Chapter 19

Single-Gene and Whole-Genome Duplications and the Evolution of Protein–Protein Interaction Networks

Grigoris Amoutzias and Yves Van de Peer

19.1 INTRODUCTION

Proteins within a cell do not function in isolation, but instead physically interact with their molecular environment, either to transduce information from the external environment to the nucleus or to form multisubunit protein complexes that act as sophisticated molecular

Evolutionary Genomics and Systems Biology, edited by Gustavo Caetano-Anollés

machines. Since the functionality of the cell depends on these physical interactions, it is no surprise that great effort is being made in cataloguing the interactome of a genome, that is, to identify and describe all protein–protein interactions (PPIs) a protein participates in. In this era of "omics" technologies and systems biology, researchers try to deal with the interactome of a given organism in a holistic approach, and try to reconstruct protein–protein interaction networks (PINs), using graph theory (Barabasi and Oltvai, 2004). In such networks, proteins are represented as nodes in a graph, and edges connect nodes that physically interact, or nodes that participate in the same complex. There is a significant amount of work on the principles of PINs, their statistical properties, and their significance for the cell (Barabasi and Oltvai, 2004), but the focus of this review is on the evolution of PPIs and more particularly the contribution of two major sources of molecular innovation, namely, single-gene and whole-genome duplications.

It is important to understand the evolution of PPIs in order to address fundamental questions about molecular biology and to use the interactome correctly. First of all, we need to understand which molecular mechanisms are responsible for innovation and the evolution of PINs, and the extent of contribution of each one of those mechanisms. Also, PINs from different organisms need to be compared, in order to understand which are the universal core protein complexes, which protein complexes are specific to a certain clade of organisms, and which are unique to one species. In this way, we will know which experimentally determined interactions can be transferred from one organism to another and which interactions are not transferable. In addition, by studying the evolution of PPIs and PINs, we will better understand the components and types of interactions that are responsible for increasing biological complexity. It is well acknowledged that organismal complexity correlates with the number and coverage of PPI domains per protein (Xia et al., 2008). Complexity also seems to correlate with an expansion of certain gene families (van Nimwegen, 2003). We need to understand whether these particular families are linked to specific types of interactions and to improve our knowledge on the relationship between organismal complexity, the various modes of duplication, and PPIs.

Here, we will first introduce the sources of molecular innovation in PINs, that is, through gene and genome duplications and mutations. Second, we will discuss and review studies from genome-scale data that provide a bird's eye view about the importance of each source of molecular innovation, and finally, we will cite some medium-scale studies that use high-quality data and provide an in-depth view about the impact of gene/genome duplication on the evolution of PPIs.

19.2 EVOLUTION OF PINs

During the last decade, the significance of gene duplications, point mutations, and domain rearrangements in shaping the regulatory and protein interaction networks has been well established (Amoutzias et al., 2004b; Babu et al., 2004; Bornberg-Bauer et al., 2005; Evlampiev and Isambert, 2008; Ispolatov et al., 2005; Pastor-Satorras et al., 2003; Wagner, 1994). PINs may evolve by two mechanisms, either by mutations such as point mutations or/and domain rearrangements on existing proteins or by gene duplication and subsequent mutations of the duplicate/s. In the former case, the number of nodes in the PIN remains stable, but the PIN is actually rewired, as some new PPI interfaces will emerge and some will be lost. In the later case, a gene duplicates, and one (or both) copy(ies) may undergo one of three fates (Figure 19.1), namely, (i) subfunctionalization, where the functions of the ancestral gene are divided among the two duplicates; (ii) neofunctionaliza-

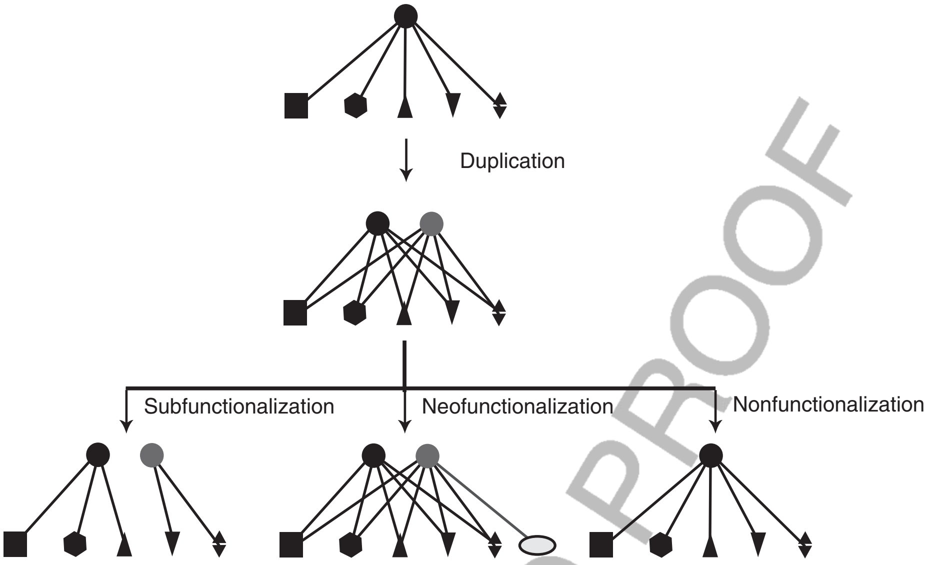

Figure 19.1 The three most common fates of a duplicated gene and its interactions. (See insert for color representation of this figure.)

tion, where one of the copies may retain the ancestral function but the other evolves a novel function; and (iii) or most frequently, nonfunctionalization, where one of the copies accumulates deleterious mutations and turns into a pseudogene. In any of the two cases where the two copies survive (subfunctionalizaton/neofunctionalization), all the initial interactions of the ancestral gene's product, that is, protein are inherited by the identical duplicate and then, depending on the extent and character of mutations, a few or all of the common interactions among the duplicates may be retained or lost, whereas some other new interactions may also emerge. Mutations on existing genes or mutations on redundant duplicates may take place at the same time in different parts of the network. Nevertheless, the latter case of gene duplication and subsequent divergence is strongly supported by genomic data, as a major contributor of molecular innovation. For example, in yeast, only one-third of the genes are characterized as singletons (Davis and Petrov, 2005). Analyses of PPIs among gene duplicates in yeast show that the duplicates diverge asymmetrically in terms of PPIs (Wagner, 2002), meaning that one of the duplicates has more PPIs than the other, although they still retain a significant number of common interactors, more than expected by chance (Musso et al., 2007).

We can consider different types of gene duplication events, depending on how many genes are duplicated simultaneously: single-gene duplications (SGDs), block duplications, and whole-genome duplications (WGDs). Block duplications are observed in cases of trisomies or chromosomal aneuploidies, where one whole-genomic block, such as a whole chromosome or part of it may duplicate or become lost. The strong negative effect of these block duplications on the fitness of individuals, as well as simulations on the evolution of genetic networks show that organisms should preferentially evolve either by single-gene or by whole-genome duplications (Wagner, 1994). These two mechanisms of duplication (SGDs/WGDs) are totally different in nature and it has been proposed that they should favor the duplication and survival of different types of genes (Davis and Petrov, 2005).

