DNA Research, 2017, 24(4), 435–444
doi: 10.1093/dnares/dsx015
Advance Access Publication Date: 19 April 2017
Full Paper

Full Paper

Intrinsic protein disorder reduces small-scale gene duplicability

Sanghita Banerjee[1,2], Felix Feyertag[1], and David Alvarez-Ponce[1,*]

[1]Department of Biology, University of Nevada, Reno, NV 89557, USA, and [2]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

*To whom correspondence should be addressed. Tel: +1 (775) 682-5735. Fax: +1 (775) 784-1302. Email: dap@unr.edu

Edited by Dr. Osamu Ohara

Received 8 December 2016; Editorial decision 22 March 2017; Accepted 28 March 2017

Abstract

Whereas the rate of gene duplication is relatively high, only certain duplications survive the filter of natural selection and can contribute to genome evolution. However, the reasons why certain genes can be retained after duplication whereas others cannot remain largely unknown. Many proteins contain intrinsically disordered regions (IDRs), whose structures fluctuate between alternative conformational states. Due to their high flexibility, IDRs often enable protein–protein interactions and are the target of post-translational modifications. Intrinsically disordered proteins (IDPs) have characteristics that might either stimulate or hamper the retention of their encoding genes after duplication. On the one hand, IDRs may enable functional diversification, thus promoting duplicate retention. On the other hand, increased IDP availability is expected to result in deleterious unspecific interactions. Here, we interrogate the proteomes of human, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Escherichia coli*, in order to ascertain the impact of protein intrinsic disorder on gene duplicability. We show that, in general, proteins encoded by duplicated genes tend to be less disordered than those encoded by singletons. The only exception is proteins encoded by ohnologs, which tend to be more disordered than those encoded by singletons or genes resulting from small-scale duplications. Our results indicate that duplication of genes encoding IDPs outside the context of whole-genome duplication (WGD) is often deleterious, but that IDRs facilitate retention of duplicates in the context of WGD. We discuss the potential evolutionary implications of our results.

Key words: protein folding, unstructured proteins, singleton, ohnologs, whole genome duplications

1. Introduction

Gene duplication is thought to be a major force driving evolutionary innovations.[1–3] Even though gene duplications occur frequently, they are often transient, and only a fraction of duplications result in fixation of the two gene copies in the population. Genes widely differ in their propensity to be retained after gene duplication (i.e. their duplicability): whereas some genes successfully duplicate very often, giving rise to large multigene families, others remain as singletons during long evolutionary periods. What factors affect gene duplicability is still a largely open question in Evolutionary Biology.[4]

Many proteins contain intrinsically disordered regions (IDRs). These regions lack a stable tertiary or secondary structure under normal physiological conditions, having a structure that constantly oscillates between alternative conformational states.[5–8] Due to their flexibility and to their enrichment in short interaction motifs,[9] IDRs are often involved in interactions with other proteins.[10,11]

In addition, intrinsically disordered proteins (IDPs, i.e. proteins with a predominance of IDRs) often have larger interaction surfaces than other proteins of similar length.[12] As a result, IDPs tend to be promiscuous in their interaction patterns, being involved in a high number of protein–protein interactions.[10,11,13] Proteins involved in signalling, including transcription factors, tend to be rich in IDRs.[14]

IDPs exhibit characteristics that might either increase or reduce the duplicability of their encoding genes. On the one hand, given their high flexibility and enrichment in interaction motifs, duplication of genes encoding IDPs is expected to result in an increased number of misinteractions—unspecific, ectopic interactions with proteins with which the protein is not supposed to interact—resulting in unwanted activation of cellular processes, interference with functional interactions and sequestration of functional proteins into non-functional complexes.[15,16] Indeed, IDP availability is often maintained at low levels, and several lines of evidence indicate that this availability is tightly regulated[15–20] and that dysregulation of IDPs often leads to disease, including neurodegeneration and cancer[14,15,21] among other deleterious effects. Remarkably, Vavouri et al.[15] found that IDPs tend to be dosage-sensitive proteins—proteins whose over-expression reduces fitness.

On the other hand, the high flexibility of IDRs may facilitate functional divergence (subfunctionalization or neofunctionalization) of gene copies after duplication, which promotes retention of the duplicates. In addition, IDRs are enriched in post-translational modification sites,[22,23] which also contribute to functional divergence of gene duplicates.[24] Consistent with this model, Montanari et al.[25] showed that yeast ohnologs—duplicates that were retained after the whole genome duplication or interspecific hybridization event that took place in an ancestor of *Saccharomyces cerevisiae*[26–28]—tend to encode proteins with a high number of IDRs. It should be noted, however, that whole genome duplication (WGD) and hybridization events maintain the stoichiometry of all molecular interactions in the cellular system,[29,30] and often result in an increased cell volume, which means that protein concentrations are not necessarily altered.[31–34] This is not the case for small-scale duplicates (SSDs), which are expected to increase abundance of the encoded proteins and to upset the balance of the interactions in which these proteins are involved.[29,35–37] Therefore, the selective pressures constraining ohnologs retention are expected to be different from those acting on other kinds of duplicates.[38,39]

Here, we interrogate the proteomes of six organisms to study the effect of protein intrinsic disorder on gene duplicability. While ohnologs tend to encode highly disordered proteins, SSDs tend to encode lowly disordered proteins. The trend is independent of covariation of disorder and duplicability with gene expression levels, protein abundances and number of protein–protein interactions. In addition, orthologs of genes that specifically duplicated in the studied species tend to encode lowly disordered proteins. Our analyses indicate that genes encoding IDPs are unlikely to undergo successful small-scale duplication, suggesting that small-scale duplication of such genes often has deleterious effects.

2. Material and methods

2.1. Quantification of protein intrinsic disorder

We retrieved the proteomes of the six studied species (human, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana* and *E. coli*) from the databases Ensembl (release 85), Ensembl Plants (release 31) and Ensembl Bacteria (release 33). For each protein-coding gene, we chose the longest encoded protein for analysis (in the event of multiple splicing variants). We used IUPred[40] to identify the disordered residues within each protein sequence. We used the IUPred-L option (long intrinsic disorder); using this option, disordered regions must encompass at least 30 consecutive amino acids predicted to be disordered. Shorter predicted disordered regions were excluded from our calculations. This software assigns to each amino acid residue a value between 0 and 1, depending on its propensity to being intrinsically disordered. We considered an amino acid residue as intrinsically disordered if the score was ≥ 0.5, a cut-off that is widely used for optimal prediction of disordered residues (e.g. Refs. [9,40,41]). For each protein, we computed the percentage of disordered residues. Additionally, we validated our main results using FoldIndex.[42]

We classified proteins based on their disorder content, as either IDPs (percentage of disordered residues $\geq 30\%$), moderately disordered proteins (MDPs, $10\% <$ percentage of disordered residues $< 30\%$) or well-structured proteins (WSPs, percentage of disordered residues $\leq 10\%$). These cut-offs are the most commonly used (see, for instance, Refs. [17,43]). Nonetheless, in order to ensure the robustness of our results to the cut-offs chosen, we repeated our analyses considering an alternative classification: IDPs (percentage of disordered residues $\geq 60\%$), MDPs ($15\% <$ percentage of disordered residues $< 60\%$) and WSPs (percentage of disordered residues $\leq 15\%$).

