Paralog dependency indirectly affects the robustness of human cells

Authors: Rohan Dandage[1,2,3,4,5] and Christian R Landry[1,2,3,4,5]

Affiliations:

1. Département de biologie
2. Département de biochimie, microbiologie et bio-informatique
3. Institut de Biologie Intégrative et des Systèmes (IBIS)
4. The Quebec Network for Research on Protein Function, Engineering, and Applications (PROTEO)
5. Centre de recherché en données massive (CRDM)

Université Laval
1030 Avenue de la médecine
Québec, Québec
G1V 0A6
Canada

Tel: 418-656-3954

Correspondence: christian.landry@bio.ulaval.ca

Abstract

The protective redundancy of paralogous genes partly relies on the fact that they carry their functions independently. However, a significant fraction of paralogous proteins may form functionally dependent pairs, for instance through heteromerization. As a consequence, one could expect these heteromeric paralogs to be less protective against deleterious mutations. To test this hypothesis, we examined the robustness landscape of gene loss-of-function by CRISPR-Cas9 in more than 450 human cell lines, which shows regions of greater deleteriousness to gene inactivation as a function of key paralog properties. We find that heteromeric paralogs are more likely to occupy such regions owing to their high expression and large number of protein-protein interaction partners. Further investigation revealed that heteromers may also be under stricter dosage balance, which may also contribute to the higher deleteriousness upon gene inactivation. Finally, we suggest that physical dependency may contribute to the deleteriousness upon loss-of-function as revealed by the correlation between the strength of interactions between paralogs and their higher deleteriousness upon loss of function.

Introduction

After a gene duplication event and before they become functionally distinct, paralogs are redundant and can mask each other's inactivating mutations (Diss et al. 2014; Brookfield 1997; Pickett and Meeks-Wagner 1995). This mutational robustness does not provide an advantage strong enough by itself to cause the maintenance of paralogs by natural selection unless mutation rate or population size are exceptionally large (van Nimwegen, Crutchfield, and Huynen 1999). Nevertheless, paralogous genes affect how biological systems globally respond to loss-of-function (LOF) mutations. For instance, the early analysis of growth rate of the yeast gene deletion collection revealed that genes with duplicates are enriched among the ones that have a weak effect on fitness when deleted (Zhenglong Gu et al. 2003). Likewise, singletons (genes with no detectable homologous sequence in the genome) tend to be overrepresented among genes whose deletion is lethal. Further studies in yeast also showed that redundancy could be maintained for millions of years, making the impact of duplication long lasting (Dean et al. 2008). A parallel observation in humans showed that genes are less likely to be involved in diseases if they have a paralog, and the probability of disease association for a gene decreases with increasing sequence similarity with its closest homolog in the genome (Hsiao and Vitkup 2008). These observations, along with smaller scale observations made in classical genetics (Pickett and Meeks-Wagner 1995; Diss et al. 2014), strongly demonstrate that redundancy allows paralogs to compensate for each other's LOF at the molecular level.

The buffering ability of paralogs is however not universal (Ihmels et al. 2007) and opposite results have been reported. For instance, Chen et al. (W.-H. Chen et al. 2013) reported an enrichment of human diseases among paralogous genes, particularly among the ones with higher functional similarity. The authors explained this result with a model in which redundancy reduces the efficacy of purifying selection, leading to the maintenance of disease alleles that could have lower penetrance, for instance through noise in gene expression. Other authors have shown that the retention of whole-genome duplicates could be biased towards genes that are more likely to bear autosomal-dominant deleterious mutations (Singh et al. 2012). In this case, the maintenance of paralogs would be associated with greater susceptibility to disease mutations, contrary to the robustness expected from gene redundancy. A better understanding of whether and how paralogs can compensate for each other's deleterious mutations therefore requires a better understanding of the mechanisms involved. This would improve our understanding of evolution and also accelerate the development of medical interventions because redundancy is often a major obstacle in this context (Lavi 2015).