19.3 SINGLE-GENE DUPLICATIONS

Single-gene duplications occur continuously within a genome. (Lynch and Conery, 2000) describe them as stochastic processes that are being fixed in a population with a frequency of up to (differs for different species) 1 out of 100 genes per million years. As noted, the most common fate of gene duplication is nonfunctionalization, where the duplicate gene will be rapidly lost. The average half-life of a single-gene duplicate has been estimated to be approximately 4 million years (Lynch and Conery, 2000). It is estimated that half the genes of a genome will be duplicated and fixed, within a timescale of 35 to 350 million years (Lynch and Conery, 2000).

19.4 WHOLE-GENOME DUPLICATIONS

Whole-genome duplications were initially thought to happen very frequently, but become fixed very rarely. Therefore, their impact on evolution has long been underestimated. Nevertheless, the advent of the genomic era revealed that actually, the fixation of WGDs is more frequent than originally thought and of major significance for speciation, radiation, and adaptation (De Bodt et al., 2005; Scannell et al., 2006; Van de Peer, 2004). Initially, this idea was hotly debated and opposed, but now it is widely accepted that almost all eukaryotic lineages such as animals, fungi, protists, and especially plants have undergone one or more rounds of WGDs in their evolutionary past. For example, in animals, two successive rounds of WGDs occurred at the origin of vertebrates (the 2R event) (Dehal and Boore, 2005; Panopoulou and Poustka, 2005) and one in the bony fish lineage (the 3R event) (Jaillon et al., 2004; Taylor et al., 2003; Vandepoele et al., 2004). In the yeast lineage, a WGD occurred around 100 million years ago (Wolfe and Shields, 1997), whereas in the ciliate Paramecium, 3 or 4 WGDs have occurred (Aury et al., 2006). In plants, one or two genome duplications are shared between all flowering plants, whereas many of them underwent additional rounds of polyploidization (Blanc and Wolfe, 2004; Cui et al., 2006; De Bodt et al., 2005; Schlueter et al., 2004; Sterck et al., 2007).

19.5 DIPLOIDIZATION PHASE

During a WGD event, either through autopolyploidy or through allopolyploidy (hybridization), all genes of a genome duplicate simultaneously and the organism appears as tetraploid. The duplicate pairs that result from such an event are called ohnologues, after Susumu Ohno, who was the first to discuss the importance of gene and genome duplications (Ohno, 1970). Usually, this tetraploid phase does not last long. Extensive genomic rearrangements and genes loss occur, as the organism returns back to its diploid state, a process called diploidization (Wolfe, 2001). During this phase, the rate of duplicate loss can be more or less constant over time, as observed in Paramecium (Aury et al., 2006) or it can be very high at the beginning and slow down later, as observed in baker's yeast (Scannell et al., 2006) and *Arabidopsis* (Maere et al., 2005). Lynch and Conery (2000) observed that the level of retention of ohnologues is unexpectedly high, compared to retention rates of single-gene duplicates. This high retention rate has been confirmed for many species that underwent one or more WGDs. Several reasons have been proposed to explain this high retention rate, such as protein dosage effects (see further), buffering of essential genes, enhancement of metabolic fluxes and rapid divergence of gene pairs (Aury et al., 2006; Chapman et al., 2006; Kondrashov et al., 2002; Lynch and Katju, 2004; Ohno, 1970; Papp et al., 2003; Veitia, 2005).

19.6 DOSAGE BALANCE HYPOTHESIS

Perhaps the most important factor for ohnologue retention that is directly linked to PPIs is the protein dosage effect. According to the dosage balance hypothesis (DBH) (Veitia et al., 2008), the stoichiometric imbalances in macromolecular complexes can have phenotypic effects, most probably fitness defects. These defects are the result of overexpression or underexpression of a protein subunit that disrupt proper formation of the complex. The DBH has also been extended to signaling and transcriptional networks, where according to theory, the balance between activators and repressors should be preserved (Birchler and Veitia, 2007). Studies have shown an overrepresentation of regulatory genes as being sensitive to haploinsufficiency (Kondrashov and Koonin, 2004; Papp et al., 2003).

Therefore, where stoichiometric balance needs to be preserved, single-gene duplications can have a detrimental effect. On the other hand, whole-genome duplications should not affect the stoichiometries, since all parts of the complex are duplicated simultaneously. To put it simply, such complexes that are sensitive to stoichiometric imbalances are bound to evolve mostly by WGDs and not by SGDs. During the diploidization phase, duplicate genes that participate in complexes and are sensitive to dosage imbalances would tend to be retained instead of being lost. Later on, mutations and their resulting genetic network rewiring will allow some of these retained duplicates to diversify or disappear, due to compensation of imbalances (Aury et al., 2006; Semon and Wolfe, 2007). There are several mechanisms that can compensate for this imbalance, at the mRNA or protein level (Veitia et al., 2008) such as the pathway and kinetics of macromolecular assembly, the topology of the complex, negative feedback regulatory loops that maintain the concentration of mRNA or protein level stable, or proteasome degradation of monomers in excess.

From the DBH alone, it is evident that the mode of duplication should have a strong effect on which categories of genes will be retained. Indeed, genes involved in signal transduction and transcription have a strong tendency to be retained after a WGD event, as has been shown for fungi, plants, and vertebrates (Blomme et al., 2006; Davis and Petrov, 2005; Maere et al., 2005). Maere et al. (2005) estimated that around two-thirds of the transcriptional regulators and half of the kinases of *Arabidopsis* are ohnologues, retained after several ancient WGD events over a period of 150 million years, whereas Blomme et al. (2006 and personal communication) estimated that half of the transcriptional regulators and two-thirds of signal transducers of human are ohnologues, retained from two genome doublings at the origin of vertebrates around 500 million years ago. It is evident that WGDs had a significant impact on these specific categories of genes, but the question remains what the impact was on the evolution of PINs. Intriguingly, TFs and kinases are participating in a certain type of interaction, termed transient.

19.7 TYPES OF INTERACTIONS

Not all interactions are of the same nature, but rather they are categorized in different types, as reviewed extensively by (Nooren and Thornton, 2003). More and more studies discuss the importance of distinguishing between the various types of interactions, in order to understand their characteristics and evolution, instead of treating them all as the same (Brown and Jurisica, 2007; Mintseris and Weng, 2005; Nooren and Thornton, 2003; Sprinzak et al., 2006; Tompa and Fuxreiter, 2008; Wilkins and Kummerfeld, 2008). One distinction is among obligate and nonobligate complexes. For instance, some proteins are not found as stable structures alone *in vivo* and take on their characteristic structure only when they become part of the complex. The complexes that they form are called obligate. On the other hand, the proteins that have a defined crystal structure, independently of their interacting partners form nonobligate complexes.

