebrates. *Cell Rep.*, 2(5):1387–1398, 2012.
- [23] Tine Blomme, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, 7(5):R43, 2006.
- [24] Michael Freeling and Brian C. Thomas. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.*, 16(7):805–814, 2006.
- [25] Lukasz Huminiecki and Carl Henrik Heldin. 2R and remodeling of vertebrate signal transduction engine. *BMC Biology*, 8(1):146, 2010.
- [26] Yuanfang Guan, Maitreya J. Dunham, and Olga G. Troyanskaya. Functional Analysis of Gene Duplications in *Saccharomyces cerevisiae*. *Genetics*, 175(2):933–943, 2007.
- [27] Steven Maere, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, 102(15):5454–5459, 2005.
- [28] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, Feb 2021.
- [29] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, Feb 2004.
- [30] A. Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences of the United States of America*, 91(10):4387–4391, May 1994. 8183919[pmid].
- [31] Sarah A. Teichmann and M. Madan Babu. Gene regulatory network growth by duplication. *Nat. Genet.*, 36(5):492–496, 2004.
- [32] Austin L. Hughes and Robert Friedman. Gene duplication and the properties of biological networks. *Journal of molecular evolution*, 61(6):758–764, Dec 2005. 16315107[pmid].
- [33] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

- [34] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, 31(1):64–68, 2002.
- [35] Laurie A. Boyer, Tong Ihn Lee, Megan F. Cole, Sarah E. Johnstone, Stuart S. Levine, Jacob P. Zucker, Matthew G. Guenther, Roshan M. Kumar, Heather L. Murray, Richard G. Jenner, David K. Gifford, Douglas A. Melton, Rudolf Jaenisch, and Richard A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005.
- [36] Lea Goentoro, Oren Shoval, Marc W. Kirschner, and Uri Alon. The incoherent feedforward loop can provide fold-change detection in gene regulation. *Molecular Cell*, 36(5):894–899, 2009.
- [37] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.
- [38] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2019.
- [39] Diana Fusco, Luigi Grassi, Bruno Bassetti, Michele Caselle, and Marco Cosentino Lagomarsino. Ordered structure of the transcription network inherited from the yeast whole-genome duplication. *BMC Syst. Biol.*, 4:77, 2010.
- [40] Mark B. Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G. Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.
- [41] Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, 29(8):1363–1375, 2019.
- [42] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, Sungho Lee, Byunghee Kang, Dabin Jeong, Yaeji Kim, Hyeon-Nae Jeon, Haein Jung, Sunhwee Nam, Michael Chung, Jong-Hoon Kim, and Insuk Lee. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, 46(D1):D380–D386, 2018.

- [43] Luiz A. Bovolenta, Marcio L. Acencio, and Ney Lemke. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, 13:405, 2012.
- [44] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.
- [45] Qiangfeng Cliff Zhang, Donald Petrey, José Ignacio Garzón, Lei Deng, and Barry Honig. PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.*, 41(Database issue):D828–833, 2013.
- [46] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47(D1):D607–D613, 2019.
- [47] Dimitra Karagkouni, Maria D. Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S. Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, Giorgos Skoufos, Thanasis Vergoulis, Theodore Dalmagas, and Artemis G. Hatzigeorgiou. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.*, 46(D1):D239–D245, 2018.
- [48] Tomas Tokar, Chiara Pastrello, Andrea E. M. Rossos, Mark Abovsky, Anne-Christin Hauschild, Mike Tsay, Richard Lu, and Igor Jurisica. mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Res.*, 46(D1):D360–D370, 2018.
- [49] Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, 49(D1):D939–D946, 2021.
- [50] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2021. *Nucleic Acids Res.*, 49(D1):D884–D891, 2021.
- [51] Arianna Bottinelli, Bruno Bassetti, Marco Cosentino Lagomarsino, and Marco Gherardi. Influence of homology and node age on the growth of protein-protein interaction networks. *Phys. Rev. E*, 86:041919, Oct 2012.

- [52] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.
- [53] Sebastian Wernicke and Florian Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.
- [54] Gavin C. Conant and Kenneth H. Wolfe. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.*, 9(12):938–950, 2008.
- [55] Luke Hakes, John W. Pinney, Simon C. Lovell, Stephen G. Oliver, and David L. Robertson. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.*, 8(10):R209, 2007.
- [56] Antonio Rosanova, Alberto Colliva, Matteo Osella, and Michele Caselle. Modelling the evolution of transcription factor binding preferences in complex eukaryotes. *Sci. Rep.*, 7(1):7596, 2017.
- [57] Jonathan J. Ward and Janet M. Thornton. Evolutionary models for formation of network motifs and modularity in the *Saccharomyces* transcription factor network. *PLoS Comput. Biol.*, 3(10):1993–2002, 2007.
- [58] M. Cosentino Lagomarsino, P. Jona, B. Bassetti, and H. Isambert. Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proceedings of the National Academy of Sciences*, 104(13):5516–5520, 2007.
- [59] Matteo Osella, Carla Bosia, Davide Corá, and Michele Caselle. The role of incoherent microRNA-mediated feedforward loops in noise buffering. *PLoS Comput. Biol.*, 7(3):e1001101, 2011.
- [60] Andrea Riba, Carla Bosia, Mariama El Baroudi, Laura Ollino, and Michele Caselle. A combination of transcriptional and microRNA regulation improves the stability of the relative concentrations of target genes. *PLoS Comput. Biol.*, 10(2):e1003490, 2014.
- [61] A. Stoltzfus. On the possibility of constructive neutral evolution. *J. Mol. Evol.*, 49(2):169–181, 1999.
- [62] Wenfeng Qian, Ben-Yang Liao, Andrew Ying-Fei Chang, and Jianzhi Zhang. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.*, 26(10):425–430, 2010.
- [63] Gavin C. Conant, James A. Birchler, and J. Chris Pires. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.*, 19:91–98, 2014.