2.1. Identification of duplicated genes

All genes were classified as singleton or duplicates. For *A. thaliana*, we obtained duplicates information from Ensembl Plants,[44] and for other eukaryotic genomes, we used the annotations available from the Ensembl database,[45] whereas for *E. coli* we generated our own annotations using similarity searches. For each eukaryotic gene, a list of paralogs (duplicates) in the same genome was obtained from Ensembl Biomart.[45] Genes with one or more annotated paralogs were deemed duplicated. Each *E. coli* protein was used as query in a BLASTP search[46] against the *E. coli* proteome. Genes with proteins producing at least one significant hit other than the query sequence (*E*-value $\leq 10^{-5}$, coverage of the query sequence $\geq 80\%$) were considered duplicated genes.

Three of the organisms included in our analyses (human, *S. cerevisiae* and *A. thaliana*) have undergone WGD events. We obtained a list of human ohnologs from the Ohnologs database,[47] a list of *S. cerevisiae* ohnologs from Gordon et al.[48] and a list of *A. thaliana* ohnologs from Blanc et al.[49] *A. thaliana* ohnologs were classified as resulting from each of the three WGD events known to have affected the *A. thaliana* lineage using the classification of Blanc et al.[49] All genes classified as duplicated but not as ohnologs were considered to be resulting from small-scale duplication.

2.2. Gene expression and protein abundance datasets

We obtained human gene expression data for 32 different tissues/organs, measured by RNA sequencing experiments, from the Human Protein Atlas.[50] For each gene, we averaged the expression level values across all 32 tissues and used the mean values in further analyses. For *D. melanogaster* and *C. elegans*, we obtained the mRNA abundance data for the whole adult body from FlyAtlas[51] and modENCODE (data from the EBI Expression Atlas, accession number E-MTAB-2812[52]), respectively. *S. cerevisiae* gene expression data were obtained from Nagalakshmi et al.[53] In the case of *A. thaliana*, we obtained gene expression datasets corresponding to 79 tissues and conditions, from Schmid et al.[54] and processed them as in

Alvarez-Ponce and Fares.[55] For each gene, the median across the 79 datasets was used. For genes matching multiple probe sets, the one resulting in a highest median was kept. Probes matching multiple genes were removed from the analysis. *E. coli* expression data were obtained from Covert et al.[56] For each gene, mRNA expression levels were averaged across three biological replicates.

In an additional analysis, for all organisms for which tissue-specific gene expression data are available (human, *D. melanogaster* and *A. thaliana*), we computed gene expression as the average across all tissues in which gene expression was detected, rather than all tissues. In human, a gene was considered to be expressed at a certain tissue if FPKM ≥ 1. In *D. melanogaster*, a gene was considered to be expressed at a certain tissue if it was detectable in at least three of the four biological replicates. In *A. thaliana*, a gene was considered to be expressed at a certain tissue if it was annotated as 'present' in at least two of the three biological replicates.

For all species, protein abundance data were obtained from the PaxDb database, version 4.0.[57] We used the whole-organism integrated datasets, which is the result of a weighted combination of the results of numerous proteomics studies.

2.3. Number of protein–protein interactions

The protein–protein interaction networks of all eukaryotic species considered in this study were obtained from the BioGRID database, version 3.4.133.[58] Only physical interactions among proteins from the same organism were considered. The *E. coli* protein–protein interaction network was obtained from Hu et al.[59] For each protein, degree was computed as the number of different proteins with which it physically interacts.

2.4. Gene orthology

Human–chicken, *D. melanogaster*–*D. grimshawi*, *C. elegans*–*C. japonica* and *A. thaliana*–*A. lyrata* orthology relationships were obtained from Ensembl Biomart.[45] *S. cerevisiae*–*C. glabrata* and *E. coli*–*M. tuberculosis* orthologies were obtained from the OrthoMCL database.[60]

3. Results

3.1. Proteins encoded by small-scale duplicated genes are less intrinsically disordered than proteins encoded by singletons

We first considered whether proteins encoded by duplicated genes differed from proteins encoded by singleton (non-duplicated) genes in terms of intrinsic disorder. For that purpose, we studied the proteomes of a wide range of organisms, including three animals (human, *D. melanogaster*, *C. elegans*), the fungus *S. cerevisiae*, the plant *A. thaliana* and the bacterium *E. coli*. For each gene, we chose the longest encoded protein for analysis, and we inferred the percent of disordered residues using IUPred.[40] In five of the six species, the disorder content of the proteins encoded by singleton genes was significantly higher than that for those encoded by duplicated ones (Fig. 1; Supplementary Table S1). For instance, in *D. melanogaster*, proteins encoded by duplicated genes exhibit a median intrinsic disorder of 7%, and proteins encoded by singleton genes exhibit a median intrinsic disorder of 18% (Mann–Whitney *U* test, $P = 1.39 \times 10^{-105}$; Fig. 1; Supplementary Table S1). The only exception was *S. cerevisiae*, where the trend was reversed: proteins encoded by duplicated genes were significantly more disordered than proteins encoded by singletons ($P = 1.51 \times 10^{-16}$; Fig. 1; Supplementary Table S1).

Montanari et al.[25] showed that in *S. cerevisiae* proteins encoded by ohnologs were considerably more disordered than those encoded by singletons. Given the potential that this trend could be affecting our observations, we decided to study separately ohnologs and duplicates resulting from small-scale duplications, in all the studied organisms known to have undergone WGD events: human,[61] *S. cerevisiae*[62] and *A. thaliana*.[63] We observed that, in all three organisms, proteins encoded by SSDs represented the least disordered class, and that, in agreement with Montanari et al.,[25] proteins encoded by ohnologs were the most disordered ones. Proteins encoded by singletons displayed an intermediate degree of disorder (Fig. 2; Supplementary Table S2). In human and *A. thaliana*, removing ohnologs from our analyses accentuated the differences between singleton and duplicated genes (Fig. 2; Supplementary Table S2). In *S. cerevisiae*, proteins encoded by singleton genes are on average more disordered than those encoded by SSDs, but the differences are not statistically significant ($P = 0.460$; Fig. 2, Supplementary Table S2).

The observation that ohnologs encode highly disordered proteins is particularly pronounced in *S. cerevisiae* (median disorder: 31.86%). In spite of the fact that ohnologs represent only ~26.6% of yeast duplicates (Supplementary Table S2), the very high disorder content of their encoded proteins results in proteins encoded by duplicated genes being on average more disordered than those encoded by singletons (Fig. 1). This does not occur in humans or *A. thaliana* (Fig. 1), in spite of the fact that ohnologs represent a similar fraction of duplicates in these species (24.0% and 26.1%, respectively; Supplementary Table S2), as in these species proteins encoded by ohnologs are not so markedly disordered (Fig. 2).

Three WGD events have been inferred in the lineage leading to *A. thaliana*.[64] We found that the degree of disorder was higher for proteins encoded by the ohnologs resulting from the most recent event than for those encoded by the ohnologs resulting from the oldest event (median disorder for the most recent class: 11.99%, median disorder for the oldest class: 9.12%; Mann–Whitney *U* test, $P = 0.003$). Proteins encoded by ohnologs originated in the other event exhibited an intermediate degree of disorder (median: 9.72%), but no significant differences were detected with the other two classes (Mann–Whitney *U* test, $P > 0.05$). Proteins encoded by all three *A. thaliana* ohnologs classes exhibited a median disorder that was higher than that for proteins encoded by singleton genes (8.06%; Table 2); however, differences were statistically significant only for the most recent class of ohnologs ($P = 1.79 \times 10^{-22}$).

