The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs

Axelle Marchant[1,2,3,4], Angel F. Cisneros[1,2,3], Alexandre K Dubé[1,2,3,4], Isabelle Gagnon-Arsenault[1,2,3,4], Diana Ascencio[1,2,3,4], Honey A. Jain[1,2,3,9], Simon Aubé[1,2,3], Chris Eberlein[1,2,3,4], Daniel Evans-Yamamoto[5,6,7], Nozomu Yachie[5,6,7,8] & Christian R. Landry[1,2,3,4*]

1. Département de Biochimie, Microbiologie et Bio-informatique, Faculté des sciences et de génie, Université Laval, Québec, Québec, G1V 0A6, Canada

2. PROTEO, Le réseau québécois de recherche sur la fonction, la structure et l'ingénierie des protéines, Université Laval, Québec, Québec, G1V 0A6, Canada

3. Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, Québec, G1V 0A6, Canada

4. Département de Biologie, Faculté des sciences et de génie, Université Laval, Québec, Québec, G1V 0A6, Canada

5. Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan.

6. Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan

7. Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan

8. Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

9. Department of Biological Sciences, Birla Institute of Technology and Sciences-Pilani, Goa-India

*Corresponding author:
Christian R Landry

1030, Avenue de la Médecine
Université Laval
Québec (Québec) G1V 0A6
Canada
Phone: 418-656-3954
christian.landry@bio.ulaval.ca

Keywords: gene duplication, protein-protein interaction networks, homomers, heteromers, pleiotropy, regulatory evolution, functional divergence.

Abstract

Paralogous proteins often arise from the duplication of genes encoding homomeric proteins. Such events lead to the formation of homomers and heteromers, thus creating new complexes after a single duplication event. We exhaustively characterize this phenomenon using the budding yeast protein-protein interaction network. We observe that heteromerizing paralogs are very frequent and less functionally diverged than non-heteromerizing ones, raising the possibility that heteromerization prevents functional divergence. Using *in silico* evolution, we show that for homomers and heteromers that share binding interfaces, mutations in one paralog have pleiotropic effects on both homomers and heteromers, resulting in highly correlated responses to selection. As a result, heteromerization could be preserved indirectly due to negative selection for the maintenance of homomers. By integrating data on gene expression and protein localization, we find that paralogs can overcome the obstacle of structural pleiotropy and develop functional divergence through regulatory evolution.

Introduction

Proteins can assemble into both stable and transient molecular complexes that perform and regulate structural, metabolic and signalling functions (Janin et al., 2008; Marsh and Teichmann, 2015; Pandey et al., 2017; Scott and Pawson, 2009; Vidal et al., 2011; Wan et al., 2015). The assembly of such complexes are necessary for protein function and thus constrain the sequence space available for protein evolution. One direct consequence of the multiple protein-protein interactions (PPIs) among proteins is that a mutation in a given gene can have pleiotropic effects on other gene functions through physical associations. Therefore, to understand how genes and cellular systems evolve, we need to consider physical interactions as part of the environmental factors shaping a gene's evolutionary trajectory (Landry et al., 2013; Levy et al., 2012).

A context in which PPIs and pleiotropy may be particularly important is during the evolution of new genes after duplication events (Amoutzias et al., 2008; Baker et al., 2013; Diss et al., 2017; Kaltenegger and Ober, 2015). Indeed, the potential for a gene to evolve new functions depends on its molecular environment. The physical environment includes a gene's paralog if the duplicates are derived from an ancestral self-interacting protein (homomer) (Figure 1). In this case, mutations in one paralog could have functional consequences for the other copy because the duplication of a homomeric protein leads not only to the formation of new homomers (HM) but also to a new heteromer (HET) (Figure 1) (Pereira-Leal et al., 2007; Wagner, 2003).

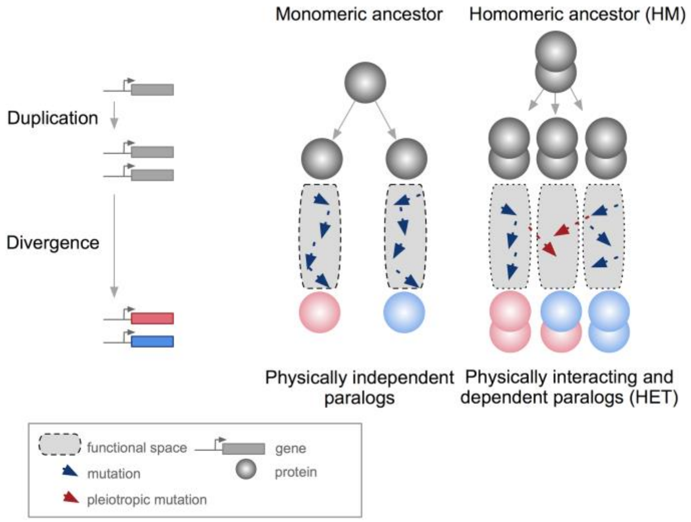

Figure 1: Mutations in paralogous proteins originating from an ancestral homomer are likely to have pleiotropic effects on each other's function due to their physical association.
Gene duplication leads to physically interacting paralogs (HET) when they derive from an ancestral homomeric protein (HM). The evolutionary fates of the physically associated paralogs tend to be interdependent because mutations in one gene can have impact on the function of the other copy through heteromerization. Following the duplication of a HM, mutations can affect both HMs and HET, making them pleiotropic mutations.

Paralogs that derive from HMs are physically associated as HETs when they arise. Subsequent evolution can lead to the maintenance or the loss of these HETs. There are several examples of paralogs that maintained the ability to form HETs and, as a consequence, they have evolved new functional relationships (Amoutzias et al., 2008; Baker et al., 2013; Kaltenegger and Ober, 2015). Examples include paralogs degenerating and becoming a repressor for the other copy (Bridgham et al., 2008), paralogs that split the functions of the ancestral HM between one of the HMs and the HET (Baker et al., 2013), that cross-stabilize and thus need each other to perform their function (Diss et al., 2017) or that evolve a new function together as a HET (Boncoeur et al., 2012). However, there are other paralogs that have lost the ability to form HETs through the divergence of their structural properties. Examples include duplicated histidine kinases (Ashenberg et al., 2011) and many duplicated heat-shock proteins (Hochberg et al., 2018), which do form HMs but appear to have lost the ability to form HETs.

One important question to examine is therefore: what are the evolutionary forces at work for the maintenance or the disruption of HETs arising from HMs. Previous studies suggest that if a paralog pair maintains its ability to form HMs, it is very likely to maintain the HET complex as well (Pereira-Leal et al., 2007). For instance, Lukatsky et al. (Lukatsky et al., 2007) showed that proteins intrinsically tend to interact with themselves and that negative selection could be needed to prevent unwanted HMs. Given this, since nascent paralogs are identical just after duplication, they would have a high propensity to assemble with each other. Hence, the formation of both HMs and HETs could be the default state after duplication and this, until specific destabilizing mutations accumulate (Ashenberg et al., 2011; Hochberg et al., 2018). Here, we hypothesize that the association of paralogs forming HETs acts as a constraint that may slow the functional divergence of paralogs by keeping gene products physically associated.

Previous studies have shown that HMs are enriched in eukaryotic PPI networks (Lynch, 2012; Pereira-Leal et al., 2007). However, the extent to which paralogs interact with each other has not been precisely quantified in any species. We therefore examine paralog assembly exhaustively in an eukaryotic interactome by collecting data from the literature and extending these with a large-scale PPI screening experiment. Second, using computational analyses, we examine the functional consequences of losing HET formation for HM forming paralogs. We then perform *in silico* evolution experiments to examine whether molecular pleiotropy could contribute to maintain interaction between paralogs that derive from ancestral HMs. We show that at least in cases where interaction interfaces are the same for HMs and HETs, selection to maintain HMs alone coupled with the pleiotropic effects of mutations may be sufficient to prevent the loss of HETs. Finally, we find that regulatory evolution, either at the level of gene transcription or protein localization, may relieve the pleiotropic constraint maintaining the interaction of paralogous proteins.

Results

Homomers and heteromers in the yeast PPI network

We collected information on *Saccharomyces cerevisiae* HMs and HETs from publicly available data (see methods) for 255 pairs of small-scale duplicates (SSDs) and 240 pairs of whole-genome duplicates (WGDs). We combined this data with our own experimental data on 155 SSDs and 131 WGDs. We performed Protein-fragment Complementation Assay experiments (referred to as PCA) to test for binary interactions of paralogs with themselves (HM) and with their sister copy (HET). PCA detects direct and near direct interactions without disturbing endogenous regulation, giving insight into the role of transcriptional regulation on the evolution of PPIs (Barshir et al., 2018; Gagnon-Arsenault et al., 2013; Rochette et al., 2014; Tarassov et al., 2008). In general, the PCA signal in our study strongly correlates with results from previous PCA experiments (Stynen et al., 2018; Tarassov et al., 2008) and other publicly available data (Figure S1). Roughly 77% of the HMs and 83% of the HETs detected in our PCA were previously reported (Figure S2, Tables S1 and S2), suggesting that most of the HMs and HETs that can be detected with available tools (and in standard conditions) have been discovered. While 91 HMs and 49 HETs reported in other studies were not detected in our PCA, our experiments discovered 48 HMs and 22 HETs not previously reported (Tables S1 and S2). The data we assembled represents a total of 595 pairs of paralogs (315 SSDs and 280 WGDs) covering 62.13% of the SSDs and 51.6% of the WGDs (Tables S1 and S2).

Using this dataset, we find that about 31% of yeast proteins (paralogs and singletons) form HMs, which agrees with previous estimates across species (Lynch, 2012). The proportion of HMs among singletons (n = 629, 25%) is lower than for all duplicates: SSDs (n = 1656, 34%, p-value < 2.2e-16), WGDs (n = 300, 32%, p-value = 5.927e-06) and those duplicated by both SSDs and WGDs (henceforth referred to as 2D, n = 76, 31%, p-value = 0.037) (Figure 2. A). It is possible that the frequency of HMs is underestimated because they were not systematically tested in previous studies (see methods). Another possibility is that some methods cannot detect them due to reasons such as low expression levels (this is also true for HETs, see below). To test for the effect of expression levels on PPI detection in the PCA assays, we measured mRNA abundance in cells growing in conditions that reflect the ones used in PCA experiments and used integrated protein abundance for the yeast proteome (Wang et al., 2012) (Tables S3 and S4). As previously observed (Celaj et al., 2017; Freschi et al., 2013), we found a correlation between PCA signal and expression level, both at the level of mRNA and protein abundance (Spearman r = 0.32, p-value < 2.2e-16 and r = 0.45, p-value < 2.2e-16 respectively). When focusing only on HMs previously reported, we also detected both correlations (Spearman r = 0.27, p-value = 2.248e-07 and r = 0.29, p-value < 2.2e-16 respectively). Associations between PCA signal and expression translate to a roughly two-fold increase in the probability of HM detection when mRNA levels change by two orders of magnitude (Figure S3. A). We also find that PCA signal for HMs is generally stronger for the most expressed paralog of a pair, confirming the effect of expression on our ability to detect HMs and HETs (Figure S3. B). Finally, we find that HMs reported in the literature but not detected by PCA have on average lower expression levels (Figure S3. B-C). We therefore conclude that some HMs (and also HETs) remain unknown because of low protein expression levels.