Another distinction can be made between transient and permanent complexes. The former have a short lifetime span and associate and dissociate *in vivo*, whereas the latter are usually disrupted by proteolysis. Obligate interactions are usually permanent, whereas nonobligate interactions can be either permanent or transient. Despite these classification efforts, (Nooren and Thornton, 2003) note that there is no simple and clear distinction between obligate and nonobligate interactions, but rather that there exists a continuum between both sorts of interaction. In addition, physiological conditions such as concentration of ions, chemicals, pH, temperature, or phosphorylation may affect the stability of an interaction.

Several recent studies have shown that a distinction should be made between transient and stable interactions, due to their different properties. Otherwise, the various signals of a PPI analysis are weakened or scrambled (Brown and Jurisica, 2007; Mintseris and Weng, 2005; Nooren and Thornton, 2003; Sprinzak et al., 2006; Tompa and Fuxreiter, 2008; Wilkins and Kummerfeld, 2008). One of the most important differences is that the interactions of stable complexes are highly conserved, even among distantly related organisms, whereas transient interactions are usually much less conserved (Brown and Jurisica, 2007; Mintseris and Weng, 2005). This difference is also reflected by the amino acid conservation level of the interacting interfaces (Mintseris and Weng, 2005). Furthermore, interacting partners of stable complexes tend to be more coexpressed than the partners of transient interactions (such as the yeast kinome), whose coexpression is not higher than random protein pairs (Brown and Jurisica, 2007). Interestingly, the human PIN seems to be dominated by transient interactions (Brown and Jurisica, 2007).

19.8 WGDs, TRANSIENT INTERACTIONS, AND ORGANISMAL COMPLEXITY

Phosphorylation is the reversible addition of a phosphate group by a protein kinase to a target protein, while phosphatases have the opposite effect and remove phosphate. Phosphorylation causes conformational changes in the structure of the target protein and may alter its functions. Between 30% and 50% of a proteome may be phosphorylated under certain physiological conditions, but interestingly, certain categories of genes, mainly retained following WGD (see higher), such as TFs and kinases are often overrepresented in phosphorylation sites (Chi et al., 2007; Heazlewood et al., 2008; Ptacek et al., 2005). These sites usually occur within fast evolving regions that are intrinsically disordered and lack regular structure, such as hinges and loops (Gnad et al., 2007; Iakoucheva et al., 2004; Peck, 2006). The phosphorylation motifs that are recognized by kinases are rather short, with a length of 8–12 amino acids (Gnad et al., 2007). Given their short length and their embedment within fast evolving regions, they must appear and disappear very rapidly. Indeed, regions of phosphorylation sites have lower conservation than the average conservation of the entire protein (Gnad et al., 2007), whereas yeast phosphoproteins, but not the actual phosphorylation sites are conserved across large evolutionary distances (Chi et al., 2007). Also, phosphorylation sites in plants are usually found outside of PFAM domains and they tend to be more conserved between orthologues than between paralogues (Nuhse et al., 2004; Peck, 2006).

Another fact that links WGDs to transient interactions is the enrichment of intrinsically disordered regions in TFs and kinases. Some regions in a protein do not have a well-defined and stable 3D structure in their native state, but instead have dynamic structures that interconvert. They are termed intrinsically disordered regions (IDRs) and may cover either a small part of or the whole of a protein (Lobley et al., 2007). Although they lack a well-defined structure, they are often involved in transient protein–protein interactions of regulatory and signaling molecules that require high specificity and low affinity. The enrichment of TFs in IDRs was demonstrated by Liu et al. (2006). Especially activation domains are mostly or totally unstructured. There is a preference for phosphorylation sites to be embedded within IDRs (Iakoucheva et al., 2004), whereas Lobley et al. (2007) have shown that transcriptional regulators and kinases are among the groups of proteins that are enriched in IDRs.

Given the fact that phosphorylation sites and IDRs are overrepresented in TFs and kinases and that WGD strongly favors the retention of these categories of genes, it becomes evident that this mode of duplication would result in the increase of transient interactions within a PIN. In addition, the rapid emergence and loss of phosphorylation sites by a few point mutations strongly suggests that WGD provides the raw material for rapid rewiring of a PIN, especially the part that is involved in transient interactions and information processing. Since transient interactions dominate gene regulation and signal transduction and given the established link between organismal complexity (at least in terms of distinct cell types) and an increase in the percentage of signal transducers and TFs within a genome (Ranea et al., 2005; van Nimwegen, 2003), it is tempting to assume that WGD is one of the major contributors of raw genetic material to increase biological complexity.

19.9 STUDIES ON PPIs OF OHNOLOGUES

A pair of retained gene duplicates may follow one of three scenarios: (i) the duplicates may retain the majority of functions of the ancestral gene, and show redundancy, (ii) they may subdivide the functions or expression of the ancestral molecule, that is, subfunctionalize, or (iii) one of the two duplicates retains the ancestral functions, whereas the other evolves new functions (neofunctionalization) (Casneuf et al., 2006). In the first scenario, the organism should become more robust to mutations, whereas in the other two scenarios, the organism should evolve new functions and possibly adapt better to new environments. Several studies (Guan et al., 2007; Hakes et al., 2007) have tried to address the question of whether the mode of duplication (SGD versus WGD) could be linked to one of the three aforementioned scenarios, by comparing the functional divergence between groups of ohnologues and groups of SG duplicates in yeast. As a measure of functional divergence, the number of common interactors, together with an integrated Bayesian analysis of diverse functional data was used, as well as a semantic distance based on Gene Ontology annotation. The results of both studies showed that for genes with the same level of sequence divergence, ohnologues diverge less in function and PPIs, compared to SG duplicates. In addition, ohnologues tend to be more dispensable than SG duplicates and also have higher synthetic lethality.

Although in our opinion these results need to be considered with caution (see next section), the analysis of Hakes et al. (2007) highlights important differences between SG and WG duplicates. Both SG and WG duplicates have the same connectivity, with an average of 10 PPIs per protein, but WG duplicates tend to share more common interactors than SG duplicates. Furthermore, for genes that participate in complexes, SG duplicates tend to be more essential than WG duplicates (21% versus 10%, respectively), whereas for genes not participating in complexes, both types have similar dispensability (9% versus 6%, respectively). WG duplicates tend to participate in complexes slightly more than SG duplicates (19% versus 14%).

Another difference between SG and WG duplicates that relates to their physical interactions is underwrapping. Crystal structures from PDB were analyzed to test whether duplicability of a gene is affected by its underwrapping (Liang et al., 2008). This term describes the solvent accessibility of the hydrogen bonds of the protein backbone. The less accessible these hydrogen bonds are to water molecules, the more functionally competent the structure is. This inaccessibility is achieved by clusters of nonpolar amino acids that wrap the hydrogen bond and protect it from water. Intramolecular hydrogen bonds that are accessible to water are called dehydrons and constitute structural vulnerabilities. Therefore, the more underwrapped proteins are, the more reliant on their interactive context they are in order to maintain structural integrity. Overexpression of highly underwrapped proteins could increase misfolding and aggregation, thus leading to dosage sensitivity. Therefore, the theory predicts that the more underwrapped proteins are, the more sensitive to dosage imbalances, which should also be reflected by their family sizes and mode of duplication.