In order to evaluate the robustness of our results to the method of prediction of intrinsic disorder used, we repeated our analyses using an alternative prediction tool, FoldIndex,[42] with similar results (Supplementary Table S3). Indeed, we observed a very strong correlation between the predictions of IUPred and those of FoldIndex (Supplementary Table S4). Of note, using FoldIndex we observed significant differences between SSDs and singletons in *S. cerevisiae* ($P = 3.70 \times 10^{-07}$; Supplementary Table S3).

3.2. Intrinsically disordered proteins are enriched in proteins encoded by singleton genes

We next classified proteins according to their disorder content into WSPs (with a percent of disordered residues $\leq 10\%$), MDPs (with a percent of disordered residues between 10% and 30%) and IDPs (percent of disordered residues $\geq 30\%$). We observed that IDPs are

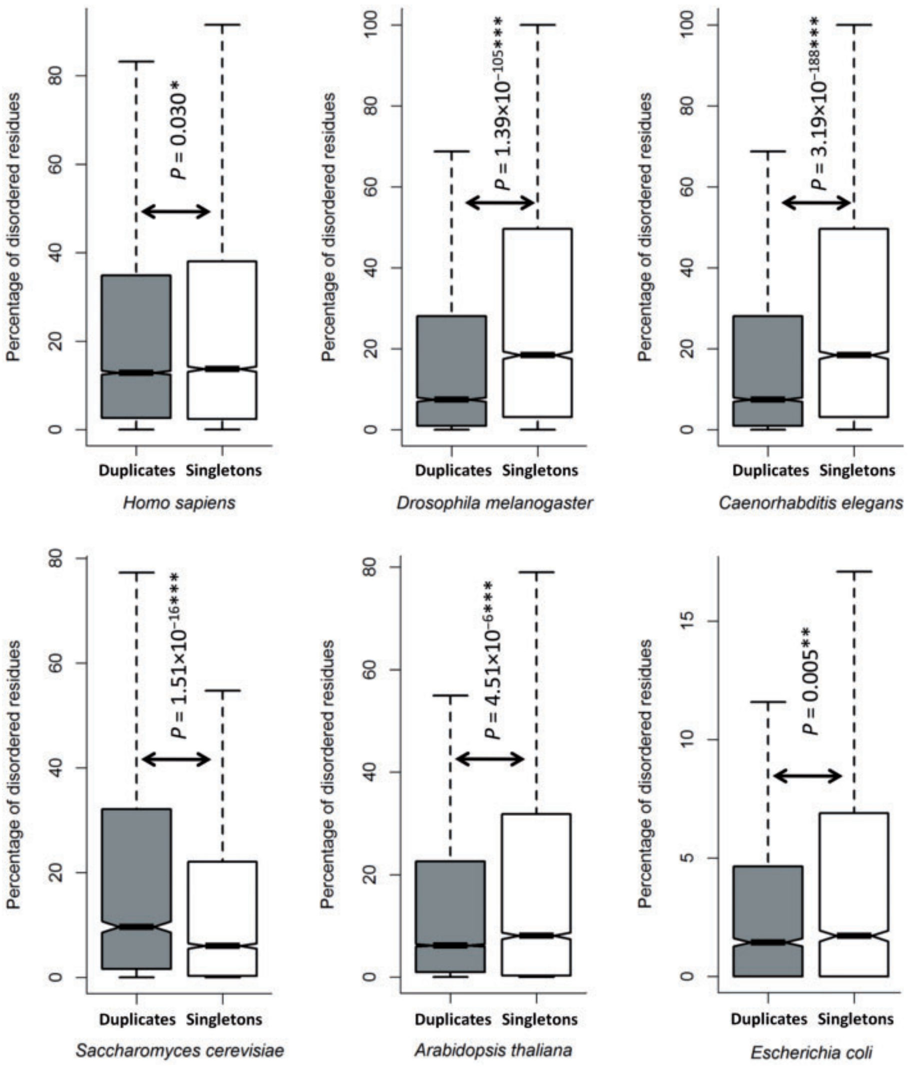

Figure 1. Differences in the percentage of disordered residues between proteins encoded by duplicated and singleton genes. P values correspond to the Wilcoxon rank sum test. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

enriched in proteins encoded by singleton genes and depleted in proteins encoded by duplicated genes (Fig. 3). For instance, in *D. melanogaster*, 39.30% of WSPs, 46.32% of MDPs and 60.72% of IDPs are encoded by singleton genes (Pearson's χ^2 test, $P = 2.2 \times 10^{-16}$; Supplementary Table S5). The only exception was again *S. cerevisiae*, where IDPs were enriched in proteins encoded by duplicated genes. However, when ohnologs and SSDs were considered separately in human, *S. cerevisiae* and *A. thaliana*, IDPs were significantly depleted in proteins encoded by SSDs in all species (Fig. 3). Furthermore, we noticed that the percentage of proteins encoded by SSDs gradually decreases from the class of WSPs to that of IDPs (Fig. 3; Supplementary Table S5). Similar results were obtained when proteins were classified using more stringent criteria (WSPs: percent of disordered residues $\leq 15\%$, MDPs: $15\% <$ percentage of disordered residues $< 60\%$, IDPs: percentage of disordered residues $\geq 60\%$) (Supplementary Table S6). Taken together, these observations are consistent with those presented in the previous section (Figs 1 and 2; Supplementary Tables S1 and S2), and indicate that genes encoding IDPs are less likely to undergo small-scale duplication than those encoding WSPs.

We found that the fraction of ohnologs is higher among IDPs than among WSPs in all species, and particularly in *S. cerevisiae* (Fig. 3). This observation is consistent with our observations that ohnologs tend to encode highly disordered proteins, especially in *S. cerevisiae* (Fig. 2).

3.3. Our observations are not due to potentially confounding factors

In some species, singletons, SSDs and ohnologs have been shown to differ in terms of expression level, protein abundance and number of protein–protein interactions.[55,65–70] In addition, these factors have been shown to correlate with proteins' disorder content in some species.[11,15,19,71–75] Combined, these trends raise the possibility that our observations (low duplicability of genes encoding IDPs) might simply be due to covariation of duplicability and intrinsic disorder with these factors. To discard this possibility, we used partial correlation analysis to evaluate the relationship between duplicability (which we represented as a binary variable taking the value of 1 for duplicated genes and 0 for singleton genes) and the percent of intrinsic disorder,