The mechanisms by which paralogs compensate for each other's LOF mutations are for most cases not known in details (Pickett and Meeks-Wagner 1995; Diss et al. 2014) but likely involve active and passive mechanisms, from transcriptional to post-translational ones. For instance, it was shown for a small fraction of paralogous gene pairs that a member of a pair is upregulated by some feedback mechanism upon the deletion of the second copy (Kafri, Bar-Even, and Pilpel 2005). Although it has important consequences,

the occurrence of this phenomenon is however very likely limited. Indeed, a systematic assessment of this mechanism at the protein level in yeast found that it could take place only for a very small set of paralogous genes (DeLuna et al. 2010).

Another potential mechanism of compensation takes place at the level of protein-protein interactions (PPI) (reviewed by (Diss et al. 2014)), whereby paralogs replace each other with respect to their binding partners through ancestrally-preserved binding ability. Evidence for this mechanism was recently reported by (Diss et al. 2013, 2017). The model proposed is that paralogs appear to have different binding partners in wild-type cells because they mutually exclude each other from binding with potential partners. This is due to differential binding affinity or expression levels of the paralogs that favors binding competition towards one paralog or the other. Upon deletion, the mutual exclusion is relieved and compensation becomes apparent. Results consistent with this observation were obtained by Ori et al. (Ori et al. 2016) in mammalian cells. The authors showed that some paralogs can replace each other through changes in expression within protein complexes, supporting the fact that paralogs have preserved the ability to interact with the same partners. Another study reported observations consistent with this model using proteomics analyses of cancer cell lines (Gonçalves et al. 2017). In this case, an increased copy number for one gene led to increased protein abundance and a decrease in abundance of its paralogs, as if a feedback mechanism was affecting the balance between paralogs. This feedback is likely due to post-translational regulation that leads to the degradation of the displaced paralogs from protein complexes, also called protein attenuation (Ishikawa et al. 2017; Taggart and Li 2018). This observation suggests that the two paralogs would have overlapping binding partners and the balance would be determined by their relative affinity and abundance, as observed in one recent meta-analysis study (Sousa et al. 2019). Finally, Rajoo et al (Rajoo et al. 2018) examined the composition of the yeast nuclear pore complex and similarly to the Diss et al study (Diss et al. 2013), found that paralogous proteins can at least partially replace each other *in situ* upon deletion and change in abundance.

A major determinant that limit the ability of paralogs to compensate is their functional divergence, which can be approximated by sequence divergence (Hsiao and Vitkup 2008; J. Li, Yuan, and Zhang 2010). Other factors could also play a role, for instance cross-dependency, which has been brought to light only recently. Deluna et al. (DeLuna et al. 2010) looked at protein abundance of yeast paralogs when their sister copies are deleted, and found that six of the 29 pairs studied displayed negative responsiveness: upon deletion, the remaining paralog showed a decreased protein abundance. In half of these cases, the paralogs heteromerized (physically interacted with each other), suggesting that protein abundance may depend on their physical interactions. The control of protein abundance through interactions was also recently elucidated in the context of human cells (Sousa et al. 2019). The consequences of these decreases in abundance were not investigated further but one could imagine that this would directly affect the compensating ability of paralogs, because the deletion of one copy of a pair leads to a LOF of the second, thereby essentially acting as a dominant negative effect. A recent study by Diss et al. (Diss et al. 2017) directly examined paralog compensation at the level of protein-protein interactions. Among more than 50 pairs of paralogs, they showed that not all

paralogs could compensate in the yeast protein interaction network. About 20 pairs showed dependency, i.e. one paralog lost some or all its interaction partners upon the loss of the second. Diss et al. found that dependent pairs were enriched for pairs that form heteromers and in some cases, the dependency could be explained by a strong decrease in protein abundance upon deletion, consistent with the observation of DeLuna et al. (DeLuna et al. 2010).

Altogether, these observations raise the possibility that heteromerization of paralogs may reflect their physical and functional dependency, which as a consequence would reduce the ability of paralogous genes to compensate for each other's loss. One could therefore predict that the protection that paralogous genes provide against the effect of LOF mutations would be contingent on whether their products form heteromeric complexes with each other or not. These genes would have fitness effects that are closer to that of single copy genes (singletons) than that of typical duplicates. Here, we examine these predictions by re-analysing a set of well-curated pairs of human paralogous genes (Lan and Pritchard 2016; Singh, Arora, and Isambert 2015) and recent large-scale genome-wide CRISPR-Cas9 screens in which the effect of gene LOF on cell proliferation was examined in more than 450 cancer cell lines (Wang et al. 2015; DepMap 2018) and a primary cell line (Shifrut et al. 2018). The meta-analysis of the effect of gene LOF on cell proliferation, mRNA expression from 374 cell lines, protein expression from 49 cell lines and protein-protein interactions (Table EV1) revealed patterns which strongly support our hypothesis that paralogs that assemble are less protective, but through factors other than heteromerization itself.