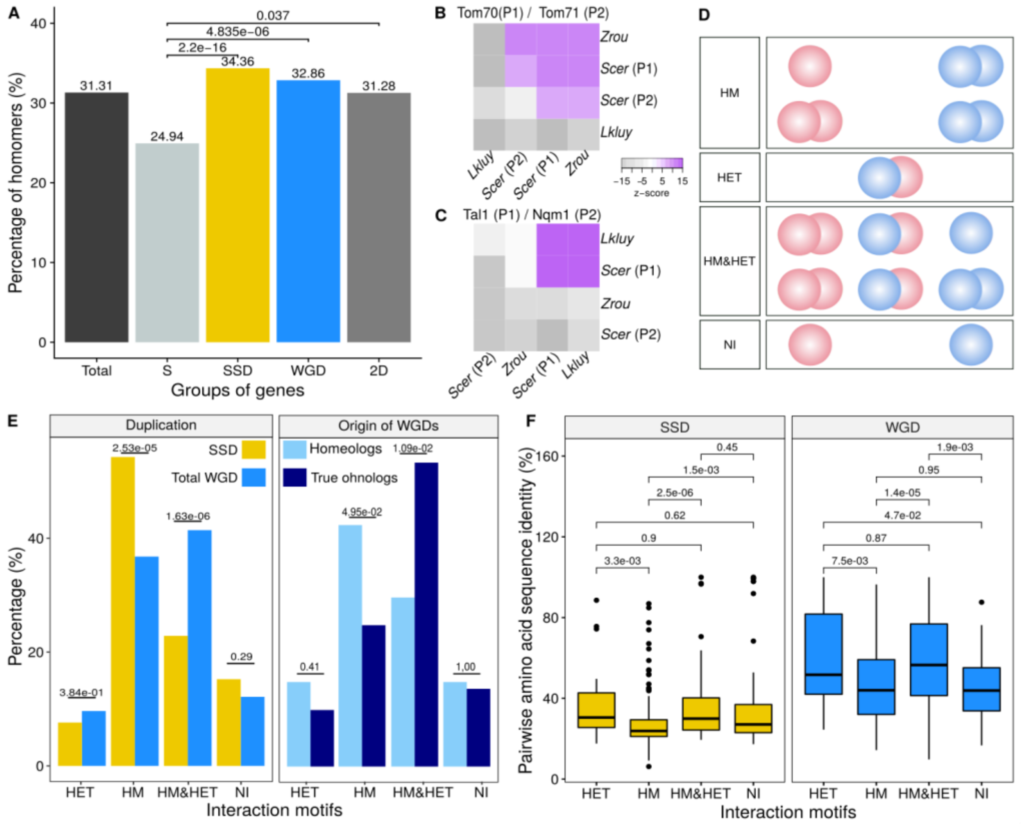

Figure 2: Homomers and heteromers of paralogs are frequent in the yeast protein interaction network. (**A**) The percentage of homomeric proteins in *S. cerevisiae* varies among singletons (S), small-scale duplicates (SSDs), whole-genome duplicates (WGDs) and genes duplicated by the two types of duplication (2D) (Chi-square test: p-value = 5.754e-15). Each category is compared with the singletons using a Fisher exact test. p-values are reported on the graph. **(B and C)** Analysis of interactions between pre-whole-genome duplication species orthologs and *S. cerevisiae* paralogs using DHFR PCA. The purple color shows significant interactions and their PCA signal intensity converted to z-scores. Experiments are performed in *S. cerevisiae*. Interactions are tested among: **(B)** *S. cerevisiae* (Scer) paralogs Tom70 (P1) & Tom71 (P2) and their orthologs in *Lachancea kluyveri* (Lkluy, SAKL0E10956g) and in *Zygosaccharomyces rouxii* (Zrou, ZYRO0G06512g) **(C)** *S. cerevisiae* paralogs Tal1 (P1) & Nqm1 (P2) and their orthologs in *L. kluyveri* (SAKL0B04642g) and in *Z. rouxii* (ZYRO0A12914g). (**D**) Paralogs show six motifs of interactions that we grouped in four categories according to their patterns. NI (non-interacting) pairs show no interaction. HM pairs show at least one homomer. HET pairs show heteromers only. HM&HET pairs show at least one homomer and the heteromer. (**E**) The percentage of motifs of interaction for SSDs (yellow) and WGDs (blue) (left panel) and between homeologs that originated from inter-species hybridization and true ohnologs from the whole-genome duplication (right panel). The p-values are from Fisher's exact tests. (**F**) Percentage of amino acid sequence identity between paralogs for each motif for SSDs and WGDs. P-values are from Wilcoxon tests.

The overrepresentation of HMs among duplicates was initially observed for human paralogs (Pérez-Bercoff et al., 2010). One possibility to explain this finding is that homomeric proteins are more likely to be maintained after duplication (Diss et al., 2017). Another explanation is that proteins forming HMs could be more expressed and therefore, easier to detect, as shown above. High expression could also increase the long term probability of genes to persist after duplication (Gout et al., 2010; Gout and Lynch, 2015). We observed that WGDs are more expressed than SSDs and that both are more expressed than singletons at mRNA and protein levels (Figure S4. A-B). However, expression does not explain completely the enrichment of HMs among duplicated proteins and therefore, this enrichment does not result entirely from reduced detection sensitivity. Both factors, expression and duplication, have significant effects on the probability of proteins to form HMs (Table S5. A). It is therefore likely that the overrepresentation of HMs of paralogs is linked to their higher expression but other factors are also involved.

Heteromers of paralogs frequently derive from ancestral homomers

The model presented in Figure 1 assumes that the ancestral protein leading to HET formed a HM before duplication. Under the principle of parsimony, we can assume that when at least one paralog forms a HM, the ancestral protein was also a HM. This was shown to be true in general by (Diss et al., 2017) who compared yeast WGDs to their orthologs from *Schizosaccharomyces pombe*. To further support this observation, we used PCA to examine HM formation for orthologs from species that diverged prior to the whole-genome duplication event (*Lachancea kluyveri* and *Zygosaccharomyces rouxii*). We looked at the mitochondrial translocon complex and at the transaldolase, which both contain HETs (see methods). We confirm that when one HM was observed in *S. cerevisiae*, at least one ortholog from pre-whole genome-duplication species also formed a HM (Figure 2. B-C). We also detected interactions between orthologs, suggesting that ability to interact has been preserved despite the millions of years of evolution separating these species.

We classified paralog pairs into four classes according to whether they show no interaction (NI, 13.78%), at least one HM but no HET (HM, 46.05%), only the HET (HET, 8.57%), or at least one of the HM and the HET (HM&HET, 31.6%) (Figure 2. D-E and S5). The number of interactions detected for all paralogous pairs is higher than expected among random pairs of proteins in large-scale screens (1% of PPIs tested are positive in (Tarassov et al., 2008)). This is in line with previous observations showing that paralogs frequently interact with each other (Ispolatov et al., 2005). Overall, most pairs forming HETs also form at least one HM (78.67%, Figure 2. E for the frequency for each type of duplication).

Previous observations showed that paralogs are enriched in protein complexes comprising more than two subunits, partly because complexes evolved by the initial establishment of self-interactions followed by duplication of the homomeric proteins (Musso et al., 2007; Pereira-Leal et al., 2007). We examined if HM&HET were part of complexes and could have evolved this way. We found 10 HETs among the yeast paralogs present in the Protein Data Bank (PDB), seven of them are in complexes with more than two distinct subunits and thus involve more proteins than just a pair of paralog. The remaining three are in complexes with only two distinct subunits and are therefore exclusively constituted of a paralog pair. We examined further the presence of HETs in complexes with more than two distinct subunits by combining data on 5,535 complexes (see methods). In 77 out of the 188 cases of HM&HET, we found both paralogs to be part of the same complex. Complexes containing more than two distinct proteins are thus unlikely to account for most of the HETs that derive from ancestral HMs. A large fraction of HM&HETs could therefore be simple heteromers of paralogs.

We observed that the correlation between HM and HET formation depends on whether paralogs are SSDs or WGDs (Figure 2. E). WGDs tend to more often form HETs when they form at least one HM, resulting in a larger proportion of HM&HET motif. We hypothesize that this could be at least partly due to the fact that most SSDs are older than WGDs (Table S1 and S2) and increasing protein sequence divergence with time could lead to the loss of the HET. We indeed find that among SSDs, those that have higher sequence identity are more likely to form HM&HET (Figure 2. F). The effect of time of duplication was also detectable for paralogs deriving from the whole-genome duplication. Recently, Marcet-Houben and Gabaldón (Marcet-Houben and Gabaldón, 2015; Wolfe, 2015) showed that WGDs likely have two distinct origins: actual duplication (generating true ohnologs) and hybridization between species (generating homeologs). For pairs where the ancestral state was HMs, we observe that true ohnologs have a tendency to form HET more frequently than for homeologs (Figure 2. E). Because homeologs had already diverged before the hybridization event, they are older than ohnologs, as shown by their lower pairwise sequence identity (Figure S5. A). This observation supports the fact that younger paralogs derived from HMs are more likely to form HETs than older ones. We further tested for an association between the age of SSDs established using gene phylogenies and their propensity to form HET, but we found no significant association, most likely due to the small number of pairs per age group (Figure S5. B).

Amino acid sequence conservation could also have a direct effect on the retention of HETs, independently of age. For instance, among WGDs (either within true ohnologs or within homeologs), which all have the same age in their own category, HM&HET pairs have higher sequence identity than HM pairs (Figure S5. C). This is also apparent for pairs of paralogs whose HM or HET structures have been solved by crystallography (Table S1). Indeed, we found that pairwise amino acid sequence identity was higher for HM&HET than for HM pairs for both entire proteins and their binding interfaces (Figure S6. A). Interestingly, the binding interface is more conserved than the rest of the protein for those forming HM&HET, suggesting a causal link between sequence identity at the interface and assembly of HM&HETs (Figure S6. B).

Heteromer formation correlates with function conservation

To test if the retention of HETs correlates with the functional similarity of HM and HM&HET paralogs, we used the similarity of Gene Ontology (GO) terms, known growth phenotypes of loss-of-function mutants and patterns of genome-wide genetic interactions. These features represent the relationship of genes with cell growth in specific conditions and the gene-gene relationships underlying cell growth. The use of GO terms could bias the analysis because these are often predicted based on sequence features. However, phenotypes and genetic interactions are derived from unbiased experiments because interactions are tested without a priori consideration of a gene’s function (Costanzo et al., 2016). We found that HM&HET pairs are more similar than HM for SSDs (Figure 3). We observe the same tendency for WGDs, although some of the comparisons are either marginally significant or non-significant (Figure 3, comparison between true ohnologs and homeologs in Figure S7). Overall, the retention of HETs after the duplication of HMs is correlated with a weaker divergence of function.

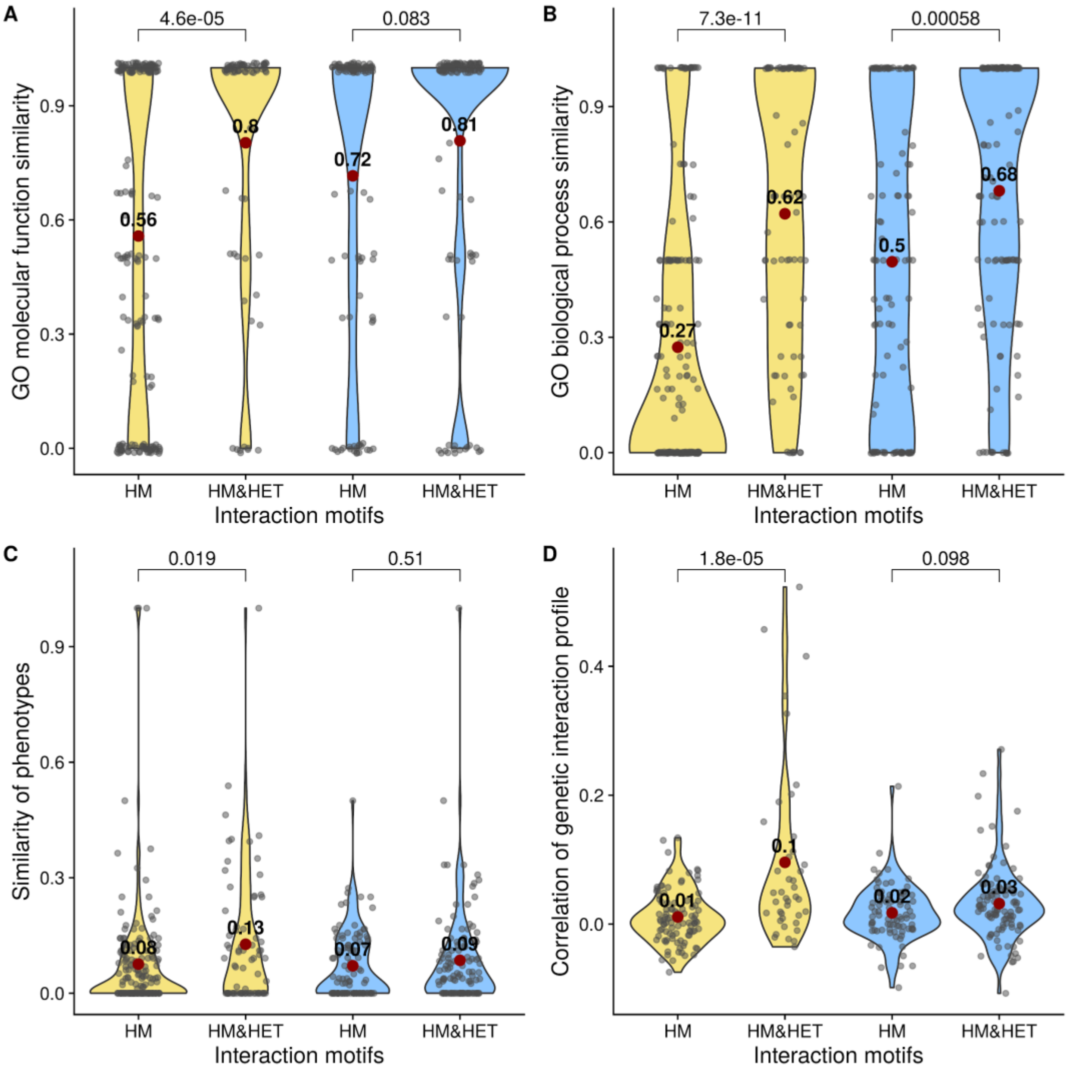

Figure 3: Heteromers are functionally more similar than homomers across paralogs that derive from homomers.