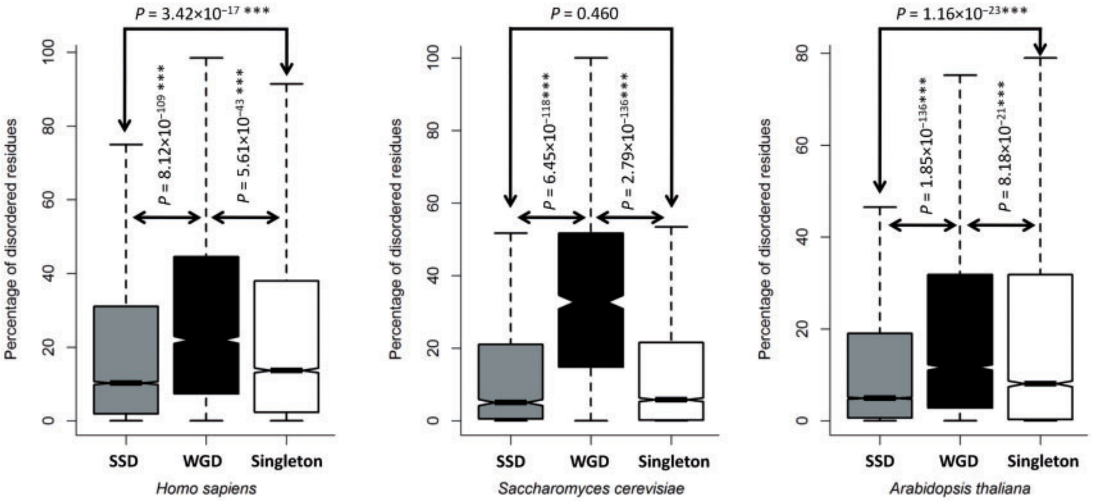

Figure 2. Differences in the percentage of disordered residues between proteins encoded by duplicates resulting from small-scale duplications (SSDs), genes resulting from whole-genome duplications (WGDs) and singleton genes. *P* values correspond to the Wilcoxon rank sum test. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

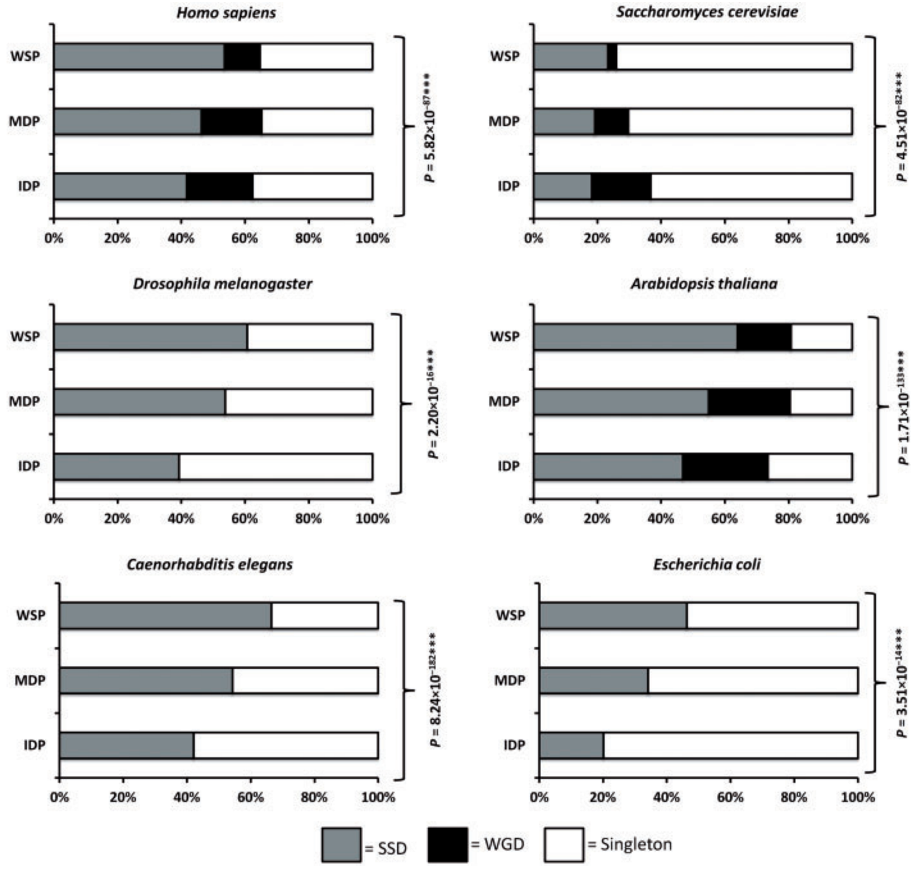

Figure 3. Proportions of duplicates and singletons among genes encoding intrinsically disordered proteins (IDPs), moderately disordered proteins (MDPs) and well-structured proteins (WSPs). In human, *S. cerevisiae* and *A. thaliana*, small-scale duplicates (SSDs), and whole genome duplicates (WGDs, or ohnologs) are considered separately. *P* values correspond to Pearson's χ^2 test. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

while controlling for all three factors (mRNA abundance, protein abundance and number of protein–protein interactions) simultaneously. In four of the species, we observed a significant association between duplicability and disorder, with duplicated genes being less disordered (Table 1). In human, the test was not significant, and in *S. cerevisiae* the partial correlation between disorder and duplicability was positive; however, removing ohnologs from the analyses resulted in significant negative partial correlations in all species (Table 1).

Table 1. Partial correlations between the percentage of disordered residues and gene duplicability

Organism	Whole dataset				Excluding ohnologs			
	N	*ρ*	*P* value	*q* value	*N*	*ρ*	*P* value	*q* value
H. sapiens	10,153	−0.001	0.920	0.920	8,057	−0.047	2.75×10^{-5}***	4.13×10^{-5}***
D. melanogaster	5,259	−0.272	4.46×10^{-49}***	2.68×10^{-48}***	—	—	—	—
C. elegans	2,474	−0.182	9.14×10^{-20}***	2.74×10^{-19}***	—	—	—	—
S. cerevisiae	4,662	0.074	4.04×10^{-7}***	8.08×10^{-7}***	4,161	−0.043	0.005**	0.005**
A. thaliana	6,642	−0.037	2.30×10^{-3}**	0.004**	4,733	−0.073	6.14×10^{-7}***	1.84×10^{-6}***
E. coli	1,176	−0.072	0.013*	0.016*	—	—	—	—

Partial Spearman's correlation coefficients (ρ) correspond to the correlation between the percent of intrinsic disorder of proteins and duplicability (encoded as a binary variable: 0 = singleton, 1 = duplicated) controlling simultaneously for mRNA abundance, protein abundance and number of protein–protein interactions. For organisms with documented whole-genome duplication events (human, *S. cerevisiae* and *A. thaliana*), we have repeated the test excluding ohnologs. *P*-values correspond to the partial correlation test. *q* values correspond to the Benjamini–Hochberg correction for multiple testing. *, *P* or $q < 0.05$; **, *P* or $q < 0.01$; ***, *P* or $q < 0.001$.

Furthermore, we examined the correlation between gene duplicability and intrinsic disorder controlling for each of the abovementioned confounding factors separately. In all cases (for all factors and species), partial correlations were significantly negative (Supplementary Table S7).

Equivalent results were obtained when, for human, *D. melanogaster* and *A. thaliana*, the expression level of each gene was computed as the average across all tissues in which it is detectably expressed, rather than across all tissues (Supplementary Table S8). The only difference is that, for *A. thaliana*, considering the entire dataset, the test is only marginally significant after correcting for multiple testing ($P = 0.037$; $q = 0.055$). Even though the magnitudes of partial correlation coefficients are small for some species (particularly when controlling for all three factors simultaneously), they are highly significant. Taken together, these results indicate that the association between duplicability and disorder is independent of expression level, protein abundance and connectivity.