- [64] Elena Kuzmin, Benjamin VanderSluis, Alex N. Nguyen Ba, Wen Wang, Elizabeth N. Koch, Matej Usaj, Anton Khmelinskii, Mojca Mattiazzi Usaj, Jolanda van Leeuwen, Oren Kraus, Amy Tresenrider, Michael Prysxlak, Ming-Che Hu, Brenda Varriano, Michael Costanzo, Michael Knop, Alan Moses, Chad L. Myers, Brenda J. Andrews, and Charles Boone. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science*, 368(6498):eaaz5667, 2020.
- [65] Michael J. Song, Barney I. Potter, Jeff J. Doyle, and Jeremy E. Coate. Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in *Arabidopsis thaliana*. *Plant Cell*, 32(5):1434–1448, 2020.
- [66] James Birchler and Reiner Veitia. Protein–protein and protein–dna dosage balance and differential paralog transcription factor retention in polyploids. *Front Plant Sci*, 2:64, 2011.
- [67] Michaël Bekaert, Patrick P. Edger, J. Chris Pires, and Gavin C. Conant. Two-Phase Resolution of Polyploidy in the *Arabidopsis* Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. *Plant Cell*, 23(5):1719–1728, 2011.
- [68] Xun Lan and Jonathan K. Pritchard. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288):1009–1013, 2016.
- [69] Jonas Ibn-Salem, Enrique M. Muro, and Miguel A. Andrade-Navarro. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Res.*, 45(1):81–91, 2017.
- [70] Wenfeng Qian and Jianzhi Zhang. Gene dosage and gene duplicability. *Genetics*, 179(4):2319–2324, 2008.
- [71] M. Madan Babu, Nicholas M. Luscombe, L. Aravind, Mark Gerstein, and Sarah A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, 14(3):283–291, 2004.
- [72] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.
- [73] Michael Lynch and John S. Conery. The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics*, 3(1-4):35–44, 2003.
- [74] Tanya Vavouri, Jennifer I. Semple, and Ben Lehner. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.*, 24(10):485–488, 2008.

- [75] Zhenglong Gu, Lars M. Steinmetz, Xun Gu, Curt Scharfe, Ronald W. Davis, and Wen-Hsiung Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66, 2003.
- [76] Christopher R. Baker, Victor Hanson-Smith, and Alexander D. Johnson. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, 342(6154):104–108, 2013.
- [77] Michal Chapal, Sefi Mintzer, Sagie Brodsky, Miri Carmi, and Naama Barkai. Resolving noise-control conflict by gene duplication. *PLoS Biol.*, 17(11):e3000289, 2019.
- [78] Johan Hallin and Christian R. Landry. Regulation plays a multifaceted role in the retention of gene duplicates. *PLoS Biol.*, 17(11):e3000519, 2019.
- [79] Carla Bosia, Matteo Osella, Mariama El Baroudi, Davide Corà, and Michele Caselle. Gene autoregulation via intronic microRNAs and its functions. *BMC Syst. Biol.*, 6:131, 2012.
- [80] Karen Niederreither and Pascal Dolle. Retinoic acid in development: towards an integrated view. *Nat Rev Genet*, 9(7):541–553, 2008.
- [81] Juliana Gutierrez-Mazariegos, Eswar Kumar Nadendla, Daniela Lima, Keely Pierzchalski, Jace W. Jones, Maureen Kane, Jun-Ichi Nishikawa, Youhei Hiro-mori, Tsuyoshi Nakanishi, Miguel M. Santos, L. Filipe C. Castro, William Bourguet, Michael Schubert, and Vincent Laudet. A Mollusk Retinoic Acid Receptor (RAR) Ortholog Sheds Light on the Evolution of Ligand Binding. *Endocrinology*, 155(11):4275–4286, 2014.
- [82] Mette Handberg-Thorsager, Juliana Gutierrez-Mazariegos, Stefan T. Arold, Eswar Kumar Nadendla, Paola Y. Bertucci, Pierre Germain, Pavel Tomancak, Keely Pierzchalski, Jace W. Jones, Ricard Albalat, Maureen A. Kane, William Bourguet, Vincent Laudet, Detlev Arendt, and Michael Schubert. The ancestral retinoic acid receptor was a low-affinity sensor triggering neuronal differentiation. *Sci Adv*, 4(2), 2018.
- [83] Hector Escriva, Stephanie Bertrand, Pierre Germain, Marc Robinson-Rechavi, Muriel Umbhauer, Jerome Cartry, Marilynne Duffraisse, Linda Holland, Hinrich Gronemeyer, and Vincent Laudet. Neofunctionalization in vertebrates: The example of retinoic acid receptors. *PLoS Genetics*, 2(7):955–965, 2006.
- [84] Yasuo Sakai, Chikara Meno, Hideta Fujii, Jinsuke Nishino, Hidetaka Shiratori, Yukio Saijoh, Janet Rossant, and Hiroshi Hamada. The retinoic acid-inactivating

enzyme cyp26 is essential for establishing an uneven distribution of retinoic acid along the antero-posterior axis within the mouse embryo. *Genes Dev*, 15(2):213–225, 2001.