The similarity score is the proportion of shared terms across pairs of paralogs for (**A**) GO molecular functions, (**B**) GO biological processes and (**C**) gene deletion phenotypes. (**D**) shows the correlation of genetic interaction profiles. p-values are from Wilcoxon tests. SSDs are shown in yellow and WGDs in blue. Mean values of the distributions are represented by red points.

Pleiotropy contributes to the maintenance of heteromers

Since molecular interactions between paralogs predate their functional divergence, it is possible that physical association by itself affects the retention of functional similarity among paralogs. Because of this, any feature of the paralogs that contributes to the maintenance of the HET state could have a strong impact on the fate of new genes emerging from the duplication of a HM. A large fraction of HMs and HETs use the same binding interface

(Bergendahl and Marsh, 2017). This shared interface could cause mutations to have pleiotropic effects on HMs and HETs (Figure 1). If we assume that HMs need to self-interact in order to perform their function, it is expected that natural selection would favor the maintenance of self-assembly. Negative selection on HM interfaces will thus also preserve HET interfaces, preventing the loss of the HET.

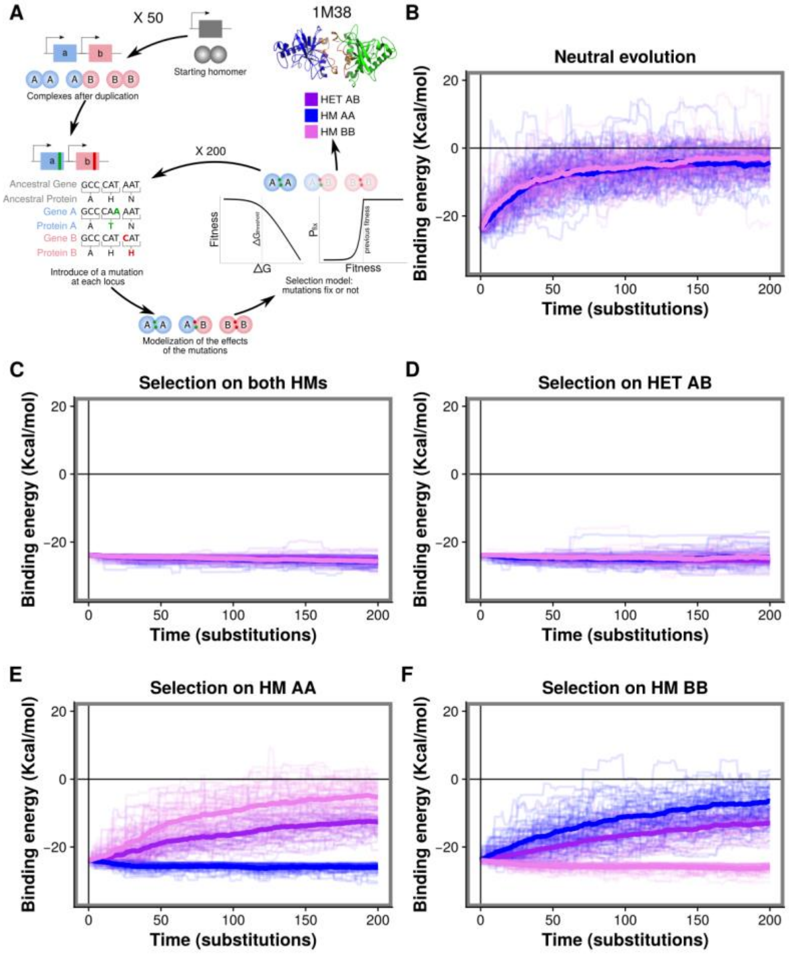

Figure 4: Selection to maintain homomers also maintains heteromers.
(**A**) General workflow of the simulation experiments. The duplication of a gene encoding a homomeric protein and the evolution of the complexes is simulated by applying mutations to the corresponding chains A and B. Only mutations that would require a single nucleotide change are allowed and stop codons are disallowed. After introducing mutations, the selection model is applied to complexes and mutations are fixed or lost. (**B to F**) The binding energy of the HMs and the HET resulting from the duplication of a HM (PDB: 1M38) is followed through time under different selection regimes applied on protein chain stability and binding energy. More positive values for binding energy indicate less favorable binding and more negative values indicate more favorable binding. (**B**) Neutral evolution: the accumulation and neutral fixation of mutations. (**C**) Selection on both HMs: selection maintains only the two HMs while the HET evolves neutrally. (**D**) Selection on HET: selection maintains the HET while the HMs evolve neutrally. (**E**) Selection on HM BB or (**F**) HM AA: selection maintains one HM while the HET and the other HM evolve neutrally. Mean binding energies among replicates are shown in thicker lines and the individual replicates

are shown with thin lines. Fifty replicate populations are monitored in each case and followed for 200 substitutions. After 200 substitutions, the neutral model corresponds to an almost complete loss of spontaneous binding. PDB structure 1M38 was visualized with PyMOLL (Schrödinger, 2015).

We tested the correlated selection model using *in silico* evolution of HM and HET protein complexes (Figure 4. A). We used a set of six representative high-quality structures of HMs (Dey et al., 2018). We evolved these HM complexes *in silico* by duplicating them and by following the binding energies of the two HMs and of the HET. We let mutations occur at the binding interface 1) in the absence of selection (neutral model), 2) in the presence of negative selection maintaining only one HM or 3) both HMs. In both cases, we apply no selection on binding energy of the HET. In the fourth scenario, we apply selection on the HET but not on the HMs to examine if selection maintaining the HET could also favor the maintenance of HMs. Mutations that have deleterious effects on the complex under selection were lost or allowed to fix with exponentially decaying probability depending on the fitness effect (see methods) (Figure 4. A).

We find that neutral evolution leads to the destabilization of all complexes (Figure 4. B), as is expected given that there are more destabilizing mutations than stabilizing ones, both in terms of binding energy and chain stability (Brender and Zhang, 2015; Guerois et al., 2002). However, selection to maintain one HM or both HMs significantly slows down the loss of the HET (Figure 4B). Interestingly, the HET is being destabilized more slowly than the second HM when only one HM is under negative selection, which could explain why for some paralog pairs, only one HM and HET are conserved (Figure S8). The reciprocal situation is also true, i.e. negative selection on HET significantly decelerates the loss of stability of HMs. These observations hold when simulating the evolution after duplication of six structures (Figure S9).

By examining the effects of single mutants (only one of the loci gets a non-synonymous mutation at the interface) on HMs and HETs, we see that the evolutionary paths followed by the simulations are caused by the correlated effects of mutations (Figure 5). The majority of mutations (77%) appear to have effects with the same directionality on both complexes, either being stabilizing or destabilizing. The remaining fraction of mutations (23%) specifically destabilizes either the HETs or HMs and these mutations tend to have small effects, further reducing the likelihood of mutations that would rapidly disrupt a specific complex. Again, the results hold for the six structures tested (Figure S10). Additionally, mutations tend to have greater effects on the binding energy of HMs compared to HETs, presumably because of the presence of a mutation on both chains simultaneously (Figure S11. A-B). Likewise, a higher percentage of double mutants (one at the interface of each duplicated locus) are fixed during the simulations when negative selection is applied to the HET rather than to the two HMs (Figure S12). This effect is more pronounced for mutants having effects with opposite effects on the HMs, which could partially cancel out when applied to chains forming a HET. The particular capability of HETs to accommodate mutations with opposite effects in each of the monomers without disrupting interactions results in a higher robustness with respect to protein complex assembly.

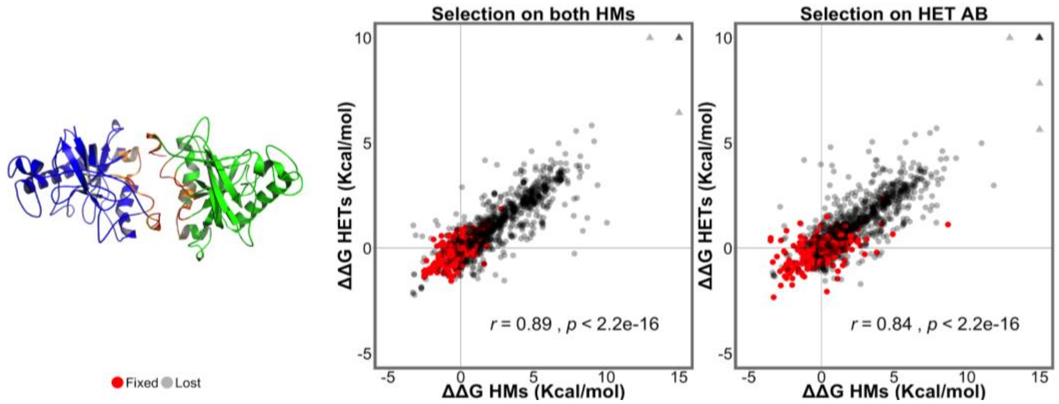

Figure 5: Mutations have pleiotropic effects on homomers and heteromers.
Effects of amino acid substitutions on one protein chain on the binding energy of homomers (x-axis) and heteromers (y-axis) when selection is applied on both homomers or on the heteromer. Fixed mutations are indicated in red whereas lost mutations are indicated in gray. Colors indicate whether the mutations were fixed (red) or lost (gray). Triangles indicate outliers that are outside the plotting region. PDB structure 1M38 was visualized with PyMOL (Schrödinger, 2015).

Regulatory evolution may break down molecular pleiotropy

The results from simulations show that the loss of HET after the duplication of a HM occurs at a slow rate and that specific mutations are required for HETs to be destabilized. However, the simulations only consider the evolution of binding interfaces, which limits the modification of interactions to a subset of all mutations that can ultimately affect PPIs. Other mechanisms responsible for the loss of paralog interactions could involve transcriptional regulation or cell compartment localization such that paralogs are not present at the same time or in the same cell compartment. To test this, we measured the correlation coefficient of expression profiles of paralogs across growth conditions using previously published data (Ihmels et al., 2004). These expression profiles are more correlated for SSDs forming HM&HET than for those forming only HM (p-value = 0.00025, Figure 6. A). A similar tendency is observed for WGDs, but the difference is not statistically significant (p-value = 0.18, Figure 6. A). The differences of expression correlation between HM&HET and HM are likely due to *cis* regulatory divergence between paralogs because we observe similar patterns when considering the similarity of transcription factor binding sites (Figure 6. B). When we split WGDs into homeologs and true ohnologs, even though the former are more co-expressed (Figure S13. A), we do not observe differences of co-expression between interaction motif (Figure S13. B). Because we found that sequence identity was correlated with the probability of observing HET and that sequence identity can be correlated with the co-expression of paralogs, we tested if these factors had an independent effect on HET formation. For SSDs, both sequence identity and co-expression show significant effects on HM&HET formation (Table S5. B), but for WGDs, only the percentage of identity seems to impact HM&HET formation (Figure 6. C, Table S5. B). Similar results were obtained for both homeologs and true ohnologs(Figure S13. C).