Results

Paralogous genes protect against the effect of gene LOF across all cell lines

We used two datasets of paralogous genes, one of relatively young paralogs, derived from small scale duplication(Lan and Pritchard 2016) and another set of relatively old paralogs most likely derived from whole-genome duplication (Data ref: Ohnolog 2018)(total of 3132 pairs of paralogs, see Methods, Dataset EV1). We first examined whether paralogous genes protect against the deleterious effects of LOF mutations in a set of 455 human cell lines from three independent CRISPR-Cas9 genome-wide LOF screens (Table EV1). Such experiments yield a CRISPR-score (CS) per gene which is an estimate of the relative depletion of guide RNAs (gRNAs) during the genome-wide CRISPR-Cas9 screening experiment. CS therefore reflects the deleteriousness of LOF on cell proliferation (Fig EV1): a lower CS value indicates more deleteriousness and vice versa. These datasets are (1) CS1 from four cell lines (Wang et al. 2015), (2) CS2 from 450 cell lines (Meyers et al. 2017; DepMap 2018), (3) CS2.1 from 450 cell lines (DepMap 2018) and (4) CS3 from 1 primary cell line (Shifrut et al. 2018) (see Dataset EV2 for cell line information, Dataset EV3 for gene-wise CS values). All the CS values capture the essentiality of the genes which in the case of cancer cell lines, are found to be largely independent of the role of the genes in cancerogenesis (Fig EV1). Because the estimation of CS of the paralogs could be confounded by gRNAs that match to more than one gene due to their sequence similarities, we recomputed scores for the CS1, CS2.1 and CS3

datasets by considering only the gRNAs that uniquely align to the genome (see Methods). Dataset CS2 and dataset CS2.1 constitute data from the same set of cell lines (biologically identical), but analysed differently. CS2 takes copy-number variation effects in each cell line into account (used directly as computed by the authors) (Meyers et al. 2017), while CS2.1 analysed by utilizing only the uniquely aligned gRNAs (see methods). CS values among datasets CS1 and CS2/CS2.1 are well correlated, indicating reproducible measurements of fitness effects across platforms, methodologies, cell lines and cell types (Appendix Fig S1). The weaker correlation with dataset CS3 values (Spearman correlation coefficient ranges from 0.19 to 0.21), however could be attributed to the difference in the physiology of the primary and cancer cell lines itself, although technical factors could also be responsible.

As expected, we find that paralogs buffer the effect of gene LOF. Genes with paralogs have relatively higher CS values than singletons (see methods for classification of singletons), for the three biologically independent datasets considered (Fig 1A). To confirm that these effects were systematic and were not driven by few cases of cell lines with strong effects, we compared the mean CS for paralogs and singletons across cell lines (see Fig 1B for analysis with CS2.1 and Appendix Fig S2A for analysis with CS2 dataset). All cell lines systematically showed stronger buffering effects for the inactivation of paralogs compared to singletons, with no exception. The same results are observed for the comparison of paralogs with genes that are not in the set of paralogs nor classified as singletons, denoted as “unclassified” (see Fig 1C for analysis with CS2.1 and Appendix Fig S2B for CS2 dataset). These results are therefore highly reproducible and cell line independent. However, the trend showed some dependence on molecular features such as mRNA expression levels, as we discuss below.

Older paralogs tend to be less protective

In order to determine the effect of paralog age on deleteriousness, we compared the essential and nonessential sets of genes in terms of their age group of duplications retrieved from Ensembl Compara (Herrero et al. 2016)(see methods). We find that, albeit with a weak difference, older paralogs are more likely to be classified as essential genes and thus have potentially more deleterious effects upon LOF than younger paralogs (Fig 1D, see methods for the classification of essential genes). This result underscores similar findings from earlier studies showing that the more diverged paralogs are, the less likely they are to buffer each other’s loss, in the context of human diseases or yeast gene deletions (Hsiao and Vitkup 2008; J. Li, Yuan, and Zhang 2010; Plata and Vitkup 2014).