3.4. Natural selection often removes genes encoding IDPs after duplication

Our observations that SSDs tend to encode lowly disordered proteins, and that IDPs are generally more likely to be encoded by singleton genes than MDPs and WSPs, are consistent with a scenario in which purifying selection limits the small-scale duplicability of genes encoding IDPs. However, an alternative scenario might also explain these observations. It is conceivable that, after duplication, genes accumulate mutations that decrease the disorder content of the encoded proteins. To distinguish between both scenarios, we performed two additional analyses.

If extra copies of genes encoding IDPs tend to be removed after gene duplication by purifying selection, one may expect the ancestral (pre-duplication) sequences of duplicated genes to encode, on average, less disordered proteins than the ancestral sequences of singleton genes. If this is the case, one would expect that orthologs in an outgroup species (e.g. *Drosophila grimshawi*) of genes that have duplicated in one of the studied species (e.g. *D. melanogaster*) would encode less disordered proteins than orthologs in the outgroup (e.g. *D. grimshawi*) of genes that have not duplicated in the species of interest (e.g. *D. melanogaster*). To test this hypothesis in *Drosophila*, we classified all *D. grimshawi* genes into three groups (Fig. 4): (A) those that have a single ortholog in *D. melanogaster* (i.e. they have not duplicated in the branch connecting *D. melanogaster* and the most recent common ancestor of *D. melanogaster* and *D. grimshawi*); (B) those that have two or more orthologs in *D. melanogaster* (i.e. they have duplicated in the *D. melanogaster* lineage); and (C) those that have no orthologs in *D. melanogaster* (either have been lost in the *D. melanogaster* lineage or originated in the *D. grimshawi* lineage). We found that proteins encoded by genes in group A were significantly more disordered than those encoded by genes in group B (median for group A: 11.46%; median for group B: 2.27%; Mann–Whitney *U* test, $P = 2.80 \times 10^{-33}$). This suggests that purifying selection removes extra copies of genes encoding highly disordered proteins after small-scale duplication. Similar results were obtained in human, *C. elegans*, *S. cerevisiae*, *A. thaliana* and *E. coli* using as outgroup, respectively, chicken, *Caenorhabditis japonica*, *Candida glabrata*, *Arabidopsis lyrata* and *Mycobacterium tuberculosis* (Table 2). For organisms that have undergone WGD events (human, *S. cerevisiae* and *A. thaliana*), we chose outgroup species that are known to have shared the same WGD histories.[61,76–79] For these organisms, similar results were obtained when the analyses were restricted to ohnologs (Supplementary Table S9) and to non-ohnologs (Supplementary Table S10). The only exception was human/chicken ohnologs: proteins encoded by genes in group A were more disordered on average than those encoded by genes in group B, but the differences were not statistically significant (Supplementary Table S9).

Alternatively, if proteins become less disordered after gene duplication, one would expect genes that have duplicated in the species of interest (e.g. *D. melanogaster*) to encode less disordered proteins than those encoded by their orthologs in the outgroup species (e.g. *D. grimshawi*) that have not undergone duplication. In order to test this possibility in *Drosophila*, we identified a total of 258 groups of orthologous genes that had duplicated in *D. melanogaster* but not in *D. grimshawi*. Each of these groups contained one *D. grimshawi* gene and more than one *D. melanogaster* gene. In each group, the *D. grimshawi* gene had multiple co-orthologs in *D. melanogaster*, and the *D. melanogaster* genes shared a single ortholog in *D. grimshawi*. We found no statistically significant differences between the degree of disorder of proteins encoded by *D. grimshawi* genes and proteins encoded by *D. melanogaster* duplicates (Table 3). In 105 out of the 258 groups, the disorder content of the *D. melanogaster* proteins was higher than the average disorder content of the *D. grimshawi* protein, whereas in 113 of the groups the *D. melanogaster* proteins were less disordered (in the other 40, the percent of disordered residues was the same in both species), which did not represent a significant departure from the 50%:50% of groups (109 and 109) with each trend randomly expected (binomial test, $P = 0.635$). Similar, non-significant differences were observed in the other studied species, except for *E. coli* and *S. cerevisiae*, in which significant differences were observed (Table 3). These observations disfavour the hypothesis that the lower disorder content of proteins encoded by duplicated genes is due to accumulation of mutations after duplication.

Table 2. Percentage of disordered residues in outgroup genes of different classes

Organism	Outgroup	Group A (non-duplicated)			Group B (duplicated)			Group C (lost)			*P*-value			*q*-value		
		N	Mean (%)	Median	*N*	Mean (%)	Median (%)	*N*	Mean (%)	Median (%)	A vs. B	A vs. C	B vs. C	A vs. B	A vs. C	B vs. C
H. sapiens	*G. gallus*	12,556	20.96	12.23	852	16.76	7.75	2,100	21.52	8.56	$4.99 \times 10^{-36***}$	$1.65 \times 10^{-10***}$	0.020*	$2.994 \times 10^{-35***}$	$3.30 \times 10^{-10***}$	0.024*
D. melanogaster	*D. grimshawi*	11,761	22.26	11.46	508	14.22	2.27	2,713	38.11	30.89	$2.80 \times 10^{-33***}$	$1.12 \times 10^{-09***}$	$1.75 \times 10^{-20**}$	$8.4 \times 10^{-33***}$	$1.68 \times 10^{-09***}$	$1.05 \times 10^{-19***}$
C. elegans	*C. japonica*	10,646	18.57	7.24	4173	15.82	5.47	15,061	24.60	11.62	$9.77 \times 10^{-4***}$	$1.76 \times 10^{-27***}$	$3.07 \times 10^{-5***}$	0.001**	$5.28 \times 10^{-27***}$	$6.14 \times 10^{-05***}$
S. cerevisiae	*C. glabrata*	3739	19.05	9.09	973	18.32	10.08	510	23.88	13.08	0.002*	0.450	0.671	0.002**	0.450	0.671
A. thaliana	*A. lyrata*	23,941	16.72	6.90	857	9.98	1.90	7,939	17.53	2.90	$5.60 \times 10^{-27***}$	$<10^{-36***}$	$2.00 \times 10^{-04***}$	$1.12 \times 10^{-26***}$	$<10^{-35***}$	$3.00 \times 10^{-04***}$
E. coli	*M. tuberculosis*	930	9.91	5.84	260	7.00	5.69	2,695	13.47	6.86	0.014*	$4.01 \times 10^{-04***}$	$4.80 \times 10^{-6***}$	0.002*	$4.81 \times 10^{-04***}$	$1.44 \times 10^{-05***}$

Group A: genes in the outgroup species that remain singleton in the studied organism. Group B: genes in the outgroup species that have duplicated in the studied organism. Group C: genes in the outgroup species that have been lost in the studied organisms. *P*-values correspond to the Wilcoxon rank sum test. *q* values correspond to the Benjamini-Hochberg correction for multiple testing. *, P or $q < 0.05$; **, P or $q < 0.01$; ***, P or $q < 0.001$.

3.5. Discussion

We have found that, in general, SSDs tend to encode proteins that are less intrinsically disordered than those encoded by singleton genes (Fig. 1), an observation that is not due to covariation of mRNA abundance, protein abundance or network centrality with both intrinsic disorder and duplicability. In addition, IDPs are generally more likely to be encoded by singleton genes than MDPs and WSPs (Fig. 3), and non-duplicated orthologs of duplicated genes tend to be lowly disordered (Fig. 4 and Table 2). The trend has been observed across a wide range of organisms, including a bacterium, a plant, a fungus, two invertebrates and a vertebrate.