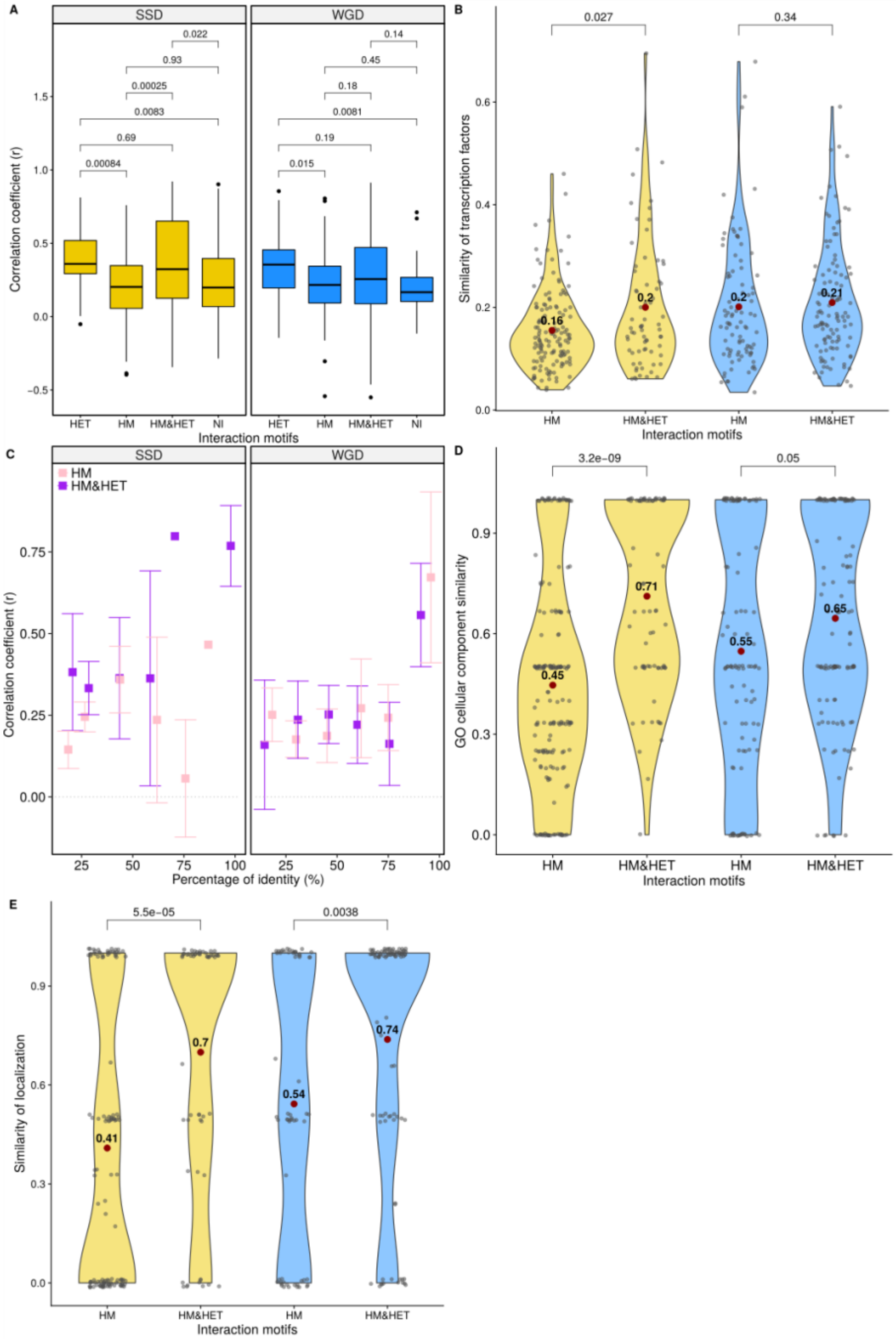

Figure 6: Paralogs expression and consequences on interaction motifs.
(**A**) Correlation coefficients (Spearman r) between the expression profile of paralog pairs across growth conditions (Ihmels et al. 2004) are compared among the different interaction motifs for SSDs (yellow) and WGDs (blue). (**B**) The similarity of transcription factor binding sites (fraction of sites shared) is compared among classes of SSDs (yellow) and WGDs (blue). (**C**) Correlation of expression between paralogs forming HM&HET (purple) or only HM (pink) as a function of their amino acid sequence identity. The data was binned into six equal categories. (**D**)

Similarity of GO cellular component and (**E**) GFP-based localization for each SSDs (yellow) and WGDs (blue). P-values are from Wilcoxon tests.

Finally, we find that HM&HET paralogs are more similar than HM for both SSDs and WGDs in terms of cellular compartments (GO) (Figure 6. D) and of cellular localization derived from experimental data (Figure 6. E). Taking into account the two origins of WGDs (homeologs and true ohnologs), we observe the same significant differences of localisation and the same tendency (but not significant) for the similarity of cellular compartments between HM and HM&HET pairs (Figure S13. D-F). The comparison of the correlation of expression profiles, of transcription factors binding sites, of GO cellular component and localisation similarities show that changes in gene and protein regulation could prevent the interaction between paralogs that derive from ancestral HMs, reducing the role of pleiotropy in maintaining their associations.

Discussion

Upon duplication, the properties of proteins are inherited from their ancestor, which may affect how paralogs subsequently evolve. Here, we examined the extent to which physical interactions between paralogs are preserved after the duplication of HMs and how these interactions affect functional divergence. Using reported PPI data, crystal structures and PCA experiments, we found that paralogs originating from ancestral HMs are more likely to functionally diverge if they do not interact with each other. We propose that non-adaptive mechanisms could play a role in the retention of physical interactions and in turn, impact functional divergence. By developing a model of *in silico* evolution of PPIs, we found that molecular pleiotropic effects of mutations on binding interfaces can constrain the maintenance of HET complexes even if they are not under selection. We hypothesize that this non-adaptive constraint could play a role in slowing down the divergence of paralogs but that it could be counteracted by regulatory evolution.

The proportions of HMs and HETs among yeast paralogs were first studied more than 15 years ago (Wagner, 2003). From this first study, it was suggested that most paralogs forming HETs do not have the ability to form HMs and thus, that evolution of new interactions was rapid. Since then, many PPI experiments have been performed (Chatr-Aryamontri et al., 2017; Kim et al., 2019; Stark et al., 2006; Stynen et al., 2018) and the resulting global picture is different. We found that most of the paralogs forming HETs also form HMs, suggesting that interactions between paralogs are inherited rather than gained *de novo.* This idea is supported by models predicting interaction losses to be much more likely than interaction gains (Gibson et al., 2009; Presser et al., 2008). It is also in line with models proposing that proteins are more likely to form HMs than to interact with other proteins because the probability of two protein interfaces to have a matching pattern is higher for a protein with itself (Lukatsky et al., 2007; Monod et al., 1965). Accordingly, the HM&HET state can be more readily achieved by the duplication of an ancestral HM and the subsequent loss of one of the two HMs than by the duplication of a monomeric protein followed by the gain of the HM and of the HET. Interaction paralogs are therefore more likely to derive from ancestral HMs, as shown by (Diss et al. 2017). For some pairs of *S. cerevisiae* paralogs presenting the HM&HET motif, we indeed detected HM formation of their orthologs from pre-whole-genome duplication species, supporting the model by which self-interaction and cross-interactions is inherited from the duplication. We did not detect HMs in both pre-whole-genome duplication species, which may reflect the incorrect expression of these proteins in *S. cerevisiae* rather than their lack of interactions. Overall, we conclude that to lose HET complexes, interacting paralogs would have to diverge in sequence of the binding interfaces, or by other mechanisms. The role of the divergence of binding interface is supported by our results showing that interacting paralogs have a higher sequence identity both at the full sequence and at the binding interface level than paralogs that do not interact.

We observed an enrichment of HMs among duplicated proteins compared to singletons. This observation was reported for various interactomes (Ispolatov et al., 2005; Pereira-Leal et al., 2007; Pérez-Bercoff et al., 2010; Yang et al., 2003). Also, analyses of PPIs from large-scale experiments have shown that interactions between paralogous proteins are more common than expected by chance alone (Ispolatov et al., 2005; Musso et al., 2007; Pereira-Leal et al., 2007). Several adaptive hypotheses have been put forward to explain the over-representations of interacting paralogous proteins. Altogether, these hypotheses imply that the retention of paralogs after duplication is more likely for HM proteins due to the enhanced probability of gain of functions (although subfunctionalization is also possible). For example, symmetrical HM proteins could have key advantages over monomeric ones for protein stability and regulation (André et al., 2008; Bergendahl and Marsh, 2017). Levy and Teichmann (Levy and Teichmann, 2013) suggested that the duplication of HM proteins serves as a seed for the growth of protein complexes. These duplications would allow the diversification of complexes by creating asymmetry and the recruitment of other proteins, leading to their specialization. It is also possible that the presence of HET itself offers a rapid way to evolve new functions. For instance, Bridgham et al. (Bridgham et al., 2008) showed that degenerative mutations in one copy of an heterodimeric transcription factor can switch its role to become a repressor. Regulatory mechanisms could therefore rapidly evolve this way (Bridgham et al., 2008; De Smet et al., 2013; Kaltenegger and Ober, 2015). Finally, Natan et al. (Natan et al., 2018) showed that cotranslational folding can be a problem for homomeric proteins because of premature assembly, particularly for proteins with interfaces closer to their N-terminus. The replacement of these HMs by HETs could solve this issue by separating the translation of the proteins to be assembled on two distinct mRNAs.

Non-adaptive mechanisms could also be at play to maintain HETs. Diss et al. (Diss et al., 2017) recently showed that in the absence of their paralog, some proteins are unstable and lose their capacity to interact with other proteins. The fact that a paralog can be unstable in the absence of its sister copy appears to be enriched among paralogs forming HET, suggesting that the individual proteins depend on each other due to their physical interactions (Diss et al., 2017). Independent observations by (DeLuna et al., 2010) also showed that the deletion of paralogs was sometimes associated with the degradation of their sister copy, particularly among HET paralogs. The Diss et al. and DeLuna et al. observations led to the proposal that paralogs could accumulate complementary degenerative mutations at the structural level after the duplication of a HM (also see Kaltenegger and Ober, 2015). This scenario would lead to the maintenance of the HET because destabilizing mutations in one chain can be compensated by stabilizing mutations in the other, keeping binding energy near the optimum. Compensatory effects cannot happen for HMs because the two copies of the protein are identical. Our simulations are consistent with the compensatory model where some pairs of mutations in the two chains of the HET have opposite effects on binding energy. On the long term, the accumulation of opposite effect mutations could maintain the HET as it remains the only functional unit capable of performing the ancestral function. However, our data suggest that most (89%) of dependent paralogs that form HET in (Diss et al., 2017) also form at least one HM, suggesting that the loss of both HMs is not required for dependency. Further experiments will be needed to fully determine the likelihood of the dependency model and in which conditions it could take place.

Our simulated evolution of the duplication of HMs leads to the proposal of a simple mechanism for the maintenance of HET that does not require adaptive mechanisms. A large fraction of HMs and HETs use the same binding interface (Bergendahl and Marsh, 2017) and as a consequence, negative selection on HM interfaces will also preserve HET interfaces. Our results show that mutations have correlated effects on HM and HET, which slows down the evolution of independent complexes. Although mutations with specific destabilization effects are available, their small magnitude would not suffice to alter greatly the balance of complexes. This limitation in terms of mutational effects that could separate complexes could therefore be another non-adaptive mechanism for the long-term maintenance of heteromers.

One of our observations is that WGDs present proportionally more HM&HET motifs than SSDs. We propose that this is at least partly due to the age of paralogs. This proposal was based on the fact that SSDs in yeast are in general older than WGDs and even among WGDs, homeologs show less frequent HM&HET than HMs compared to true ohnologs. However, the mode of duplication itself could also impact HET maintenance. For instance, upon a whole-genome duplication event, all subunits of complexes are duplicated at the same time, which can lead to differential rates of gene retention across functional categories. Consistent with this, previous studies have found that WGDs are enriched in proteins that are subunits of protein complexes when compared with SSDs (Hakes et al., 2007; Papp et al., 2003). A possible explanation for this trend is that while the duplication of a single subunit of a complex may be deleterious by perturbing stoichiometry, duplication of all subunits at once may not cause such disadvantage since balance is preserved (Birchler and Veitia, 2012; Rice and McLysaght, 2017). Another possibility is that WGDs are maintained due to their dosage effects, which results in selection for the maintenance of their protein function (Gout and Lynch, 2015; Thompson et al., 2016) and at the same time, favors the maintenance of HETs.

We noticed a significant fraction of paralogs forming only HMs but not HET, including some cases of recent duplicates, indicating the forces that maintain HETs can be overcome. Duplicate genes in yeast and other model systems often diverge quickly in terms of transcriptional regulation (Li et al., 2005; Thompson et al., 2013) due to *cis* regulatory mutations (Dong et al., 2011). Because transcriptional divergence of paralogs can directly change PPI profiles, expression changes would be able to rapidly change a motif from HM&HET to HM. Indeed, Gagnon-Arsenault et al. (Gagnon-Arsenault et al., 2013) showed that switching the coding sequences between paralogous loci was sometimes sufficient to change PPI specificity in living cells. Protein localization can also be an important factor affecting the ability of proteins to interact (Rochette et al., 2014). Here, we found that paralogs that derive from HMs and that have lost their ability to form HETs are less co-regulated and co-localized.

This divergence suggests that regulatory evolution could play a role in relieving duplicated homomeric proteins from the correlated effects of mutations affecting shared protein interfaces. Although we did not explore this possibility, it is likely that pairs of paralogs that are co-regulated and co-localized but that do not form HET show other mechanisms that prevent their association. For instance, it is possible that paralogs could gain or lose interaction domains (Nasir et al., 2014), which could potentially bypass the constraints imposed by homologous interaction interfaces and drive the functional differentiation of paralogs.

Overall, our analyses show that the duplication of self-interacting proteins creates paralogs whose evolution is constrained in ways that are not expected for monomeric paralogs. Further analyses of structural data as well as transcriptional and localization data will allow to disentangle the causal or correlated roles of these individual mechanisms in the evolution of PPIs and in the functional evolution of paralogs.

