Intrinsic protein disorder reduces small-scale gene duplicability

Sanghita Banerjee1,2, Felix Feyertag1, and David Alvarez-Ponce1,*

1Department of Biology, University of Nevada, Reno, NV 89557, USA, and 2Machine 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

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.3

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.\textsuperscript{12} As a result, IDPs tend to be promiscuous in their interaction patterns, being involved in a high number of protein–protein interactions.\textsuperscript{10,11,13} Proteins involved in signalling, including transcription factors, tend to be rich in IDRs.\textsuperscript{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.\textsuperscript{15,16} Indeed, IDP availability is often maintained at low levels, and several lines of evidence indicate that this availability is tightly regulated\textsuperscript{15–20} and that dysregulation of IDPs often leads to disease, including neurodegeneration and cancer.\textsuperscript{14,15,21} Among other deleterious effects. Remarkably, Vavouri et al.\textsuperscript{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 neo-functionalization) of gene copies after duplication, which promotes retention of the duplicates. In addition, IDRs are enriched in post-translational modifications.\textsuperscript{22,23} which also contribute to functional divergence of gene duplicates.\textsuperscript{24} Consistent with this model, Montanari et al.\textsuperscript{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—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.\textsuperscript{29,30} and often result in an increased cell volume, which means that protein concentrations are not necessarily altered.\textsuperscript{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.\textsuperscript{29,35–37} Therefore, the selective pressures constraining ohnologs retention are expected to be different from those acting on other kinds of duplicates.\textsuperscript{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.

\section{2. Material and methods}

\subsection*{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\textsuperscript{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.\textsuperscript{42}

We classified proteins based on their disorder content, as either IDPs (percentage of disordered residues ≥ 50%), moderately disordered proteins (MDPs, 10% < percentage of disordered residues < 30%) or well-structured proteins (WSPs, percentage of disordered residues ≤ 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 ≥ 60%), MDPs (15% < percentage of disordered residues < 60%) and WSPs (percentage of disordered residues ≤ 15%).

\subsection*{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.\textsuperscript{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\textsuperscript{71} and modENCODE (data from the EBI Expression Atlas, accession number E-MTAB-2812\textsuperscript{72}), respectively. S. cerevisiae gene expression data were obtained from Nagalakshmi et al.\textsuperscript{13} In the case of A. thaliana, we obtained gene expression datasets corresponding to 79 tissues and conditions, from Schmid et al.\textsuperscript{54} and processed them as in
Alvarez-Ponce and Fares. 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. 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 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. 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. Only physical interactions among proteins from the same organism were considered. The *E. coli* protein–protein interaction network was obtained from Hu et al. 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. *S. cerevisiae–C. glabrata* and *E. coli–M. tuberculosis* orthologies were obtained from the OrthoMCL database.

### 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. 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 × 10^-10; 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 × 10^-16; Fig. 1; Supplementary Table S1).

Montanari et al. 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, *S. cerevisiae* and *A. thaliana*. We observed that, in all three organisms, proteins encoded by SSDs represented the least disordered class, and that, in agreement with Montanari et al., 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*. 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 × 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, 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 × 10^-10; 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 ≤ 10%), MDPs (with a percent of disordered residues between 10% and 30%) and IDPs (percent of disordered residues ≥ 30%). We observed that IDPs are
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 $\chi^2$ 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.\(^5\)\(^5\),\(^6\)\(^5\),\(^6\)\(^5\),\(^7\)\(^0\),\(^7\)\(^1\),\(^7\)\(^2\)\(^7\)\(^3\)\(^7\)\(^4\) 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,
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).

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 < 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 < 0.05; **P < 0.01; ***P < 0.001.
Furthermore, we examined the correlation between gene duplicability and intrinsic disorder controlling for each of the above-mentioned 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 removes extra copies of encoding highly disordered proteins after small-scale duplication. Similar results were obtained in human, *C. elegans*, *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

---

**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^{-4} |
| *D. melanogaster* | 5,259 | 0.272 | 4.46 × 10^{-4}** | 2.68 × 10^{-2}** | — | — | — | — |
| *C. elegans* | 2,474 | 0.074 | 0.005** | 0.005** | 4,161 | 0.043 | 0.005** | 0.005** |
| *S. cerevisiae* | 4,662 | 0.037 | 2.30 × 10^{-3}** | 0.004** | 4,733 | 0.073 | 6.14 × 10^{-7}** | 1.84 × 10^{-6}** |
| *A. thaliana* | 6,642 | 0.013* | 0.016 | — | — | — | — | — |
| *E. coli* | 1,176 | 0.072 | — | — | — | — | — | — |

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* < 0.05; **, *P* < 0.01; ***,**P* < 0.001.
Table 2. Percentage of disordered residues in outgroup genes of different classes

| Organism          | Group A (non-duplicated) | Group B (duplicated) | Group C (lost) | \( q \)-value |
|-------------------|--------------------------|----------------------|---------------|-------------|
|                    | N                        | Mean                 | Median        |             |
| **H. sapiens**     | 13,556                   | 12.56                | 11.66         |             |
|                    |                          | 14.72                | 14.82         |             |
|                    |                          | 15.82                | 16.08         | 1.44        |
| **G. gallus**      | 11,761                   | 22.26                | 11.46         |             |
|                    |                          | 22.98                | 20.83         |             |
|                    |                          | 23.22                | 23.13         |             |
| **C. elegans**     | 10,646                   | 18.57                | 7.24          |             |
|                    |                          | 18.82                | 18.82         |             |
|                    |                          | 19.10                | 19.10         |             |
| **C. japonica**    | 23,941                   | 16.72                | 6.90          |             |
|                    |                          | 16.72                | 16.72         | 0.014       |
|                    |                          | 13.47                | 13.47         |             |

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 disfavor the hypothesis that the lower disorder content of proteins encoded by duplicated genes is due to accumulation of mutations after duplication.

### 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. found that yeast dosage-sensitive genes (those that impact fitness negatively upon over-expression) tend to encode IDPs.

According to the interaction promiscuity hypothesis, 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. 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. 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. It should be noted, however, that not all gene duplications result in increased protein abundances, 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 with each trend randomly expected (binomial test, \( P = 0.635 \)).
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. grimmhawi | 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. italicana | 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.