Liang et al. (2008) compiled protein structures from PDB and calculated the underwrapping extent of each protein. Next, the gene family size was determined. The authors found a negative correlation between underwrapping and duplicability, showing that indeed underwrapping makes genes more sensitive to dosage effects and hinders SGDs. They also compared the yeast SG duplicates against the WG duplicates and found that WG duplicates could tolerate higher underwrapping. It was also noted that the underwrapping effect and therefore the dosage imbalance effect was strong for simple unicellular organisms, but less strong for more complex organisms. Five major reasons were suggested for this loss of sensitivity to dosage imbalance: (i) more efficient regulatory networks that can compensate for higher expression, (ii) alternative splicing as an escape route, (iii) higher allostery in complex organisms, (iv) smaller effective population size, that allows slightly deleterious dosage imbalances to become fixed, and (v) positive selection.

Another study tried to determine the effect of WGD on the homodimeric interactions. By using network motifs and a mathematical model on the gain and loss of interactions, Presser et al. (2008) analyzed the interactions among the yeast ohnologues and conclude that the pre-WGD genome had proteins that tended to self-interact, more than after the WGD. The WGD probably caused a rewiring effect. It is possible that mutations changed some of the ancient homodimers to obligate heterodimers. Other groups also suggested a model of PIN evolution, where redundant duplicate homodimers would evolve to heterodimers by mutations (Amoutzias et al., 2004b; Ispolatov et al., 2005; Pereira-Leal et al., 2007).

19.10 CONCERNS ABOUT THE METHODS OF ANALYSIS AND THE QUALITY OF THE DATA

Analyses on large-scale datasets are valuable for observing trends at the genome level. Nevertheless, we need to bear in mind that there are several issues that complicate such analyses, like the dimension of time, the fact that different organisms are under different constraints and live in different environments, the quality and coverage of the data and the biases of each dataset, among others.

One example of how time complicates an analysis refers to two studies mentioned earlier, about the functional differences among SGDs and WGDs (Guan et al., 2007; Hakes et al., 2007). All ohnologues of yeast have the same age, around 100 million years, whereas the SG duplicates have various ages. Ideally, one should use SG duplicates as old as WG duplicates, but then, probably the number of SG duplicates would not be sufficient for statistical analysis. If the SGD dataset is dominated by very old duplicates, then the observed difference in functional divergence could be just a function of time and not due to the mode of duplication. Indeed, Guan et al. (2007) recognize the importance of dating the duplicates to increase the confidence in future analyses and discuss in depth the potential pitfalls and how they could be avoided. Future analyses with more species and functional data are needed to fully resolve this problem.

Another issue related to time is this of the rate of gene loss during the diploidization phase. In yeast and in plants, this rate is initially very high, but is declining over time (Maere et al., 2005; Semon and Wolfe, 2007). In addition, the rate of loss is different for various GO categories in plants (Maere et al., 2005). Therefore, depending on how old the event is, we should observe different outcomes, in terms of gene content for a genome.

Occasionally, analyses appear to contradict common beliefs and stir up discussion about our understanding of a process. A very interesting case is the one about the effect of gene dosage on the evolution of protein complexes, as discussed by (Freeling and Thomas, 2006; Pereira-Leal and Teichmann, 2005). Although the DBH posits that protein complexes should preferentially evolve by WGDs and not by SGDs, Pereira-Leal and Teichmann (2005) show the opposite. These authors found that the predominant mechanism of protein complex creation was not duplication, actually. Nevertheless, there was a small, but significant portion of the complexes that evolved by duplication from other complexes. Pereira-Leal and Teichmann (2005) classified homologous complexes as concurrent and parallel. Concurrent complexes share some of the components, whereas in parallel complexes, two homologous complexes will have similar but no shared components. Bioinformatics analyses showed that the most reasonable scenario is one where new homologous complexes arise by a slow process of step-wise duplications, whereas the ancient whole-genome duplication that occurred in yeast did not have a significant impact on complex creation. The new homologous complexes retained the general function of the ancestral complexes, but evolved new specificities. This finding seems to contradict the DBH, which predicts whole-genome duplication as the most favorable mechanism, in order to avoid dosage imbalance, unless mechanisms for compensation of the dosage imbalance are in action. In addition, Hakes et al. (2007) found that for genes that participate in complexes, onhologues are more dispensable than single-gene duplicates, which appears to contradict the DBH again. Mintseris and Weng (2005) found evidence for the effect of dosage imbalance on stable protein complexes, but could not address this issue regarding the transient interactions.

Many large-scale studies are based on literature-curated data as well as on high-throughput experiments and try to provide a global snapshot of the effect of whole-genome duplication. Although the literature-curated data are considered as the gold standard and are expected to have low coverage but describing highly confident interactions, on the other hand, many of the large-scale experiments used for some of the analyses are incomplete, with errors and biases. For example, the yeast and human interactomes are 50% and 10% complete, with an estimated total of 38–75,000 interactions in yeast and 154–369,000 interactions in humans (Hart et al., 2006). A more recent analysis converges on the number of yeast interactions, but doubles the number of human interactions (Stumpf et al., 2008). In addition, the false positive rates for any experimental high-throughput (HTP) method may fluctuate between 30% and 80% (Hart et al., 2006). Therefore, it is of no surprise that experts in the field suggest that in the future, the interactome should be treated like a genome, with multiple coverage, to account for mistakes.

Another major concern about HTP interactome data is the bias of certain technologies toward detecting the interactions of certain gene categories (extensively reviewed by (Lalonde et al., 2008)). Methods that have been adapted for HTP screening include the yeast-two-hybrid (Y2H), the mating-based split-ubiquitin (mbSUS), and affinity purification of protein complexes followed by mass-spectroscopy identification of proteins (AP-MS). The various methods differ in their sensitivity, specificity, and ability to detect interactions over a broad spectrum of affinities. Also, some methods detect direct physical interactions (e.g., Y2H), whereas others determine the presence of one protein within a protein complex (AP-MS) or its vicinity (FRET). Y2H and AP-MS also differ in their ability to detect PPIs with different kinetics and binding affinities. AP-MS, for example, will be biased in favor of stable complexes. Y2H is more capable of detecting binary transient interactions. In one of the first evaluations of the various interactome datasets, many biases were identified in the yeast interaction data (von Mering et al., 2002), related to certain cellular environments, more ancient, conserved, or highly expressed proteins. Lalonde et al. (2008) discuss common problems of these HTP technologies that include (i) limited number of replica tests, (ii) the interactions are assayed in an all-or-nothing scheme, ignoring binding affinities, (iii) proteins are often overexpressed, thus modifying the relative concentration of potential partners, (iv) heterologous systems may be used, and (v) analysis of interactions in cellular extracts that may bring together proteins from different compartments. Therefore, HTP data include potential interactions, together with *in vivo* interactions. Without a doubt, HTP screens are necessary to obtain an overview of the potential interactome, but low-throughput, carefully prepared follow-up studies are needed to verify the initial data (Lalonde et al., 2008).