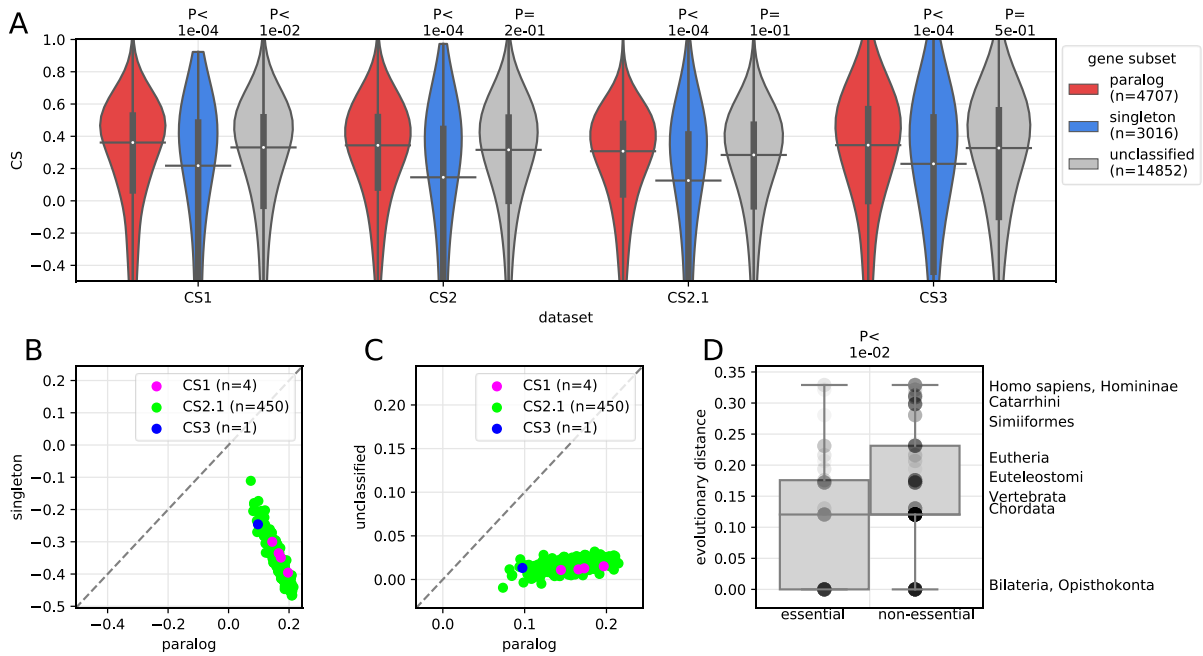


Fig 1: The LOF of paralogs is less deleterious than that of singletons in human cell lines.

A) LOF data derived from genome-wide CRISPR-Cas9 screening experiments. Deleteriousness of LOF of a gene on cell proliferation is estimated from the depletion of gRNAs during the experiment. The extent of depletion is measured as a CRISPR-score (CS, see Methods). CS values across cell lines from three biologically independent datasets — CS1 (Wang et al. 2015), CS2/CS2.1 (Meyers et al. 2017; DepMap 2018) and CS3 (Shifrut et al. 2018) are shown. Genes that are not in the paralog datasets but that were not identified as singletons in the stringent identification of singletons are denoted as "Unclassified". Relatively higher CS of paralogs compared to singletons indicate that they are relatively less deleteriousness. P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

B) Comparisons of CS values between paralogs and singletons and **C)** between paralogs and unclassified genes (neither clearly a paralog nor a singleton). CS data for 4 (CS1) + 450 (CS2.1) + 1 (CS3) cell lines is shown. Each point represents the mean CS for a class (singleton, paralog or unclassified) in an individual cell line. All points are below the diagonal (dashed gray line), showing that the effect is systematic and largely cell-line independent. Similar plots are shown for CS2 dataset in Appendix Fig S2.