Material and Methods

1. Characterization of paralogs in *S. cerevisiae* genome

1.1 Classification of paralogs by mechanism of duplication

We classified duplicated genes in three categories according to their mechanism of duplication: small-scale duplicate, SSD; whole-genome duplicate, WGD (Byrne and Wolfe, 2005); and double duplicate, 2D (SSD and WGD). We removed WGDs from the paralogs defined in (Guan et al., 2007) to generate the list of SSDs. If one of the two paralogs of a SSD pair is associated to another paralog in a WGD pair, this paralog was considered as 2D (Tables S6 and S7). To decrease the potential bias from multiple duplication events, we removed those from the data on interaction motifs. We used data from (Marcet-Houben and Gabaldón, 2015) to identify WGDs that likely originated from allopolyploidization (homeologs) and true ohnologs.

1.2 Sequence similarity

The amino-acid sequences of each pair of proteins were obtained from the Saccharomyces Genome Database (SGD) (S288C strain version 2007-03-01) and their pairwise identity were computed using the *pairwise Alignment* function (default parameters) from the R Biostrings package (Pagès et al., 2018). The percentage of identity was estimated using the *pid* function (option type = “PID1”) from the same package.

1.3 Age of duplication

We estimated the age of SSDs using gene phylogenies. If duplication of gene A led to the formation of two paralogs A1/A2, before the speciation event separating two species, orthologs A1 are expected to be more similar than paralogs A1/A2 (Li et al., 2003). Based on this principle, we estimated the age of the SSDs using gene phylogenies extracted from PhylomeDB (Huerta-Cepas et al., 2008) and imported in the R Tidytree package (Yu, 2019; Yu et al., 2018, 2017). For a gene A1, the common node between A1 and its paralog A2 was identified and the next more recent node leading to A1 was retained. Among the descending clades of the retained node, the more distant species from *S. cerevisiae* was selected and an age group was defined following a species tree (see Figures S14 and S15, Tables S8 and S1). If the two paralogs were not found in the same tree, the node assigned was the oldest for each tree. If the age group determined from the gene phylogeny of paralog A1 did not match the age group of paralog A2, it is potentially due to one paralog that disappeared in the more distant clade; therefore, the oldest age group was retained. The age group assignment was then validated by checking whether for at least one species of this group, the ortholog gene had at least two paralogs. If it was not the case, the age group assignment was decreased by one and retested until a species was found with at least two paralogs within the group. This custom method was tested for WGDs that occured in the common ancestor of post-WGD species or pre-KLE (*Kluyveromyces*, *Lachancea*, and *Eremothecium*) branch defined by the age groups number 1 and 2 respectively (Figure S14). The error rate was estimated to be less than 1%.

1.4 Function, transcription factor binding sites, localization and protein complexes

We obtained GO terms (GO slim) from SGD (Cherry et al., 2012) in September 2018. We removed terms corresponding to missing data and created a list of annotations for each gene. These lists were compared to measure the extent of similarity between two members of a pair. We calculated the similarity of molecular function, cellular component and biological process taking the number of GO terms in common divided by the total number of unique GO terms of the two paralogs combined (Jaccard index). We also compared the same way the transcription factor binding sites using YEASTRACT data (Teixeira et al., 2018, 2006), cellular localizations extracted from YeastGFP database (Huh et al., 2003), and set of phenotypes associated with the deletion of the paralogs (data from SGD in September 2018). We kept only information with specific phenotypes (a feature observed and a direction of change relative to wild type). We compared the pairwise correlation of genetic interaction profiles using the genetic interaction profile similarity (measured by Pearson's correlation coefficient) of non-essential genes available in TheCellMap database (version of March 2016) (Usaj et al., 2017). We used the median of correlation coefficients if more than one value was available for a given pair. Non-redundant set of protein complexes was derived from the Complex Portal (Meldal et al., 2015), the CYC2008 catalogue (Pu et al., 2009, 2007) and Benschop et al., (Benschop et al., 2010).

2. HMs and HETs from the literature and databases

We used HM and HET reported in BioGRID version BIOGRID-3.5.166 (Chatr-Aryamontri et al., 2017, 2013). We selected physical interactions by using data derived from the following detection methods: Affinity Capture-MS, Affinity Capture-Western, Reconstituted Complex, Two-hybrid, Biochemical Activity, Co-crystal Structure, Far Western, FRET, Protein-peptide, PCA and Affinity Capture-Luminescence. It is possible that some HMs or HETs are absent from the database because they have been tested but not detected. This negative information is not reported. We therefore attempted to discriminate non-tested interactions from negative interactions the following way. We obtained the list of Pubmed IDs from studies in which at least one HM is reported to identify those using methods that can detect HMs. We then examined every study individually to examine if a given HM was reported and inferred that the HM existed if reported (coded 1). If it was absent but other HMs were reported and that we confirmed the presence of the protein as a bait and a prey, we considered the HM absent (coded 0). If the protein was not present in both baits and preys, we considered the HM as not tested (coded NA). We proceeded in the same way for HET of paralogs. If the HM was detected in the Protein Data Bank (PDB) Data (Berman et al., 2000), we inferred that it was present (coded 1). If the HM was not detected but the monomer was reported, it likely that there is no HM for this protein and it was thus considered non-HM (coded 0). If there was no monomer and no HM, the data was considered as missing (coded NA). We proceeded the same way for HETs. Data on genome-wide HM screens were obtained from (Kim et al., 2019; Stynen et al., 2018). The two methods relied on Protein-fragment complementation assays (PCA), the first one using the dihydrofolate reductase (DHFR) enzyme and the second on a fluorescent protein as reporter (also known as Bimolecular fluorescence complementation or BiFC). We discarded results from proteins marked as problematic by (Rochette et al., 2014; Tarassov et al., 2008) from (Stynen et al., 2018) and false positive identified by (Kim et al., 2019). These were considered as missing data from these studies. We examined all proteins tested and considered them as HM if they were reported as positive (coded 1) and considered non-HM if tested but not reported as positive (coded 0).

3. Experimental Protein-fragment complementation assay

We performed a screen usingPCA based on DHFR (Tarassov et al., 2008) following standard procedures (Rochette et al., 2014; Tarassov et al., 2008)

3.1 DHFR strains

We identified 453 pairs of WGDs and 188 pairs of SSDs that were present in the Yeast Protein Interactome Collection (Tarassov et al., 2008) and another set of 143 strains constructed by (Diss et al., 2017). We retrieved strains from the collection (Tarassov et al., 2008) and we tested their growth on YPD supplemented with nourseothricin (NAT) for DHFR F[1,2] strains and hygromycin B (HygB) for DHFR F[3] strains. We confirmed the presence of the DHFR fragments at the correct location by colony PCR using specific forward Oligo-C targeting a few hundred base pairs upstream of the fusion and the reverse complement oligonucleotide ADH-term (Table S9). Cells from colonies were lysed in 40 µL of 20 mM NaOH for 20 min at 95°C. Tubes were centrifuged for 5 min at 4000 rpm and 2.5 µL of supernatant was added to a PCR mix composed of 16.85 µL of DNAse free water, 2.5 µL of 10X buffer (Bioshop, Burlington, Canada), 1.5 µL of 25 mM MgCl2, 0.5 µL of 10 mM dNTP (BIO BASIC, Markham, Canada), 0.15 µL of 5 U/µL Taq DNA polymerase (BioShop, Burlington, Canada), 0.5 µL of 10 µM Oligo-C and 0.5 µL 10 µM ADH-term. The initial denaturation was performed for 5 min at 95°C and was followed by 35 cycles of 30 sec of denaturation at 94°C, 30 sec of annealing at 55°C, 1 min of extension at 72°C and 3 min of a final extension at 72°C. We confirmed 1769 of the 1904 strains from the DHFR collection and 117 strains out of the 143 from (Diss et al., 2017) (Table S6, S7, and S9)

The missing or non-validated strains were constructed *de novo* using the standard DHFR strain construction protocol (Michnick et al., 2016; Rochette et al., 2015). The DHFR fragments and associated resistance cassettes were amplified from plasmids pAG25-linker-DHFR-F[1,2]-ADHterm (marked with nourseothricin N-acetyl-transferase) and pAG32-linker-DHFR-F[3]-ADHterm (marked with hygromycin B phosphotransferase) (Tarassov et al., 2008) using oligonucleotides defined in (Table S9). PCR mix was composed of 16.45 µL of DNAse free water, 1 µL plasmid at 10 ng/µL, 5 µL of 5X Kapa Buffer, 0.75 µL of 10 mM dNTPs, 0.3 µL of Kapa HiFi HotStart DNA polymerase at 1 U/µL and 0.75 µL of both forward and reverse oligos at 10 µM. The initial denaturation was performed for 5 min at 95°C and was followed by 32 cycles of 20 sec of denaturation at 98°C, 15 sec of annealing at 64.4°C, 2.5 min of extension at 72°C and 5 min of a final extension at 72°C.

We performed strain construction in BY4741 (MAT**a** *his3Δ leu2Δ met15Δ ura3Δ*) and BY4742 (MATα *his3Δ leu2Δ lys2Δ ura3Δ*) competent cells prepared as in (Gagnon-Arsenault et al., 2013) for the DHFR F[1,2] and DHFR F[3] fusions respectively. Competent cells (20 µL) were combined with 8 µL of PCR product (~0.5-1 µg/µL) and 100µL of Plate Mixture (PEG3350 40%, 100 mM of LiOAc, 10 mM of Tris-Cl at pH 7.5 and 1 mM of EDTA). The mixture was vortexed and incubated at room temperature without agitation for 30 min. Heat shock was then performed after adding 15 µL of DMSO and mixing thoroughly by incubating in a water bath at 42°C for 15-20 min. Following the heat shock, cells were spun down at 400g for 3 min. Supernatant was removed by aspiration and cell pellets were resuspended in 100 µL of YPD (1% yeast extract, 2% tryptone, 2% glucose). Cells were allowed to recover from heat shock for 4 hours at 30°C before being plated on YPD plates (YPD with 2% agar) with 100 µg/mL of NAT for DHFR F[1,2] strains or with 250 µg/mL of HygB for DHFR F[3] strains. Transformations were incubated at 30°C for 3 days (Table S10). The correct integration of DHFR fragments was confirmed by colony PCR as described above. At the end, we were able to reconstruct and validate 152 new strains (Table S6 and S7). From all available strains, we selected pairs of paralogs for which we had both proteins tagged with both DHFR fragments (four differents strains per pairs). This resulted in 1268 strains corresponding to 317 pairs of paralogs (Table S6 and S7). We next discarded pairs considered as forming false positives by (Tarassov et al., 2008), which resulted in 286 pairs.

3.2 Construction of DHFR plasmids for orthologous genes

For the plasmid-based PCA, Gateway cloning-compatible destination plasmids pDEST-DHFR F[1,2] (TRP1 and LEU2) and pDEST-DHFR F[3] (TRP1 and LEU2) were constructed based on the CEN/ARS low-copy yeast two-hybrid (Y2H) destination plasmids pDEST-AD (TRP1) and pDEST-DB (LEU2) (Rual et al., 2005). A DNA fragment having I-CeuI restriction site was amplified using DEY001 and DEY002 primers (Table S9) without template and another fragment having PI-PspI/I-SceI restriction site was amplified using DEY003 and DEY004 primers (Table S9) without template. pDEST-AD and pDEST-DB plasmids were each digested by PacI and SacI and mixed with the I-CeuI fragment (destined to the PacI locus) and PI-PspI/I-SceI fragment (destined to the SacI locus) for Gibson DNA assembly (Gibson et al., 2009) to generate pDN0501 (TRP1) and pDN0502 (LEU2). Four DNA fragments were then prepared to construct the pDEST-DHFR F[1,2] vectors: (i) a fragment containing ADH1 promoter; (ii) a fragment containing Gateway destination site; (iii) a DHFR F[1,2] fragment; and (iv) a backbone plasmid fragment. The ADH1 promoter fragment was amplified from pDN0501 using DEY005 and DEY006 primers (Table S9) and the Gateway destination site fragment was amplified from pDN0501 using DEY007 and DEY008 primers (Table S9). The DHFR-F[1,2] fragment was amplified from pAG25-linker-DHFR-F[1,2]-ADHterm (Tarassov et al., 2008) using DEY009 and DEY010 primers (Table S9). The backbone fragment was prepared by restriction digestion of pDN0501 or pDN0502 using I-CeuI and PI-PspI and purified by size-selection. The four fragments were assembled by Gibson DNA assembly where each fragment pair was overlapping with more than 30 bp, producing pHMA1001 (TRP1) or pHMA1003 (LEU2). The PstI–SacI region of the plasmids was finally replaced with a DNA fragment containing an amino acid flexible polypeptide linker (GGGGS) prepared by PstI/SacI double digestion of a synthetic DNA fragment DEY011 to produce pDEST-DHFR F[1,2] (TRP1) and pDEST-DHFR F[1,2] (LEU2). The DHFR F[3] fragment was then amplified from pAG32-linker-DHFR-F[3]-ADHterm with DEY012 and DEY013 primers (Table S9), digested by SpeI and PI-PspI, and used to replace the SpeI–PI-PspI region of the pDEST-DHFR F[1,2] plasmids, producing pDEST-DHFR F[3] (TRP1) and pDEST-DHFR F[3] (LEU2) plasmids. In this study, we used pDEST-DHFR F[1,2] (TRP1) and pDEST-DHFR F[3] (LEU2) for the plasmid-based DHFR PCA. After Gateway LR cloning of Entry Clones to these destination plasmids, the expression plasmids encode protein fused to the DHFR fragments via an NPAFLYKVVGGGSTS linker.