Given all the problems that blight HTP interaction data, we can wonder to what extent the conclusions of duplication-interaction related analyses are robust? At this early phase in the era of interactomics, caution is definitely advised, but it should not give way to pessimism or nihilism. Most of the bioinformatics analyses are performed on yeast genomic and functional data. Yeast is indeed the best choice for such analyses, because there is a significant number of related yeast species (some predating the WGD) that have been sequenced and thus helped to provide a very confident and carefully analyzed dataset of ohnologues (Byrne and Wolfe, 2005; Kellis et al., 2004). In addition, the yeast interactome has the highest coverage compared to other organisms, with an estimated 50%. Analyses by two groups found that, in terms of functional distance, PPI data are in agreement with an integrated Bayesian analysis of diverse functional data and semantic distance based on Gene Ontology annotation (Guan et al., 2007; Hakes et al., 2007). Therefore, most of the analyses about PPIs should at least be able to capture some strong global trends.

19.11 THE IMPORTANCE OF MEDIUM-SCALE STUDIES: THE CASE OF DIMERIZATION

Although large-scale studies provide a bird's eye view on the evolution of PINs, they are complicated by several problems that are not so strongly present in medium-scale studies. Usually, at the level of a protein family or a certain pathway, these medium-scale studies use data with higher quality and coverage and provide a deeper insight into the mechanisms and processes that govern the evolution of PPIs. One well-studied case related to regulatory PINs is the evolution of dimerizing interactions in metazoan TFs.

Dimerization is defined as the formation of a functional protein complex composed of two subunits (Klemm et al., 1998). In signal transduction pathways, dimeric interactions usually are not very stable. Rather, they are dynamic and act as reversible switches in the process of information flow (Nooren and Thornton, 2003). Dimerization is observed in many signal transduction and regulatory gene families (Klemm et al., 1998; Marianayagam et al., 2004). In TFs, two monomers need to dimerize in order to bind DNA and depending on the choice of partner and the cellular context, each unique TF dimer triggers a sequence of regulatory events that lead to a particular cellular fate.

The best-studied TF families that form homotypic dimers (dimers among homologous proteins) are the bHLH, bZIP, Nuclear Receptors, MADS-box, HD-ZIP, and NF-kB, as well as the STATs. These TFs create a large number of dimers with distinct biological properties (more than 500 in human and up to 2500 when alternative splicing is accounted for) and form elaborate control circuits that are central to the evolution and generation of organismal complexity (Amoutzias et al., 2008). Dimerizing TFs regulate a very wide range of processes, such as the cell cycle, reproduction, development, homeostasis, metabolism, immunity, inflammation, and programmed cell death (Amoutzias et al., 2008). Members of these dimerizing TF families mediate their DNA-binding and dimerization activities via highly conserved domains. In all families, the DNA-binding domain is the most conserved portion of the protein, whereas the dimerization domain (which usually lies downstream from the DNA-binding domain) is less conserved (Amoutzias et al., 2008). These domains are shared among all the members of each family and therefore are often used in phylogenetic analyses. Usually, specific functions, such as recognition of DNA elements and dimerization are tightly linked to the phylogenetic clustering. Other regions of the proteins can contain transcriptional activation/repression domains, various functional domains, or phosphorylation sites; but these elements are usually not as highly conserved.

Some of the most important implications of TF dimerization are reviewed in (Amoutzias et al., 2008; Klemm et al., 1998; Marianayagam et al., 2004), with perhaps the most important being differential regulation. One TF monomer can have multiple binding partners and thus form dimers that possess distinct properties and perform specific functions, thereby mediating differential gene regulation. In this case, the concentration of each monomer in the cell, its posttranslational modifications (e.g., phosphorylation), and its binding affinity for other monomers will determine which dimer will form, and thus which signaling process will prevail over the others. Very good examples are the Myc–Max and Mad–Max heterodimers that define whether a large number of targeted genes will be expressed or silenced, respectively (Grandori et al., 2000; Luscher, 2001).

N genes of a given TF family could in theory generate N homodimers $+ (N \times (N - 1)/2)$ unique heterodimers, assuming negligible binding specificity among the monomers, a lack of cell- or tissue-specific expression patterns, and little alternative splicing. Therefore, for the 51 bZIPs, 118 bHLHs, and 48 NRs in humans, there is the potential to form 1326, 7021, and 1176 unique dimers, respectively. Theoretically, TF dimerization could make a huge contribution to gene regulation flexibility and complexity, given the fact that there are approximately 2000–3000 human sequence-specific TFs (Kummerfeld and Teichmann, 2006; van Nimwegen, 2003). In practice, the specificity of monomer–monomer interactions limits the available binding options. Protein-array experiments and reliable predictions based on biophysical constraints on leucine zipper (LZ) interactions lead to estimates of approximately 350 unique bZIP dimers (Fong et al., 2004; Grigoryan and Keating, 2006; Newman and Keating, 2003; Vinson et al., 2006). Strong evidence also indicates specificity of dimerization in the bHLHs, NRs, HD-ZIPs, MADS-box, and the plant bZIPs (Amoutzias et al., 2007a; Amoutzias et al., 2004a; Amoutzias et al., 2007b; de Folter et al., 2005; Ehlert et al., 2006; Johannesson et al., 2001; Veron et al., 2007). A very interesting finding is that the paralogues of any given phylogenetic subgroup in bZIPs, bHLHs, NRs, MADS-box, HD-ZIPs share, to a high degree, their various dimerization partners (Amoutzias et al., 2007a; Amoutzias et al., 2004a; Amoutzias et al., 2007b; Johannesson et al., 2001; Newman and Keating, 2003). This results from the evolution of the TF families.

19.12 EVOLUTION OF DIMERIZATION NETWORKS

Although some of the DNA-binding folds of dimerizing TFs are also found in prokaryotes (e.g., HTH), all of the TF families that we have discussed so far are specific to eukaryotes. Some are found in the metazoan, fungal and plant lineages (bHLH, bZIP, and MADS-box), whereas others are specific to plants (HD-ZIP) or the metazoa/opisthokonta (NF-κB, NR, and STAT) (Amoutzias et al., 2008). Although several ancient TF families are found in all three of these eukaryotic lineages, some have undergone significant lineage-specific expansion in only one (i.e., MADS-box TFs in plants) or two (i.e., bHLH and bZIP in metazoa and plants) of the lineages, independently.