Taken together, these observations indicate that duplicates encoding IDPs are less likely to be retained after small-scale gene duplication than genes encoding WSPs or MDPs. This is consistent with a scenario in which small-scale duplication of genes encoding IDPs is often deleterious (more often than those encoding WSPs or MDPs), and duplicates are thus often removed by purifying selection (more often than those encoding WSPs or MDPs). Compatible with this scenario, Vavouri et al.[15] found that yeast dosage-sensitive genes (those that impact fitness negatively upon over-expression) tend to encode IDPs.

According to the interaction promiscuity hypothesis,[15] given their high structural flexibility and enrichment in interaction domains, an increased concentration of any IDP is expected to result in an increased number of misinteractions (i.e. unwanted un-specific interactions). Many proteins exhibit both physiological targets and non-physiological ones, with which they unavoidably interact with low affinity. Even if a protein's affinity for non-physiological targets is low, an increase in the protein's concentration is expected to increase the number of non-physiological interactions—due to mass action, any two proteins will interact if present at sufficiently high concentrations. This is expected to especially apply to IDPs, which are particularly flexible and rich in promiscuous short linear motifs, and are thus expected to be promiscuous in their patterns of interaction.[15] Misinteractions can have a number of deleterious (or even cytotoxic) effects, by producing (i) a waste of functional proteins, some of which can become sequestered in non-functional complexes (molecular titration); (ii) interference with functional interactions, and/or (iii) unwanted initiation of cellular processes.[75] As expected from the potential deleterious effects of IDP dysregulation, several observations indicate that the availability of IDPs is tightly regulated by a variety of mechanisms, including increased mRNA decay rates and increased proteolytic degradation.[15–20] It should be noted, however, that not all gene duplications result in increased protein abundances,[80] and that not all IDPs produce deleterious effects upon over-expression (see Ref. [16] and references therein).

We found that ohnologs tend to encode proteins that are more disordered than those encoded by singletons or SSDs (Fig. 2). In addition, the fraction of proteins encoded by ohnologs is higher among IDPs than among WSPs (Fig. 3). These observations are in agreement

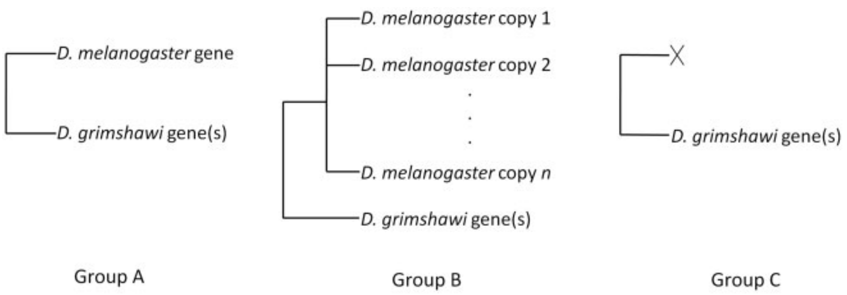

Figure 4. Classification of genes in the outgroup species according to the duplication status of the orthologs in the species of interest. Group A: genes in the outgroup species that remain singleton in the studied organism. Group B: genes in the outgroup species that have duplicated in the studied organism. Group C: genes in the outgroup species that have been lost in the studied organisms. The figure depicts an example in which *D. melanogaster* is the studied species and *D. grimshawi* is the outgroup species. If purifying selection tends to remove the duplicates of genes encoding highly disordered proteins, then we expect proteins in group A to be significantly more disordered than those in group B.

Table 3. Cases in which proteins encoded by duplicated genes in the studied species are more or less disordered than proteins encoded by their non-duplicated orthologs in outgroup species

Organism	Outgroup	Case I	Case II	Case III	*P* value	*q* value
H. sapiens	*G. gallus*	282	251	34	0.1937	0.3880
D. melanogaster	*D. grimshawi*	105	113	40	0.6355	0.9290
C. elegans	*C. japonica*	248	251	70	0.9287	0.9290
S. cerevisiae	*C. glabrata*	77	119	12	0.0002***	0.0012**
A. thaliana	*A. lyrata*	200	204	91	0.8814	0.9290
E. coli	*M. tuberculosis*	20	44	1	0.0026**	0.0078**

Case I: number of cases in which proteins encoded by duplicated genes in the organism of interest are more disordered than proteins encoded by their non-duplicated ortholog in the outgroup species. Case II: number of cases in which proteins encoded by duplicated genes in the organism of interest are less disordered than proteins encoded by their non-duplicated ortholog in the outgroup species. Case III: number of cases in which proteins encoded by duplicated genes in the organism of interest are as disordered as proteins encoded by their non-duplicated ortholog in the outgroup species. *P*-values correspond to the binomial test (comparison of cases I and II vs. the 50%:50% expected by chance). *Q*-values correspond to the Benjamini–Hochberg correction for multiple testing. *, *P* or $q < 0.05$; **, *P* or $q < 0.01$; ***, *P* or $q < 0.001$.

with prior observations in yeasts that proteins encoded by ohnologs tend to be more disordered than those encoded by singleton genes.[25] However, these observations appear to be at odds with our observations that, overall, duplicated genes tend to encode lowly disordered proteins (Fig. 1). It should be noted, nonetheless, that ohnologs duplicated in a very specific context, in which all genes duplicated simultaneously. After a WGD event, the stoichiometry of all protein–protein interactions is maintained.[29,30] In addition, WGD is thought to be often accompanied by an increase in cell volume,[31–34] meaning that the concentration of each protein after WGD may be similar to that before WGD. Therefore, duplication of ohnologs probably did not have the same deleterious effects expected for small-scale duplications (which alter the stoichiometry of the system and result in increased protein concentrations). Being free of these negative effects, ohnologs probably were able to exploit the duplication-promoting effects of IDRs–IDRs, and/or the post-translational modification sites in which they are enriched, may have facilitated functional diversification, which may have promoted retention of genes encoding IDPs after WGD.[24,25] Remarkably, and consistent with our model, ohnologs (which tend to encode highly disordered proteins; Fig. 2; Ref. [25]) are unlikely to duplicate by mechanisms other than WGD,[81] and copy-number variation of these genes is often associated with disease.[82] Marcet-Houben and Gabaldón[28] have recently proposed an alternative mechanism for the presence of 'ohnologs' in the *S. cerevisiae* lineage: a recent hybridization of two closely related yeasts (if this is true, yeast genes thus far considered 'ohnologs' should actually be considered 'synologs'; Ref. [83]). Nonetheless, hybridization of closely related species is also expected to result in increased cell size and to respect the stoichiometry of all interactions.

Montanari et al.[25] observed that after WGD, yeast ohnologs tend to experience a net loss in their disorder content (a behaviour, however, that was not observed in all genes). This raises the possibility that, given enough time, the differences between ohnologs and the other genes would disappear, or even invert (resulting in ohnologs encoding the less disordered proteins). Our analyses confirm, however, that this is not the case for any of the species analyzed in our study (Fig. 2; Supplementary Table S2).