We obtained the orthologous gene sequences for the mitochondrial translocon complex and the transaldolase proteins of *Lachancea kluyveri* and *Zygosaccharomyces rouxii* from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005). Each ORF was amplified using oligonucleotides listed in Table S9. We used 300 ng of purified PCR product to set a BPII recombination reactions (5 µL) into the Gateway Entry Vector pDONR201 (150 ng) according to the manufacturer's instructions (Invitrogen #11789-020, Carlsbad, USA). BPII reaction mix wa incubated overnight at 25°C. The reaction was inactivated with proteinase K. The whole reaction was used to transform MC1061 competent *E. coli* cells (Invitrogen #C663-03, Carlsbad, USA), followed by selection on solid 2YT media (1% Yeast extract, 1.6% Tryptone, 0.2% Glucose, 0.5% NaCl and 2% Agar) with 50 mg/L of kanamycin (BioShop #KAN201.10, Burlington, Canada) at 37°C. Positive clones were detected by PCR using an ORF specific oligonucleotide and a general pDONR201 primer (Table S9). We then extracted the positive clone using minipreps for downstream application.

LRII reaction were performed by mixing 150 ng of the Entry Clone and 150 ng of expression plasmids (pDEST-DHFR F[1,2]-TRP1 or pDEST-DHFR F[3]-LEU2) according to manufacturer's instructions (Invitrogen #11791019, Carlsbad, USA). The reactions were incubated overnight at 25°C and inactivated with proteinase K. We used the whole reaction to transform MC1061 competent *E. coli* cells, followed by selection on solid 2YT media supplemented with 100 mg/L ampicillin (BioShop #AMP201.25, Burlington, Canada) at 37°C. Positive clones were confirmed by PCR using the ORF specific primer and a plasmid universal primer. The sequence-verified expression plasmids bearing the orthologous fusions with DHFR F[1,2] and DHFR F[3] fragments were used to transform the yeast strains YY3094 (MATa *leu2-3,112 trp1-901 his3-200 ura3-52 gal4Δ gal80Δ LYS2::P*$_{GAL1}$*-HIS3 MET2::P*$_{GAL7}$*-lacZ cyh2*^{R} *can1Δ::P*$_{CMV}$*-rtTA-KanMX4*) and YY3095 (MATα *leu2-3,112 trp1-901 his3-200 ura3-52 gal4Δ gal80Δ LYS2::P*$_{GAL1}$*-HIS3 MET2::P*$_{GAL7}$*-lacZ cyh2*^{R} *can1Δ::T*$_{ADH1}$*-P*$_{tetO2}$*-Cre-T*$_{CYC1}$*-KanMX4*), respectively. The strains YY3094 and YY3095 were generated from BFG-Y2H toolkit strains RY1010 and RY1030 (Yachie et al., 2016), respectively, by restoring their wild type *ADE2* genes. The *ADE2* gene was restored by homologous recombination of the wild type sequence cassette amplified from the laboratory strain BY4741 using primers DEY014 and DEY015 (Table S9); SC -ade plates (Table S10) were used to screen successful transformants.

3.3 DHFR PCA experiments

Three DHFR PCA experiments were performed, hereafter referred to as PCA1, PCA2 and PCA3. The configuration of strains on plates and screening was performed using robotically manipulated pin tools (BM5-SC1, S&P Robotics Inc., Toronto, Canada (Rochette et al., 2015)). We first organized haploid strains in 384 colony arrays containing a border of control strains using a cherry-picking 96-pin tool (Figure S16). We constructed four haploid arrays corresponding to paralog 1 and 2 (P1 and P2) and mating type: MAT**a** P1-DHFR F[1,2]; MAT**a** P2-DHFR F[1,2] (on NAT media, Table S10); MATα P1-DHFR F[3] ; MATα P2-DHFR F[3] (on HygB media, Table S10). Border control strains known to show interaction by PCA (MAT**a** *LSM8*-DHFR F[1-2] and MATα *CDC39*-DHFR F[3]) were incorporated respectively in all MAT**a** DHFR F[1,2] and MATα DHFR F[3] plates in the first and last columns and rows. The strains were organized as described in Figure S16. The two haploid P1 and P2 384 plates of the same mating type were condensated into a 1536 colony array using a 384-pintool. The two 1536 arrays (one bait, one prey) were crossed on YPD to systematically test P1-DHFR F[1,2] / P1-DHFR F[3], P1-DHFR F[1,2] / P2-DHFR F[3] , P2-DHFR F[1,2] / P1-DHFR F[3] and P2-DHFR F[1,2] / P2-DHFR F[3] interactions in adjacent positions. We performed two rounds of diploid selection (S1 to S2) by replicating the YPD plates onto YPD containing both NAT+HygB and growing for 48 hours. The resulting 1536 diploid plates were replicated twice on DMSO control plates (for PCA1 and 2) and twice on the selective MTX media (for all runs) (Table S10). Five 1536 PCA plates (PCA1-plate1, PCA1-plate2, PCA2, PCA3-plate1 and PCA3-plate2) were generated this way. We tested the interactions between 286 pairs in five to twenty replicates each (Table S1) .

We also used the robotic platform to generate three bait and three prey 1536 arrays for the DHFR PCA assays on plasmids, testing each pairwise interaction at least four times. We mated all the preys and bait strains on YPD media at room temperature for 24 hours. We performed two successive steps of diploid selection (SC -leu -trp -ade, Table S10) followed by two steps in MTX media (SC -leu -trp -ade MTX, Table S10). We incubated the diploid selection plates and the first MTX plates at 30°C for 48 hours.

3.4. Analysis of DHFR PCA assays

3.4.1 Image analysis and colony size quantification

All images were analysed the same way, including images from (Stynen et al., 2018). Images of plates were taken with a EOS Rebel T5i camera (Canon, Tokyo, Japan) every two hours during the entire course of the experiment. Incubation and imaging was performed in a spImager custom platform (S&P Robotics Inc, Toronto, Canada). We kept images after 2 days of growth for diploid selection plates (S2 and S4) and after 4 days of growth for DMSO and MTX plates. Images were analysed using *gitter* (R package version 1.1.1 (Wagih and Parts, 2014) to quantify colony sizes defining a square around the colony center and measuring the foreground pixel intensity minus the background pixel intensity.

3.4.2 Data filtering

For the images from (Stynen et al., 2018), we filtered data based on the diploid selection plates. Colonies smaller than 200 pixels were considered as missing data rather than as non interacting strains. For PCA1, 2 and 3, colonies flagged as irregular by gitter (as S (colony spill or edge interference) or S, C (low colony circularity) flags), that did not grow on the last diploid selection step (S2 or S4) or DMSO media (smaller than quantile 25 minus the interquartile range) were considered as missing data. We considered only bait-prey pairs with at least 4 replicates and used median of colony sizes as PCA signal. The data was finally filtered based on the completeness of paralogous pairs so we could test HMs and HETs systematically. Median colony sizes were $\log_2$ transformed after adding a value of 1 to all data to obtain PCA scores. The result of (Stynen et al., 2018) and PCA1,2 and 3 were strongly correlated, with an overall pearson correlation of 0.578 (p.value < 2.2e-16) (Figure S1. B).

3.4.3 Detection of protein-protein interactions

The distribution of PCA score was modeled per duplication type (SSDs and WGDs) and per interaction tested (HM or HET) as in (Diss et al., 2017) with the *normalmixEM* function (default parameters) available in the R mixtools package (Benaglia et al., 2009). The background signal on MTX was used as a null distribution to which interactions were compared. The size of colonies (PCA scores (PCA_s)) were converted to z-scores using the mean (μ_b) and standard deviation (sd_b) of the background distribution ($Z_s = (PCA_s - \mu_b)/sd_b$). PPI were considered as detected if Z_s of the bait-prey pair was greater than 2.5 (Figure S17) (Chrétien et al., 2018).

We observed 38 cases in which only one of the two possible HET interaction was detected (P1-DHFR F[1,2] x P2-DHFR F[3] or P2-DHFR F[1,2] x P1-DHFR F[3]). It is typical for PCA assays to detect interactions in one orientation or the other (See (Tarassov et al., 2008)). However, this could also be caused by one of the four strains having an abnormal fusion sequence. We verified by PCR and sequenced the fusion sequences to make sure this was not the case. The correct strains were conserved and the other ones were re-constructed and retested. Finally, only 14 cases of unidirectional HET were observed in our final results. For all other 69 cases, both were detected.

3.4.4 Dataset integration

The PCA data was integrated with other data as described above. The overlaps among the different datasets and PCA are shown in Figure S2.

4. Gene expression in MTX conditions

4.1 Cell cultures for RNAseq

We used the diploid strain used for the border controls in the DHFR PCA (MAT**a**/α *LSM8*-DHFR F[1,2]/*LSM8 CDC39*/*CDC39*-DHFR F[3]) to measure expression profile in MTX conditions. Three overnight pre-cultures were grown separately in 5 ml of YPD+NAT+HygB (Table S10) at 30°C with shaking at 250 rpm. A second set of pre-cultures were grown starting from a dilution of OD_{600} = 0.01 in 50 ml in the same condition to OD_{600} of 0.8 to 1. Final cultures were started at OD_{600} = 0.03 in 250 ml in two different synthetic media supplemented with MTX or DMSO (Table S10) at 30°C with shaking at 250 rpm. These cultures were transferred to 5 x 50 ml tubes when they reached OD_{600} of 0.6 to 0.7 and centrifuged at 1008 RCF at 4°C for 1 min. The supernatant was discarded and cell pellets were frozen in liquid nitrogen and stored at -80°C until processing. RNA extraction, library generation and amplification were performed as described in (Eberlein et al., 2019). Briefly, the Quantseq 3' mRNA kit (Lexogen, Vienna, Austria) was used for library preparation (Moll et al., 2014) following the manufacturer's protocol. The PCR cycles number during library amplification was adjusted to 16. The six libraries were pooled and sequenced on a single Ion Torrent (ThermoFisher Scientific, Waltham, United States) chip for a total of 7,784,644 reads on average per library. Barcodes associated to the samples of this study are listed in Table S3.

4.2 RNAseq analysis

Read quality statistics were retrieved from the program FastQC (Andrews, 2010). Reads were cleaned using cutadapt (Martin, 2011). We removed the first 12 bp, trimmed the poly-A tail from the 3' end, trimmed low-quality ends using a cutoff of 15 (phred quality + 33) and discarded reads shorter than 30 bp. The number of reads before and after cleaning can be found in Table S3. Raw sequences can be downloaded under the NCBI BioProject ID PRJNA480398.

Cleaned reads were aligned on the reference genome of S288c from SGD (S288C_reference_genome_R64-2-1_20150113.fsa version) using bwa (Li and Durbin, 2009). Because we used a 3'mRNA-Seq Library, reads mapped largely to 3'UTRs. We increased the window of genes annotated in the SGD annotation (saccharomyces_cerevisiae_R64-2-1_20150113.gff version) using the UTR annotation from (Nagalakshmi et al., 2008). Based on this genes-UTR annotation as reference, the number of mapped reads per genes was estimated using htseq-count of the Python package HTSeq (Anders et al., 2015) and reported in Table S3.

4.3 Correlation of gene expression profiles

The correlation of expression profiles for paralogs was calculated using Pearson's correlation from the large-scale normalized expression data from *S. cerevisiae* (Ihmels et al., 2004) over 1000 mRNA expression profiles from different conditions and different cell cycle phases.

5. Structural analyses

5.1. Sequence conservation in interfaces of yeast complexes

5.1.1. Identification of crystal structures

The reference proteome of *Saccharomyces cerevisiae* assembly R64-1-1 was downloaded on April 16th, 2018 from the Ensembl database at (http://useast.ensembl.org/info/data/ftp/index.html) (Zerbino et al., 2018). The sequences of paralogs classified as SSD or WGD (Byrne and Wolfe, 2005; Guan et al., 2007) were searched using BLASTP (version 2.6.0+) (Camacho et al., 2009) to all the protein chains contained in the Protein Data Bank (PDB) downloaded on September 21st, 2017 (Berman et al., 2000). Due to the high sequence identity of some paralogs (up to 95%), their structures were assigned as protein chains from the PDB that had a 100% sequence identity and an E-value lower than 0.000001.