The integration of genomic and functional data for the three largest families of dimerizing TFs in metazoa, the bHLHs (Amoutzias et al., 2004a), NRs (Amoutzias et al., 2007a) and, especially, the bZIPs (Amoutzias et al., 2007a) delineates, to some extent, the evolution of DNA-binding and dimerization specificity during the major phases of animal macroevolution and shows the effects of SGDs and WGDs (Figure 19.2). From the genomes of fungi, diploblastic cnidarians (*Nematostella vectensis*), invertebrates (insects), and vertebrates (fishes and human), we can now understand more about the events that occurred during the emergence of bilaterian and vertebrate animals (Figure 19.2). Briefly, major gene duplications, point mutations, and domain rearrangements occurred at the origin

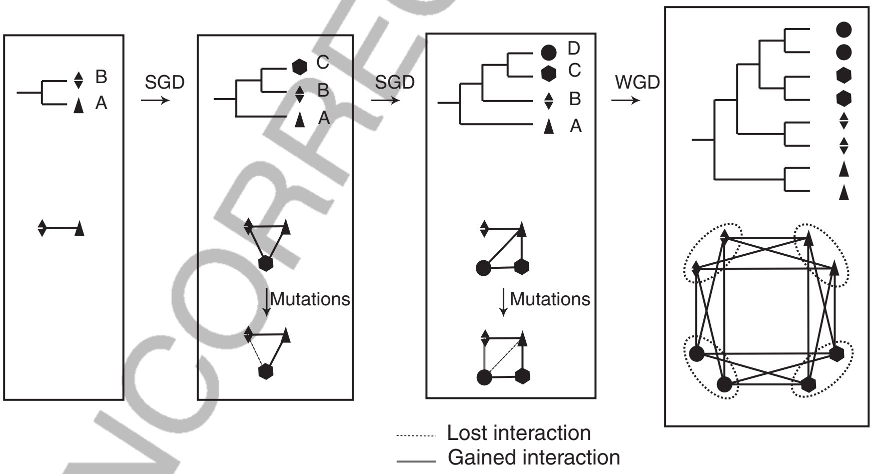

Figure 19.2 A general example of how TF dimerizing interactions have evolved during the major phases of animal macroevolution. SGD events and mutations (point mutations or domain rearrangements) created the various subfamilies, each one with a distinct interaction pattern. The subfamilies and their core dimerization network were formed by the time the bilaterian animals appeared. The two rounds of WGDs at the dawn of vertebrate evolution created more paralogues for each subfamily, but overall, these paralogues have, until today, retained the dimerizing interaction pattern of their ancestral molecules. (See insert for color representation of this figure.)

of metazoa, approximately 1 billion years ago. These events shaped the repertoire of gene subfamilies and the interactions among them. (Miyata and Suga, 2001) have hypothesized that the emergence of multicellularity was accompanied by a phase of large-scale duplications and domain rearrangements for the signaling families in general, but so far, there exists no hard evidence for a WGD during this early period. By the time the urbilaterian ancestor arose around 650 MYA, a highly conserved core dimerization network had already been formed, with most of the subfamilies present. The genes that evolved during this period seem to have been shaped by single-gene duplications. Its has been hypothesized that the emergence and rapid radiation of bilateria is not due to a large-scale duplication event, rather due to the duplication and rewiring of a few key gene families (Davidson, 2006; Miyata and Suga, 2001). Nevertheless, (Spring, 2003) suggests the opposite that is a genome duplication that had as a consequence the emergence of bilateria. Later, two rounds of whole-genome duplications occurred at the origin of vertebrates (2R event) around 550 MYA (Dehal and Boore 2005; Panopoulou and Poustka, 2005) and added more paralogues to each subfamily, but overall, they did not create many new subfamilies. These highly similar paralogues possess very similar DNA-binding and dimerization specificities until today in humans. The evolution of the bZIP, bHLH, and NR networks strongly supports this picture (Amoutzias et al., 2007a; Amoutzias et al., 2004a).

The bZIPs are a well-studied case of how dimerization and DNA-binding evolved in animals, owing to a plethora of functional data, especially their high quality and coverage of dimerization data. The human bZIP dimerization network was reconstructed from a protein-array technology, which does not show the biases that afflict Y2H assays. This array was used to monitor all possible bZIP interactions of one protein against all other proteins (Newman and Keating, 2003). The interaction array showed high symmetry and reproducibility and was in good agreement with the literature. In addition, rules that were derived from this dataset for predicting specificity were in good overall agreement with rules derived from independent studies.

Current data show that the genome of the last common ancestor of eumetazoa contained genes for many dimerizing bZIP subfamilies (Amoutzias et al., 2007a). Most of these subfamilies must have emerged after the divergence of the fungi and before that of the cnidaria. Only 1 of the 19 human bZIP subfamilies was shared with the fungi, whereas 13 are shared with cnidaria. In addition, these 13 subfamilies recognize all 6 DNA elements bound by the human bZIPs. Therefore, specificity of DNA binding mainly evolved during this period. Several of these ancient bZIP subfamilies subsequently duplicated and, while retaining their DNA-binding affinity for certain motifs, started to diverge at the dimerization domain, thus gaining and losing interactions with other bZIP subfamilies. This change in dimerization specificity could allow the new combinations of monomers to recognize new DNA motifs and thus increase the regulatory capacity of the genome. Of note, many bZIPs are involved in developmental processes. By the time the common ancestor of bilateria arose (just before the hypothesized Cambrian explosion), 17 of the 19 bZIP subfamilies were present and formed a complex core dimerization network, conserved in many vertebrate and invertebrate bilaterians. There is no evidence so far in the literature for a WGD during this period, although (Spring, 2003) suggests the opposite. Until this time, most of the subfamilies must have consisted of only one gene.

At the origin of vertebrates, around 550 MYA, all of the 19 bZIP subfamilies were present. Then, the two rounds of whole-genome duplication (the 2R event) that the vertebrate ancestor underwent created more paralogues for each subfamily. These paralogues not only retained the DNA-binding specificity of their ancestral molecule but also retained most of its dimerizing interactions until today. The paralogues evidently diverged outside of the

DNA-binding and dimerization domains, thus making new interactions with other signal transduction molecules. At least 35 from the 51 human bZIPs are retained duplicates from the 2R, based on phylogenetic analysis. The high retention of WG duplicates is a general trend observed in vertebrate TFs (Blomme et al., 2006). Subsequent lineage specific SGDs and losses of TFs also occurred in the vertebrate lineage, though to a limited extent.

The same scenario has been proposed for the evolution of the bHLH dimerization network in metazoa (Amoutzias et al., 2004b). Again, single-gene duplication and domain rearrangements formed the various subfamilies somewhere between the origin of metazoa and the origin of bilaterian animals. During that time, the core topology of the bHLH network was formed. Later on, 2R increased the number of paralogues for each subfamily, but overall, the majority of dimerizing interactions among the paralogues have been conserved until today.