Despite the general tendency of ohnologs to encode highly disordered proteins, among ohnologs, those that underwent subsequent duplications tend to be lowly disordered (Supplementary Table S9). This reinforces our model that genes encoding IDPs are less likely to undergo duplication. This applies even if they are ohnologs, because gene families that stem from WGD and that encode IDPs are less likely to undergo further expansion than those that do not encode IDPs.

Acknowledgements

We are grateful to Mario A. Fares for helpful discussions during the development of this work, and to Sandip Chakraborty for critically reading the manuscript. We are also grateful to two anonymous referees for their helpful comments.

Conflict of interest

None declared.

Funding

This work was supported by funds from the University of Nevada, Reno awarded to DAP.

Supplementary data

Supplementary data are available at *DNARES* Online.

References

1. Zhang, J.Z. 2003, Evolution by gene duplication: an update, *Trends Ecol. Evol.* **18**, 292–8.
2. Lynch, M. and Conery, J.S. 2000, The evolutionary fate and consequences of duplicate genes, *Science*, **290**, 1151–5.
3. Ohno, S. 1970, *Evolution by gene duplication*. Springer-Verlag: Berlin.
4. Kondrashov, F.A. and Kondrashov, A.S. 2006, Role of selection in fixation of gene duplications, *J. Theor. Biol.*, **239**, 141–51.
5. Oldfield, C.J. and Dunker, A.K. 2014, Intrinsically disordered proteins and intrinsically disordered protein regions, *Annu. Rev. Biochem.*, **83**, 553–84.
6. Dunker, A.K., Lawson, J.D., Brown, C.J., et al. 2001, Intrinsically disordered protein, *J. Mol. Graph. Model.*, **19**, 26–59.
7. Sugase, K., Dyson, H.J. and Wright, P.E. 2007, Mechanism of coupled folding and binding of an intrinsically disordered protein, *Nature*, **447**, 1021–5.
8. Wright, P.E. and Dyson, H.J. 1999, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J. Mol. Biol.*, **293**, 321–31.
9. Fuxreiter, M., Tompa, P. and Simon, I. 2007, Local structural disorder imparts plasticity on linear motifs, *Bioinformatics*, **23**, 950–6.
10. Dyson, H.J. and Wright, P.E. 2002, Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol.*, **12**, 54–60.
11. Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M. and Uversky, V.N. 2005, Flexible nets. The roles of intrinsic disorder in protein interaction networks, *FEBS J.*, **272**, 5129–48.
12. Gunasekaran, K., Tsai, C.J., Kumar, S., Zanuy, D. and Nussinov, R. 2003, Extended disordered proteins: targeting function with less scaffold, *Trends Biochem. Sci.*, **28**, 81–5.
13. Kim, P.M., Sboner, A., Xia, Y. and Gerstein, M. 2008, The role of disorder in interaction networks: a structural analysis, *Mol. Syst. Biol.*, **4**, 179.
14. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. 2002, Intrinsic disorder in cell-signaling and cancer-associated proteins, *J. Mol. Biol.*, **323**, 573–84.
15. Vavouri, T., Semple, J.I., Garcia-Verdugo, R. and Lehner, B. 2009, Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity, *Cell*, **138**, 198–208.
16. Babu, M.M., van der Lee, R., de Groot, N.S. and Gsponer, J. 2011, Intrinsically disordered proteins: regulation and disease, *Curr. Opin. Struct. Biol.*, **21**, 432–40.
17. Gsponer, J., Futschik, M.E., Teichmann, S.A. and Babu, M.M. 2008, Tight regulation of unstructured proteins: from transcript synthesis to protein degradation, *Science*, **322**, 1365–8.
18. Edwards, Y.J., Lobley, A.E., Pentony, M.M. and Jones, D.T. 2009, Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data, *Genome Biol.*, **10**, R50.
19. Chen, J., Liang, H. and Fernández, A. 2008, Protein structure protection commits gene expression patterns, *Genome Biol.*, **9**, R107.
20. Radivojac, P., Vacic, V., Haynes, C., et al. 2010, Identification, analysis, and prediction of protein ubiquitination sites, *Proteins*, **78**, 365–80.
21. Uversky, V.N., Oldfield, C.J. and Dunker, A.K. 2008, Intrinsically disordered proteins in human diseases: introducing the D2 concept, *Annu. Rev. Biophys.*, **37**, 215–46.
22. Niklas, K.J., Bondos, S.E., Dunker, A.K. and Newman, S.A. 2015, Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications, *Front. Cell. Dev. Biol.*, **3**, 8.
23. Kurotani, A. and Sakurai, T. 2015, In silico analysis of correlations between protein disorder and post-translational modifications in algae, *Int. J. Mol. Sci.*, **16**, 19812–35.
24. Amoutzias, G.D., He, Y., Gordon, J., Mossialos, D., Oliver, S.G. and Van de Peer, Y. 2010, Posttranslational regulation impacts the fate of duplicated genes, *Proc. Natl. Acad. Sci. USA*, **107**, 2967–71.
25. Montanari, F., Shields, D.C. and Khaldi, N. 2011, Differences in the number of intrinsically disordered regions between yeast duplicated proteins, and their relationship with functional divergence, *PLoS One*, **6**, e24989.
26. Wolfe, K. 2000, Robustness – it's not where you think it is, *Nat. Genet.*, **25**, 3–4.
27. Wolfe, K.H. and Shields, D.C. 1997, Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, **387**, 708–13.
28. Marcet-Houben, M. and 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.
29. Veitia, R.A. 2004, Gene dosage balance in cellular pathways: implications for dominance and gene duplicability, *Genetics*, **168**, 569–74.
30. Veitia, R.A. 2005, Paralogs in polyploids: one for all and all for one? *Plant Cell*, **17**, 4–11.
31. Bomblies, K. and Madlung, A. 2014, Polyploidy in the Arabidopsis genus, *Chromosome Res.*, **22**, 117–34.
32. Cavalier-Smith, T. 1978, Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox, *J. Cell. Sci.*, **34**, 247–78.
33. Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S. and Fink, G.R. 1999, Ploidy regulation of gene expression, *Science*, **285**, 251–4.
34. Hennaut, C., Hilger, F. and Grenson, M. 1970, Space limitation for permease insertion in the cytoplasmic membrane of *Saccharomyces cerevisiae*, *Biochem. Biophys. Res. Commun.*, **39**, 666–71.
35. Papp, B., Pal, C. and Hurst, L.D. 2003, Dosage sensitivity and the evolution of gene families in yeast, *Nature*, **424**, 194–7.
36. Birchler, J.A., Bhadra, U., Bhadra, M.P. and Auger, D.L. 2001, Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits, *Dev. Biol.*, **234**, 275–88.
37. Veitia, R.A. 2002, Exploring the etiology of haploinsufficiency, *Bioessays*, **24**, 175–84.
38. Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. 2007, Natural history and evolutionary principles of gene duplication in fungi, *Nature*, **449**, 54–61.
39. Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G. and Robertson, D.L. 2007, All duplicates are not equal: the difference between small-scale and genome duplication, *Genome Biol.*, **8**, R209.
40. Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. 2005, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, **21**, 3433–4.
41. Hegedus, T., Serohijos, A.W., Dokholyan, N.V., He, L. and Riordan, J.R. 2008, Computational studies reveal phosphorylation-dependent changes in the unstructured R domain of CFTR, *J. Mol. Biol.*, **378**, 1052–63.
42. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., et al. 2005, FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435–8.
43. Van der Lee, R., Buljan, M., Lang, B., et al. 2014, Classification of intrinsically disordered regions and proteins, *Chem. Rev.*, **114**, 6589–631.
44. Bolser, D., Staines, D.M., Pritchard, E. and Kersey, P. 2016, Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data, *Methods Mol. Biol.*, **1374**, 115–40.
45. Kinsella, R.J., Kahari, A., Haider, S., et al. 2011, Ensembl BioMarts: a hub for data retrieval across taxonomic space, *Database (Oxford)*, **2011**, bar030.

46. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
47. Singh, P.P., Arora, J. and Isambert, H. 2015, Identification of Ohnolog genes originating from whole genome duplication in early vertebrates, based on Synteny comparison across multiple genomes, *PLoS Comput. Biol.*, **11**, e1004394.
48. Gordon, J.L., Byrne, K.P. and Wolfe, K.H. 2009, Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome, *PLoS Genet.*, **5**, e1000485.
49. Blanc, G., Hokamp, K. and Wolfe, K.H. 2003, A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome, *Genome Res.*, **13**, 137–44.
50. Uhlén, M., Fagerberg, L., Hallström, B.M., et al. 2015, Proteomics. Tissue-based map of the human proteome, *Science*, **347**, 1260419.
51. Chintapalli, V.R., Wang, J. and Dow, J.A. 2007, Using FlyAtlas to identify better Drosophila melanogaster models of human disease, *Nat. Genet.*, **39**, 715–20.
52. Petryszak, R., Keays, M., Tang, Y.A., et al. 2016, Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants, *Nucleic Acids Res.*, **44**, D746–52.
53. Nagalakshmi, U., Wang, Z., Waern, K., et al. 2008, The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*, **320**, 1344–9.
54. Schmid, M., Davison, T.S., Henz, S.R., et al. 2005, A gene expression map of Arabidopsis thaliana development, *Nat. Genet.*, **37**, 501–6.
55. Alvarez-Ponce, D. and Fares, M.A. 2012, Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network, *Genome Biol. Evol.*, **4**, 1263–74.
56. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. 2004, Integrating high-throughput and computational data elucidates bacterial networks, *Nature*, **429**, 92–6.
57. Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D. and von Mering, C. 2015, Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines, *Proteomics*, **15**, 3163–8.
58. Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., et al. 2015, The BioGRID interaction database: 2015 update, *Nucleic Acids Res.*, **43**, D470–8.
59. Hu, P., Janga, S.C., Babu, M., et al. 2009, Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins, *PLoS Biol.*, **7**, e1000096.
60. Li, L., Stoeckert, C.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res.*, **13**, 2178–89.
61. Dehal, P. and Boore, J.L. 2005, Two rounds of whole genome duplication in the ancestral vertebrate, *PLoS Biol.*, **3**, e314.
62. Wolfe, K.H. 2015, Origin of the yeast whole-genome duplication, *PLoS Biol.*, **13**, e1002221.
63. Blanc, G., Barakat, A., Guyot, R., Cooke, R. and Delseny, M. 2000, Extensive duplication and reshuffling in the Arabidopsis genome, *Plant Cell*, **12**, 1093–101.
64. Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M. and Van de Peer, Y. 2002, The hidden duplication past of Arabidopsis thaliana, *Proc. Natl. Acad. Sci. USA*, **99**, 13627–32.
65. Wang, Y., Wang, X. and Paterson, A.H. 2012, Genome and gene duplications and gene expression divergence: a view from plants, *Ann. N. Y. Acad. Sci.*, **1256**, 1–14.
66. Rogozin, I.B. 2014, Complexity of gene expression evolution after duplication: protein dosage rebalancing, *Genet. Res. Int.*, **2014**, 516508.
67. Makova, K.D. and Li, W.H. 2003, Divergence in the spatial pattern of gene expression between human duplicate genes, *Genome Res.*, **13**, 1638–45.
68. Gu, X., Zhang, Z. and Huang, W. 2005, Rapid evolution of expression and regulatory divergences after yeast gene duplication, *Proc.Natl. Acad. Sci. USA*, **102**, 707–12.
69. Li, W.H., Yang, J. and Gu, X. 2005, Expression divergence between duplicate genes, *Trends Genet.*, **21**, 602–7.
70. D'Antonio, M. and Ciccarelli, F.D. 2011, Modification of gene duplicability during the evolution of protein interaction network, *PLoS Comput. Biol.*, **7**, e1002029.
71. Singh, G.P. and Dash, D. 2008, How expression level influences the disorderness of proteins, *Biochem. Biophys. Res. Commun.*, **371**, 401–4.
72. Paliy, O., Gargac, S.M., Cheng, Y., Uversky, V.N. and Dunker, A.K. 2008, Protein disorder is positively correlated with gene expression in Escherichia coli, *J. Proteome Res.*, **7**, 2234–45.
73. Singh, G.P., Ganapathi, M. and Dash, D. 2007, Role of intrinsic disorder in transient interactions of hub proteins, *Proteins*, **66**, 761–5.
74. Ma, L., Pang, C.N., Li, S.S. and Wilkins, M.R. 2010, Proteins deleterious on overexpression are associated with high intrinsic disorder, specific interaction domains, and low abundance, *J. Proteome Res.*, **9**, 1218–25.
75. Yang, J.R., Liao, B.Y., Zhuang, S.M. and Zhang, J. 2012, Protein misinteraction avoidance causes highly expressed proteins to evolve slowly, *Proc. Natl. Acad. Sci. USA*, **109**, E831–40.
76. Kellis, M., Birren, B.W. and Lander, E.S. 2004, Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*, *Nature*, **428**, 617–24.
77. Hufton, A.L., Groth, D., Vingron, M., Lehrach, H., Poustka, A.J. and Panopoulou, G. 2008, Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement, *Genome Res.*, **18**, 1582–91.
78. Ahmad, K.M., Kokošar, J., Guo, X., Gu, Z., Ishchuk, O.P. and Piškur, J. 2014, Genome structure and dynamics of the yeast pathogen Candida glabrata, *FEMS Yeast Res.*, **14**, 529–35.
79. Hu, T.T., Pattyn, P., Bakker, E.G., et al. 2011, The Arabidopsis lyrata genome sequence and the basis of rapid genome size change, *Nat. Genet.*, **43**, 476–81.
80. Cardoso-Moreira, M., Arguello, J.R., Gottipati, S., Harshman, L.G., Grenier, J.K. and Clark, A.G. 2016, Evidence for the fixation of gene duplications by positive selection in Drosophila, *Genome Res.*, **26**, 787–98.
81. Makino, T. and McLysaght, A. 2010, Ohnologs in the human genome are dosage balanced and frequently associated with disease, *Proc. Natl. Acad. Sci. USA*, **107**, 9270–4.
82. McLysaght, A., Makino, T., Grayton, H.M., et al. 2014, Ohnologs are overrepresented in pathogenic copy number mutations, *Proc. Natl. Acad. Sci. USA*, **111**, 361–6.
83. Gogarten, J.P. 1994, Which is the most conserved group of proteins? Homology–orthology, paralogy, xenology, and the fusion of independent lineages, *J. Mol. Evol.*, **39**, 541–3.