5.1.2. Identification of interfaces

Residue positions involved in protein interaction interfaces were defined based on the distance of residues to the other chain (Tsai et al., 1996). Contacting residues are defined as those whose two closest non-hydrogen atoms are separated by a distance smaller than the sum of their van der Waals radii plus 0.5 Å. Nearby residues are those whose alpha carbons are located at a distance smaller than 6 Å. All distances were measured using the Biopython library (version 1.70) (Cock et al., 2009).

5.1.3. Sequence conservation within interfaces

The dataset of PDB files was then filtered to include only the crystallographic structures with the highest resolution available for each complex involving direct contacts between subunits of the paralogs. Full-length protein sequences from the reference proteome were then aligned to their matching chains from the PDB with MUSCLE version 3.8.31 (Edgar, 2004) to assign the structural data to the residues in the full-length chain. These full-length chains were then aligned to their paralogs and sequences from PhylomeDB phylogenies (Huerta-Cepas et al., 2008) with MUSCLE version 3.8.31. Sequence identity was calculated within the rim and core regions of the interface. PDB identifiers for structures included in this analysis are shown in Table S11. Pairs of paralogs for which the crystallized domain was only present in one of the proteins were not considered for this analysis.

5.2. Simulations of coevolution of protein complexes

5.2.1 Mutation sampling during evolution of protein interfaces

Simulations were carried out with high quality crystal structures of homodimeric proteins from PDB (Berman et al., 2000). Four of them (PDB: 1M38, 2JKY, 3D8X, 4FGW) were taken from the above dataset of structures that matched yeast paralogs and two others from the same tier of high quality structures (PDB: 1A82, 2O1V). The simulations model the duplication of the gene encoding the homodimer, giving rise to separate copies that can accumulate different mutations, leading to the formation of HMs and HETs as in Figure 1.

Mutations were introduced using a transition matrix whose substitution probabilities consider the codon code and allow only substitutions that would require a single base change in the underlying codons (Thorvaldsen, 2016). Due to the degenerate nature of the genetic code, the model also allows synonymous mutations. Thus, the model explores the effects of mutations in both chains, as well as mutations in only one chain. The framework assumes equal mutation rates at both loci, as it proposes a mutation at each locus after every step in the simulation, with 50 replicates of 200 steps of substitution in each simulation. Restricting the mutations to the interface maintains sequence identity above 40%, which has been described previously as the threshold at which protein fold remains similar (Addou et al., 2009; Todd et al., 2001; Wilson et al., 2000).

5.2.2 Implementation of selection

Simulations were carried out using the FoldX suite version 4 (Guerois et al., 2002; Schymkowitz et al., 2005). Starting structures were repaired with the RepairPDB function, mutations were simulated with BuildModel followed by the Optimize function, and estimations of protein stability and binding energy of the complex were done with the Stability and AnalyseComplex functions, respectively. Effects of mutations on complex fitness were calculated using methods previously described (Kachroo et al., 2015). The fitness of a complex was calculated from three components based on the stability of protein chains and the binding energy of the complex using equation 1:

$$x_i^k = -\log\left[e^{\beta\left(\Delta G_i^k - \Delta G_{threshold}^k\right)} + 1\right] \qquad (1)$$

where *i* is the index of the current substitution, *k* is the index of one of the model's three energetic parameters (stability of chain A, stability of chain B, or binding energy of the complex), x_i^k is the fitness component of the *k*th parameter for the *i*th substitution, β is a parameter that determines smoothness of the fitness curve, ΔG_i^k is the free energy value of the *k*th free energy parameter (stability of chain A, stability of chain B, or binding energy of the complex) for the *i*th substitution, and $\Delta G_{threshold}^k$ is a threshold around which the *k*th fitness component starts to decrease. The total fitness of the complex after the *i*th mutation was calculated as the sum of the three computed values for x_i^k, as shown in equation 2:

$$x_i = \sum_{k=1}^{3} x_i^k \qquad (2)$$

The fitness values of complexes were then used to calculate the probability of fixation ($p_{fixation}$) of the substitutions using the Metropolis criterion, as in equation 3:

$$p_{fixation} = \begin{cases} 1, & x_j > x_i \\ e^{-2N(x_i - x_j)}, & x_j \leq x_i \end{cases} \qquad (3)$$

where $p_{fixation}$ is the probability of fixation, x_j is the total fitness value for the complex after *j* substitutions; x_i is the total fitness value for the complex after *i* substitutions, with *j* = *i* + *1*; and N is the population size, which influences the efficiency of selection.

Different selection scenarios were examined depending on the complexes whose binding energy and chain stabilities were under selection: neutral evolution (no selection applied on chain stability and the binding energy of the complex), selection on one homodimer, selection on the two homodimers, and selection on the heterodimer. β was set to 10, N was set to 1000 and the $\Delta G^{k}_{threshold}$ were set to 99.9% of the starting values for each complex, following the parameters described in (Kachroo et al., 2015). For the simulations with neutral evolution, β was set to 1.

5.2.3 Analyses of simulations

The results from the simulations were then analyzed by distinguishing mutational steps with only one non-synonymous mutation (single mutants, between 29% and 34% of the steps in the simulations) from steps with two non-synonymous mutations (double mutants, between 61% and 68% of the steps). The global data were used to follow the evolution of binding energies of the complexes over time, which is shown in Figure 4. The effects of mutations in HM and HET were compared using the single mutants (Figure 5). The double mutants were used to compare the rates of mutation fixation based on their effects on the HMs (Figure S10).

Author contributions

CRL, AM and AFC designed this study. AM, AKD, IGA, DA, SA, CE and DEY performed the experiments. AFC performed the *in silico* evolution experiments and the analysis of protein structures. AM, AFC, HAJ and CRL analysed the results. CRL and NY supervised the research. AM, AFC and CRL wrote the manuscript with input from all authors.

Acknowledgements

This work was supported by Canadian Institutes of Health Research grants 299432, 324265 and 387697 to CRL AM was supported by a FRQS postdoctoral scholarship. AFC was supported by fellowships from PROTEO, MITACS, and Université Laval, as well as joint funding from MEES and AMEXCID. SA was supported by a NSERC undergraduate scholarship. CRL holds the Canada Research Chair in Evolutionary Cells and Systems Biology. We thank SW Michnick for sharing data before publication. The authors thank Philippe Després, Rohan Dandage, Johan Hallin and Anna Fijarczyk for comments on the paper, Rong Shi for useful discussions, and Stéphane Larose for assistance on data management.

Competing interests

The authors have no competing interests to declare.

References

Addou S, Rentzsch R, Lee D, Orengo CA. 2009. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. *J Mol Biol* **387**:416–430. doi:10.1016/j.jmb.2008.12.045

Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG. 2008. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci* **33**:220–229. doi:10.1016/j.tibs.2008.02.002

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**:166–169. doi:10.1093/bioinformatics/btu638

André I, Strauss CEM, Kaplan DB, Bradley P, Baker D. 2008. Emergence of symmetry in homooligomeric biological assemblies. *Proc Natl Acad Sci U S A* **105**:16148–16152. doi:10.1073/pnas.0807576105

Andrews S. 2010. FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashenberg O, Rozen-Gagnon K, Laub MT, Keating AE. 2011. Determinants of homodimerization specificity in histidine kinases. *J Mol Biol* **413**:222–235. doi:10.1016/j.jmb.2011.08.011

Baker CR, Hanson-Smith V, Johnson AD. 2013. Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science* **342**:104–108. doi:10.1126/science.1240810

Barshir R, Hekselman I, Shemesh N, Sharon M, Novack L, Yeger-Lotem E. 2018. Role of duplicate genes in determining the tissue-selectivity of hereditary diseases. *PLoS Genet* **14**:e1007327. doi:10.1371/journal.pgen.1007327

Benaglia T, Chauveau D, Hunter D, Young D. 2009. mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software, Articles* **32**:1–29. doi:10.18637/jss.v032.i06

Benschop JJ, Brabers N, van Leenen D, Bakker LV, van Deutekom HWM, van Berkum NL, Apweiler E, Lijnzaad P, Holstege FCP, Kemmeren P. 2010. A consensus of core protein complex compositions for Saccharomyces cerevisiae. *Mol Cell* **38**:916–928. doi:10.1016/j.molcel.2010.06.002

Bergendahl LT, Marsh JA. 2017. Functional determinants of protein assembly into homomeric complexes. *Sci Rep* **7**:4932. doi:10.1038/s41598-017-05084-8

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**:235–242. doi:10.1093/nar/28.1.235

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A* **109**:14746–14753. doi:10.1073/pnas.1207726109

Boncoeur E, Durmort C, Bernay B, Ebel C, Di Guilmi AM, Croizé J, Vernet T, Jault J-M. 2012. PatA and PatB form a functional heterodimeric ABC multidrug efflux transporter responsible for the resistance of Streptococcus pneumoniae to fluoroquinolones. *Biochemistry* **51**:7755–7765. doi:10.1021/bi300762p

Brender JR, Zhang Y. 2015. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLoS Comput Biol* **11**:e1004494. doi:10.1371/journal.pcbi.1004494

Bridgham JT, Brown JE, Rodríguez-Marí A, Catchen JM, Thornton JW. 2008. Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors. *PLoS Genet* **4**. doi:10.1371/journal.pgen.1000191

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**:1456–1461. doi:10.1101/gr.3672305

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421. doi:10.1186/1471-2105-10-421

Celaj A, Schlecht U, Smith JD, Xu W, Suresh S, Miranda M, Aparicio AM, Proctor M, Davis RW, Roth FP, St Onge RP. 2017. Quantitative analysis of protein interaction network dynamics in yeast. *Mol Syst Biol* **13**:934. doi:10.15252/msb.20177532

Chatr-Aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M. 2013. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* **41**:D816–D823. doi:10.1093/nar/gks1158

Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz B-J, Dolinski K, Tyers M. 2017. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**:D369–D379. doi:10.1093/nar/gkw1102

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* **40**:D700–5. doi:10.1093/nar/gkr1029

Chrétien A-È, Gagnon-Arsenault I, Dubé AK, Barbeau X, Després PC, Lamothe C, Dion-Côté A-M, Lagüe P, Landry CR. 2018. Extended Linkers Improve the Detection of Protein-protein Interactions (PPIs) by Dihydrofolate Reductase Protein-fragment Complementation Assay (DHFR PCA) in Living Cells. *Mol Cell Proteomics* **17**:373–383. doi:10.1074/mcp.TIR117.000385

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422–1423. doi:10.1093/bioinformatics/btp163

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, Pelechano V, Styles EB, Billmann M, van Leeuwen J, van Dyk N, Lin Z-Y, Kuzmin E, Nelson J, Piotrowski JS, Srikumar T, Bahr S, Chen Y, Deshpande R, Kurat CF, Li SC, Li Z, Usaj MM, Okada H, Pascoe N, San Luis B-J, Sharifpoor S, Shuteriqi E, Simpkins SW, Snider J, Suresh HG, Tan Y, Zhu H, Malod-Dognin N, Janjic V, Przulj N, Troyanskaya OG, Stagljar I, Xia T, Ohya Y, Gingras A-C, Raught B, Boutros M, Steinmetz LM, Moore CL, Rosebrock AP, Caudy AA, Myers CL, Andrews B, Boone C. 2016. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**. doi:10.1126/science.aaf1420

DeLuna A, Springer M, Kirschner MW, Kishony R. 2010. Need-Based Up-Regulation of Protein Levels in Response to Deletion of Their Duplicate Genes. *PLoS Biol* **8**:e1000347. doi:10.1371/journal.pbio.1000347

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A* **110**:2898–2903. doi:10.1073/pnas.1300127110

Dey S, Ritchie DW, Levy ED. 2018. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat Methods* **15**:67–72. doi:10.1038/nmeth.4510

Diss G, Gagnon-Arsenault I, Dion-Coté A-M, Vignaud H, Ascencio D, Berger CM, Landry CR. 2017. Gene duplication can impart fragility, not robustness in the yeast protein interaction network. *Science* **355**:630–634. doi:10.1126/science.aai7685

Dong D, Yuan Z, Zhang Z. 2011. Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic Acids Res* **39**:837–847. doi:10.1093/nar/gkq874

Eberlein C, Hénault M, Fijarczyk A, Charron G, Bouvier M, Kohn LM, Anderson JB, Landry CR. 2019. Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat Commun* **10**:923. doi:10.1038/s41467-019-08809-7