19.13 CONCLUSIONS

As more PPI data are generated in small or large-scale experiments, a conceptual model of the interactome is gradually being formalized and refined, composed of binary interactions and multisubunit complexes. Many binary interactions are transient and involved in information processing, such as transcriptional regulation or signal transduction (Brown and Jurisica, 2007; Mintseris and Weng, 2005). On the other hand, multisubunit complexes form sophisticated molecular machines that are composed of core, module and attachment proteins (Gavin et al., 2006; Krogan et al., 2006). Usually, most of the members of a complex are tightly coregulated and form the immature complex that waits for the final components to be expressed in the right occasion and thus complete the complex, a principle termed as "just in time assembly" (de Lichtenberg et al., 2005). Protein complexes seem to form the highly conserved part of the PIN, whereas the binary transient interactions form an evolutionarily flexible coat around the conserved core.

From the analyses that we have discussed so far, WGDs seem to have a strong effect on that part of the PIN that is composed of transient interactions. WGDs provide the raw material for rapid evolution and rewiring of interactions that are involved in information processing, like phosphorylation. Nevertheless, the effect of SGDs on the evolution of PPIs should not be underestimated. Medium-scale studies on metazoan dimerizing networks (Amoutzias et al., 2004b; Amoutzias et al., 2008; Amoutzias et al., 2007b) together with large-scale studies on PPIs (Guan et al., 2007; Hakes et al., 2007; Pereira-Leal and Teichmann, 2005) show that paralogues from SGDs underwent more drastic changes in their PPIs, compared to paralogues from WGDs. In addition, more protein complexes seem to have evolved by step-wise gene duplications rather than WGDs (Pereira-Leal and Teichmann, 2005). Is it possible that SGDs are involved in the rewiring of the conserved PIN core and thus are linked to major innovations in evolution, whereas WGDs provide the raw material for rapid rewiring of PINs and thus rapid adaption and species radiation? In our opinion, it is not clear yet if what we have observed so far is an effect of the time of duplication, or is inherent to the mode of duplication, although current data point toward this latter possibility. Future work and more data on genomes and interactomes will undoubtedly shed further light on these fundamental questions.

ACKNOWLEDGMENT

Grigoris Amoutzias is supported by an EMBO long-term fellowship.

REFERENCES

Amoutzias, G.D., Pichler, E.E., Mian, N., De Graaf, D., Imsiridou, A., Robinson-Rechavi, M., Bornberg-Bauer, E., Robertson, D.L., and Oliver, S.G., 2007a. A protein interaction atlas for the nuclear receptors: properties and quality of a hub-based dimerisation network. *BMC Systems Biol*. **1**: 34.

Amoutzias, G.D., Robertson, D.L., and Bornberg-Bauer, E., 2004a. The evolution of protein interaction networks in regulatory proteins. *Comp. Func. Genomics* **5**: 79–84.

Amoutzias, G.D., Robertson, D.L., Oliver, S.G., and Bornberg-Bauer, E., 2004b. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep*. **5**: 274–279.

Amoutzias, G.D., Robertson, D.L., Van de Peer, Y., and Oliver, S.G., 2008. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci.* **33**: 220–229.

Amoutzias, G.D., Veron, A.S., Weiner, J. 3rd., Robinson-Rechavi, M., Bornberg-Bauer, E., Oliver, S.G., and Robertson, D.L., 2007b. One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol. Biol. Evol.* **24**: 827–835.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., and Jubin, C. et al., 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**: 171–178.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., and Teichmann, S.A., 2004. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**: 283–291.

Barabasi, A.L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**: 101–113.

Birchler, J.A. and Veitia, R.A., 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* **19**: 395–402.

Blanc, G. and Wolfe, K.H., 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**: 1667–1678.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y., 2006. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol*. **7**: R43.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A., and Weiner, J., 3rd. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.* **62**: 435–445.

Brown, K.R. and Jurisica, I., 2007. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol*. **8**: R95.

Byrne, K.P. and Wolfe, K.H., 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. **15**: 1456–1461.

Casneuf, T., De Bodt, S., Raes, J., Maere, S., and Van de Peer, Y., 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol*. **7**: R13.

Chapman, B.A., Bowers, J.E., Feltus, F.A., and Paterson, A. H., 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc. Natl. Acad. Sci. USA* **103**: 2730–2735.

Chi, A., Huttenhower, C., Geer, L.Y., Coon, J.J., Syka, J.E., Bai, D.L., Shabanowitz, J., Burke, D.J., Troyanskaya, O. G., and Hunt, D.F., 2007. Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. USA* **104**: 2193–2198.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W., 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res*. **16**: 738–749.

Davidson, E.H., 2006. *The regulatory genome*, Academic Press.

Davis, J.C. and Petrov, D.A., 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet*. **21**: 548–551.

De Bodt, S., Maere, S., and Van de Peer, Y., 2005. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20**: 591–597.

de Folter, S., Immink, R.G., Kieffer, M., Parenicova, L., Henz, S.R., Weigel, D., Busscher, M., Kooiker, M., Colombo, L., Kater, M.M., Davies, B., and Angenent, G.C., 2005. Comprehensive interaction map of the *Arabidopsis* MADS box transcription factors. *Plant Cell* **17**: 1424–1433.

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P., 2005. Dynamic complex formation during the yeast cell cycle. *Science* **307**: 724–727.

Dehal, P. and Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. **3**: e314.

Ehlert, A., Weltmeier, F., Wang, X., Mayer, C.S., Smeekens, S., Vicente-Carbajosa, J., and Droge-Laser, W., 2006. Two-hybrid protein–protein interaction analysis in *Arabidopsis* protoplasts: establishment of a heterodimerization map of group C and group S bZIP transcription factors. *Plant J*. **46**: 890–900.

Evlampiev, K. and Isambert, H., 2008. Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc. Natl. Acad. Sci. USA* **105**: 9863–9868.

Fong, J.H., Keating, A.E., and Singh, M., 2004. Predicting specificity in bZIP coiled-coil protein interactions. *Genome Biol*. **5**: R11.

Freeling, M. and Thomas, B.C., 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res*. **16**: 805–814.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., and Boesche, M. et al., 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**: 631–636.

Gnad, F., Ren, S., Cox, J., Olsen, J.V., Macek, B., Oroshi, M., and Mann, M., 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*. **8**: R250.

Grandori, C., Cowley, S.M., James, L.P., and Eisenman, R.N., 2000. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16**: 653–699.

Grigoryan, G. and Keating, A.E., 2006. Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.* **355**: 1125–1142.

Guan, Y., Dunham, M.J., and Troyanskaya, O.G., 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* **175**: 933–943.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G., and Robertson, D.L., 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol*. **8**: R209.

Hart, G.T., Ramani, A.K., and Marcotte, E.M., 2006. How complete are current yeast and human protein-interaction networks? *Genome Biol*. **7**: 120.