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

Freschi L, Torres-Quiroz F, Dubé AK, Landry CR. 2013. qPCA: a scalable assay to measure the perturbation of protein-protein interactions in living cells. *Mol Biosyst* **9**:36–43. doi:10.1039/c2mb25265a

Gagnon-Arsenault I, Marois Blanchet F-C, Rochette S, Diss G, Dubé AK, Landry CR. 2013. Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. *J Proteomics*, Special Issue: From protein structures to clinical applications **81**:112–125. doi:10.1016/j.jprot.2012.09.038

Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA 3rd, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**:343–345. doi:10.1038/nmeth.1318

Gout J-F, Kahn D, Duret L, Paramecium Post-Genomics Consortium. 2010. Correction: The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLoS Genet* **6**:10.1371. doi:10.1371/annotation/c55d5089-ba2f-449d-8696-2bc8395978db

Gout J-F, Lynch M. 2015. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. *Mol Biol Evol* **32**:2141–2148. doi:10.1093/molbev/msv095

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional Analysis of Gene Duplications in Saccharomyces cerevisiae. *Genetics* **175**:933–943. doi:10.1534/genetics.106.064329

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**:369–387. doi:10.1016/S0022-2836(02)00442-4

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol* **8**:R209. doi:10.1186/gb-2007-8-10-r209

Hochberg GKA, Shepherd DA, Marklund EG, Santhanagoplan I, Degiacomi MT, Laganowsky A, Allison TM, Basha E, Marty MT, Galpin MR, Struwe WB, Baldwin AJ, Vierling E, Benesch JLP. 2018. Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. *Science* **359**:930–935. doi:10.1126/science.aam7229

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res* **36**:D491–6. doi:10.1093/nar/gkm899

Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. *Nature* **425**:686–691. doi:10.1038/nature02026

Ihmels J, Bergmann S, Barkai N. 2004. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**:1993–2003. doi:10.1093/bioinformatics/bth166

Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res* **33**:3629–3635. doi:10.1093/nar/gki678

Janin J, Bahadur RP, Chakrabarti P. 2008. Protein–protein interaction and quaternary structure. *Q Rev Biophys* **41**:133–180. doi:10.1017/S0033583508004708

Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. 2015. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**:921–925. doi:10.1126/science.aaa0769

Kaltenegger E, Ober D. 2015. Paralogue Interference Affects the Dynamics after Gene Duplication. *Trends Plant Sci* **20**:814–821. doi:10.1016/j.tplants.2015.10.003

Kim Y, Jung JP, Pack C-G, Huh W-K. 2019. Global analysis of protein homomerization in Saccharomyces cerevisiae. *Genome Res* **29**:135–145. doi:10.1101/gr.231860.117

Landry CR, Levy ED, Abd Rabbo D, Tarassov K, Michnick SW. 2013. Extracting insight from noisy cellular networks. *Cell* **155**:983–989. doi:10.1016/j.cell.2013.11.003

Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A* **109**:20461–20466. doi:10.1073/pnas.1209312109

Levy ED, Teichmann SA. 2013. Chapter Two - Structural, Evolutionary, and Assembly Principles of Protein Oligomerization In: Giraldo J, Ciruela F, editors. Progress in Molecular Biology and Translational Science. Academic Press. pp. 25–51. doi:10.1016/B978-0-12-386931-9.00002-7

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. doi:10.1093/bioinformatics/btp324

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**:2178–2189. doi:10.1101/gr.1224503

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet* **21**:602–607. doi:10.1016/j.tig.2005.08.006

Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI. 2007. Structural similarity enhances interaction propensity of proteins. *J Mol Biol* **365**:1596–1606. doi:10.1016/j.jmb.2006.11.020

Lynch M. 2012. The evolution of multimeric protein assemblages. *Mol Biol Evol* **29**:1353–1366. doi:10.1093/molbev/msr300

Marcet-Houben M, Gabaldón T. 2015. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker’s Yeast Lineage. *PLoS Biol* **13**:e1002220. doi:10.1371/journal.pbio.1002220

Marsh JA, Teichmann SA. 2015. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem* **84**:551–575. doi:10.1146/annurev-biochem-060614-034142

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**:10–12. doi:10.14806/ej.17.1.200

Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN, Ricard-Blum S, Roechert B, Skyzypek MS, Tiwari M, Velankar S, Wong ED, Hermjakob H, Orchard S. 2015. The complex portal--an encyclopaedia of macromolecular complexes. *Nucleic Acids Res* **43**:D479–84. doi:10.1093/nar/gku975

Michnick SW, Levy ED, Landry CR, Kowarzyk J, Messier V. 2016. The Dihydrofolate Reductase Protein-Fragment Complementation Assay: A Survival-Selection Assay for Large-Scale Analysis of Protein-Protein Interactions. *Cold Spring Harb Protoc* **2016**. doi:10.1101/pdb.prot090027

Moll P, Ante M, Seitz A, Reda T. 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat Methods* **11**:972. doi:10.1038/nmeth.f.376

Monod J, Wyman J, Changeux JP. 1965. On the nature of allosteric transitions: a plausible model. *J Mol Biol* **12**:88–118. doi:10.1016/S0022-2836(65)80285-6

Musso G, Zhang Z, Emili A. 2007. Retention of protein complex membership by ancient duplicated gene products in budding yeast. *Trends Genet* **23**:266–269. doi:10.1016/j.tig.2007.03.012

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**:1344–1349. doi:10.1126/science.1158441

Nasir A, Kim KM, Caetano-Anollés G. 2014. Global patterns of protein domain gain and loss in superkingdoms. *PLoS Comput Biol* **10**:e1003452. doi:10.1371/journal.pcbi.1003452

Natan E, Endoh T, Haim-Vilmovsky L, Flock T, Chalancon G, Hopper JTS, Kintses B, Horvath P, Daruka L, Fekete G, Pál C, Papp B, Oszi E, Magyar Z, Marsh JA, Elcock AH, Babu MM, Robinson CV, Sugimoto N, Teichmann SA. 2018. Cotranslational protein assembly imposes evolutionary constraints on homomeric proteins. *Nat Struct Mol Biol* **25**:279–288. doi:10.1038/s41594-018-0029-5

Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2018. Biostrings: Efficient manipulation of biological strings version 2.50.1 from Bioconductor. https://rdrr.io/bioc/Biostrings/

Pandey AV, Henderson CJ, Ishii Y, Kranendonk M, Backes WL, Zanger UM. 2017. Editorial: Role of Protein-Protein Interactions in Metabolism: Genetics, Structure, Function. *Front Pharmacol* **8**:881. doi:10.3389/fphar.2017.00881

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194–197. doi:10.1038/nature01771

Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* **8**:R51. doi:10.1186/gb-2007-8-4-r51

Pérez-Bercoff A, Makino T, McLysaght A. 2010. Duplicability of self-interacting human genes. *BMC Evol Biol* **10**:160. doi:10.1186/1471-2148-10-160

Presser A, Elowitz MB, Kellis M, Kishony R. 2008. The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proc Natl Acad Sci U S A* **105**:950–954. doi:10.1073/pnas.0707293105

Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ. 2007. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **7**:944–960. doi:10.1002/pmic.200600636

Pu S, Wong J, Turner B, Cho E, Wodak SJ. 2009. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* **37**:825–831. doi:10.1093/nar/gkn1005

Rice AM, McLysaght A. 2017. Dosage-sensitive genes in evolution and disease. *BMC Biol* **15**:78. doi:10.1186/s12915-017-0418-y

Rochette S, Diss G, Filteau M, Leducq J-B, Dubé AK, Landry CR. 2015. Genome-wide Protein-protein Interaction Screening by Protein-fragment Complementation Assay (PCA) in Living Cells. *J Vis Exp*. doi:10.3791/52255

Rochette S, Gagnon-Arsenault I, Diss G, Landry CR. 2014. Modulation of the yeast protein interactome in response to DNA damage. *J Proteomics*, Special Issue: Can Proteomics Fill the Gap Between Genomics and Phenotypes? **100**:25–36. doi:10.1016/j.jprot.2013.11.007

Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. 2005.

Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**:1173–1178. doi:10.1038/nature04209

Schrödinger LLC. 2015. The PyMOL Molecular Graphics System, Version 1.8.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res* **33**:W382–8. doi:10.1093/nar/gki387

Scott JD, Pawson T. 2009. Cell Signaling in Space and Time: Where Proteins Come Together and When They're Apart. *Science* **326**:1220–1224. doi:10.1126/science.1175668

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**:D535–9. doi:10.1093/nar/gkj109

Stynen B, Abd-Rabbo D, Kowarzyk J, Miller-Fleming L, Aulakh SK, Garneau P, Ralser M, Michnick SW. 2018. Changes of Cell Biochemical States Are Revealed in Protein Homomeric Complex Dynamics. *Cell*. doi:10.1016/j.cell.2018.09.050

Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. 2008. An in Vivo Map of the Yeast Protein Interactome. *Science* **320**:1465–1470. doi:10.1126/science.1153878

Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sá-Correia I. 2006. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res* **34**:D446–51. doi:10.1093/nar/gkj013

Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, Cavalheiro M, Antunes M, Lemos A, Pedreira T, Sá-Correia I. 2018. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Res* **46**:D348–D353. doi:10.1093/nar/gkx842

Thompson A, Zakon HH, Kirkpatrick M. 2016. Compensatory Drift and the Evolutionary Dynamics of Dosage-Sensitive Duplicate Genes. *Genetics* **202**:765–774. doi:10.1534/genetics.115.178137

Thompson DA, Roy S, Chan M, Styczynsky MP, Pfiffner J, French C, Socha A, Thielke A, Napolitano S, Muller P, Kellis M, Konieczka JH, Wapinski I, Regev A. 2013. Evolutionary principles of modular gene regulation in yeasts. *Elife* **2**:e00603. doi:10.7554/eLife.00603

Thorvaldsen S. 2016. A Mutation Model from First Principles of the Genetic Code. *IEEE/ACM Trans Comput Biol Bioinform* **13**:878–886. doi:10.1109/TCBB.2015.2489641

Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**:1113–1143. doi:10.1006/jmbi.2001.4513

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. 1996. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit Rev Biochem Mol Biol* **31**:127–152. doi:10.3109/10409239609106582

Usaj M, Tan Y, Wang W, VanderSluis B, Zou A, Myers CL, Costanzo M, Andrews B, Boone C. 2017. TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. *G3* **7**:1539–1549. doi:10.1534/g3.117.040220

Vidal M, Cusick ME, Barabási A-L. 2011. Interactome Networks and Human Disease. *Cell* **144**:986–998. doi:10.1016/j.cell.2011.02.016

Wagih O, Parts L. 2014. gitter: A Robust and Accurate Method for Quantification of Colony Sizes From Plate Images. *G3* **4**:547–552. doi:10.1534/g3.113.009431

Wagner A. 2003. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London B: Biological Sciences* **270**:457–466. doi:10.1098/rspb.2002.2269

Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, Chessman K, Pal S, Cromar G, Papoulas O, Ni Z, Boutz DR, Stoilova S, Havugimana PC, Guo X, Malty RH, Sarov M, Greenblatt J, Babu M, Derry WB, Tillier ER, Wallingford JB, Parkinson J, Marcotte EM, Emili A. 2015. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**:339–344. doi:10.1038/nature14877

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* **11**:492–500. doi:10.1074/mcp.O111.014704

Wilson CA, Kreychman J, Gerstein M. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**:233–249. doi:10.1006/jmbi.2000.3550

Wolfe KH. 2015. Origin of the Yeast Whole-Genome Duplication. *PLoS Biol* **13**:e1002221. doi:10.1371/journal.pbio.1002221

Yachie N, Petsalaki E, Mellor JC, Weile J, Jacob Y, Verby M, Ozturk SB, Li S, Cote AG, Mosca R, Knapp JJ, Ko M, Yu A, Gebbia M, Sahni N, Yi S, Tyagi T, Sheykhkarimli D, Roth JF, Wong C, Musa L, Snider J, Liu Y-C, Yu H, Braun P, Stagljar I, Hao T, Calderwood MA, Pelletier L, Aloy P, Hill DE, Vidal M, Roth FP. 2016. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol Syst Biol* **12**:863. doi:10.15252/msb.20156660

Yang J, Lusk R, Li W-H. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A* **100**:15661–15665. doi:10.1073/pnas.2536672100

Yu G. 2019. A Tidy Tool for Phylogenetic Tree Data Manipulation [R package tidytree version 0.1.9]. *Comprehensive R Archive Network*.

Yu G, Lam TT-Y, Zhu H, Guan Y. 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**:3041–3043. doi:10.1093/molbev/msy194

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**:28–36. doi:10.1111/2041-210X.12628

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2018. Ensembl 2018. *Nucleic Acids Res* **46**:D754–D761. doi:10.1093/nar/gkx1098