Heazlewood, J.L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., and Schulze, W.X., 2008. PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res*. **36**: D1015–1021.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T. R., Sikes, J.G., Obradovic, Z., and Dunker, A.K., 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. **32**: 1037–1049.

Ispolatov, I., Yuryev, A., Mazo, I., and Maslov, S., 2005. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res*. **33**: 3629–3635.

Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., and Stange-Thomann, N. et al., 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**: 946–957.

Johannesson, H., Wang, Y., and Engstrom, P., 2001. DNA-binding and dimerization preferences of Arabidopsis homeodomain-leucine zipper transcription factors *in vitro*. *Plant Mol. Biol.* **45**: 63–73.

Kellis, M., Birren, B.W., and Lander, E.S., 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**: 617–624.

Klemm, J.D., Schreiber, S.L., and Crabtree, G.R., 1998. Dimerization as a regulatory mechanism in signal transduction. *Annu Rev. Immunol.* **16**: 569–592.

Kondrashov, F.A. and Koonin, E.V., 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet*. **20**: 287–290.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E. V., 2002. Selection in the evolution of gene duplications. *Genome Biol*. **3**: RESEARCH0008.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., and Guo, X. et al., 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**: 637–643.

Kummerfeld, S.K. and Teichmann, S.A., 2006. DBD: a transcription factor prediction database. *Nucleic Acids Res*. **34**: D74–81.

Lalonde, S., Ehrhardt, D.W., Loque, D., Chen, J., Rhee, S.Y., and Frommer, W.B., 2008. Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations. *Plant J*. **53**: 610–635.

Liang, H., Plazonic, K.R., Chen, J., Li, W.H., and Fernandez, A., 2008. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet*. **4**: e11.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K., 2006. Intrinsic disorder in transcription factors. *Biochemistry* **45**: 6873–6888.

Lobley, A., Swindells, M.B., Orengo, C.A., and Jones, D.T., 2007. Inferring function using patterns of native disorder in proteins. *PLoS Comput Biol*. **3**: e162.

Luscher, B., 2001. Function and regulation of the transcription factors of the Myc/Max/Mad network. *Gene* **277**: 1–14.

Lynch, M. and Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lynch, M. and Katju, V., 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet*. **20**: 544–549.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y., 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**: 5454–5459.

Marianayagam, N.J., Sunde, M., and Matthews, J.M., 2004. The power of two: protein dimerization in biology. *Trends Biochem Sci*. **29**: 618–625.

Mintseris, J. and Weng, Z., 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl. Acad. Sci. USA* **102**: 10930–10935.

Miyata, T. and Suga, H., 2001. Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays* **23**: 1018–1027.

Musso, G., Zhang, Z., and Emili, A., 2007. Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends Genet*. **23**: 266–269.

Newman, J.R. and Keating, A.E., 2003. Comprehensive identification of human bZIP interactions with coiled-coil arrays. *Science* **300**: 2097–2101.

Nooren, I.M. and Thornton, J.M., 2003. Diversity of protein–protein interactions. *EMBO J*. **22**: 3486–3492.

Nuhse, T.S., Stensballe, A., Jensen, O.N., and Peck, S.C., 2004. Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* **16**: 2394–2405.

Ohno, S., 1970. *Evolution by gene duplication*, Springer, Berlin.

Panopoulou, G. and Poustka, A.J., 2005. Timing and mechanism of ancient vertebrate genome duplications: the adventure of a hypothesis. *Trends Genet*. **21**: 559–567.

Papp, B., Pal, C., and Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.

Pastor-Satorras, R., Smith, E., and Sole, R.V., 2003. Evolving protein interaction networks through gene duplication. *J Theor. Biol.* **222**: 199–210.

Peck, S.C., 2006. Phosphoproteomics in *Arabidopsis*: moving from empirical to predictive science. *J. Exp. Bot*. **57**: 1523–1527.

Pereira-Leal, J.B., Levy, E.D., Kamp, C., and Teichmann, S. A., 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol*. **8**: R51.

Pereira-Leal, J.B. and Teichmann, S.A., 2005. Novel specificities emerge by stepwise duplication of functional modules. *Genome Res*. **15**: 552–559.

Presser, A., Elowitz, M.B., Kellis, M., and Kishony, R., 2008. The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication. *Proc. Natl. Acad. Sci. USA* **105**: 950–954.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., and Zhu, X. et al., 2005. Global analysis of protein phosphorylation in yeast. *Nature* **438**: 679–684.

Ranea, J.A., Grant, A., Thornton, J.M., and Orengo, C.A., 2005. Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet*. **21**: 21–25.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., and Wolfe, K.H., 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341–345.

Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C., 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47**: 868–876.

Semon, M. and Wolfe, K.H., 2007. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* **17**: 505–512.

Spring, J., 2003. Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. *J. Struct. Funct. Genomics* **3**: 19–25.

Sprinzak, E., Altuvia, Y., and Margalit, H., 2006. Characterization and prediction of protein–protein interactions within and between complexes. *Proc. Natl. Acad. Sci. USA* **103**: 14718–14723.

Sterck, L., Rombauts, S., Vandepoele, K., Rouze, P., and Van de Peer, Y., 2007. How many genes are there in plants (. . . and why are they there)? *Curr. Opin. Plant Biol.* **10**: 199–203.

Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M., and Wiuf, C., 2008. Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**: 6959–6964.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y., 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res*. **13**: 382–390.

Tompa, P. and Fuxreiter, M., 2008. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem Sci*. **33**: 2–8.

Van de Peer, Y., 2004. Computational approaches to unveiling ancient genome duplications. *Nat. Rev Genet.* **5**: 752–763.

van Nimwegen, E., 2003. Scaling laws in the functional content of genomes. *Trends Genet*. **19**: 479–484.

Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A., and Van de Peer, Y., 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. USA* **101**: 1638–1643.

Veitia, R.A., 2005. Paralogs in polyploids: one for all and all for one? *Plant Cell* **17**: 4–11.

Veitia, R.A., Bottani, S., and Birchler, J.A., 2008. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet*. **24**: 390–397.

Veron, A.S., Kaufmann, K., and Bornberg-Bauer, E., 2007. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol. Biol. Evol.* **24**: 670–678.

Vinson, C., Acharya, A., and Taparowsky, E.J., 2006. Deciphering B-ZIP transcription factor interactions *in vitro* and *in vivo*. *Biochim. Biophys. Acta* **1759**: 4–12.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P., 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**: 399–403.

Wagner, A., 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91**: 4387–4391.

Wagner, A., 2002. Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.* **19**: 1760–1768.

Wilkins, M.R. and Kummerfeld, S.K., 2008. Sticking together? Falling apart? Exploring the dynamics of the interactome. *Trends Biochem. Sci.* **33**: 195–200.

Wolfe, K.H., 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet*. **2**: 333–341.

Wolfe, K.H. and Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.

Xia, K., Fu, Z., Hou, L., and Han, J.D., 2008. Impacts of protein–protein interaction domains on organism and network complexity. *Genome Res*. **19**: 1500–1508.