D) Older paralogs tend to be more essential than younger one, therefore less protective (i.e. more deleterious upon LOF), than younger ones. On the y-axis, the age groups are ordered in increasing distance of phylogenetic node of duplication relative to common ancestor, i.e. Opisthokonta. Sets of essential and non-essential genes were derived from the union of gene sets reported by DepMap (DepMap 2018) and BAGEL (Hart and Moffat 2016) (See Methods). P-value from a two-sided Mann-Whitney U test is shown.

Heteromeric paralogs emerge from ancestral homomers

The model in which paralogous genes are dependent on each other considers that interacting paralogs derive from ancestral homomeric proteins (Kaltenegger and Ober 2015; Baker, Hanson-Smith, and Johnson 2013; Bridgham et al. 2008; Diss et al. 2017). We can assume that when the two paralogs individually form a homomer, the ancestral protein was most likely also a homomer. Therefore, we can infer that heteromers of paralogs are derived from ancestral homomers, if each paralog also forms a homomer.

We use two sources of PPI (Livstone et al. 2011; Chatr-Aryamontri et al. 2015) and IntAct (Orchard et al. 2014), to define homomeric genes or heteromeric gene pairs based on PPI (see Methods). Further, the subsets are defined based on all PPI (henceforth this dataset will be referred to as 'all PPI') or direct physical interactions only (henceforth this dataset will be referred to as 'direct PPI'). Homomers, in the context of this study, refer to the assembly of a protein with itself while heteromers refer to paralogous proteins that assemble with other proteins. Considering all PPIs (see methods for the difference between 'all PPI' and 'direct PPI'), paralogs are 8.13 times more likely to form heteromeric pairs (Fisher's exact test, P-value < 1.4e-14) if they also both form homomers than if none of them does. The likelihood is 48.88 times for heteromers defined by 'direct PPI's only (P-value < 5.5e-18) (see Appendix Table S1 for the numbers of pairs in each category). We can therefore generally assume that pairs of heteromers are more likely to derive from ancestral homomers, consistent with previous observations (Wagner 2003; Pereira-Leal et al. 2007).

Paralogs that form heteromers have stronger effects on cell proliferation when inactivated

Next, we investigated the effect of LOF of paralogs that form heteromers and those that do not. Consistent with the dependency hypothesis, the LOF of paralogous heteromers seem to cause relatively more deleterious effect on cell proliferation than the LOF of non-heteromeric paralogs, across in all 4 CS datasets (Fig 2A, similar analysis with 'direct PPI's is shown in Appendix Fig S3). We also observe that the effect is consistent across cell lines by looking at the mean CS of paralogous heteromers or non heteromers genes within each cell line (Fig 2B), with a majority of cell lines showing stronger effects for the LOF of paralogs forming heteromers. This trend is clearly observed across all the CS datasets and irrespective of the source of the PPI used for the definition of the heteromeric paralogs (similar analysis as that of the Fig 2B with all the rest of the combinations of the PPI sources and CS datasets is shown in Appendix Fig S4). A similar analysis with paralogs that are both heteromers and homomers compared with paralogs which are only homomers shows that interacting paralogous are relatively more deleterious (Fig 2C). This trend is also clearly observed across all the CS datasets and irrespective of the source of the PPI used for the definition of the subsets of paralogs (similar analysis as that of the Fig 2B with all the rest of the combinations of the PPI sources and CS datasets is shown in Appendix Fig S4). The effects are therefore not due to homomerization but due to heteromerization (Fig 2C). These results support the hypothesis that interacting pairs of paralogs are less likely to buffer each other's LOF.

One potential confounding factor with this analysis is the fact that the frequency of heteromers could covary with the age of paralogs, which we showed above to affect at least partially the essentiality of the gene (Fig 1D). Heteromers are indeed older than the non-heteromeric paralogs (Fig 2D), albeit only in the case of the heteromers defined by 'all PPI's. We therefore looked at CS values of paralog LOF corrected for age, by using age groups. We observed that for all age groups, except for two, CS values for heteromers are indeed lower than for non-heteromers, suggesting that this effect is largely independent from age (Fig 2E). The reason for inconsistency in case of two age groups is however unclear but it may be due to the DNA sequence divergence between paralogs and the ability of gRNA to target one gene specifically.

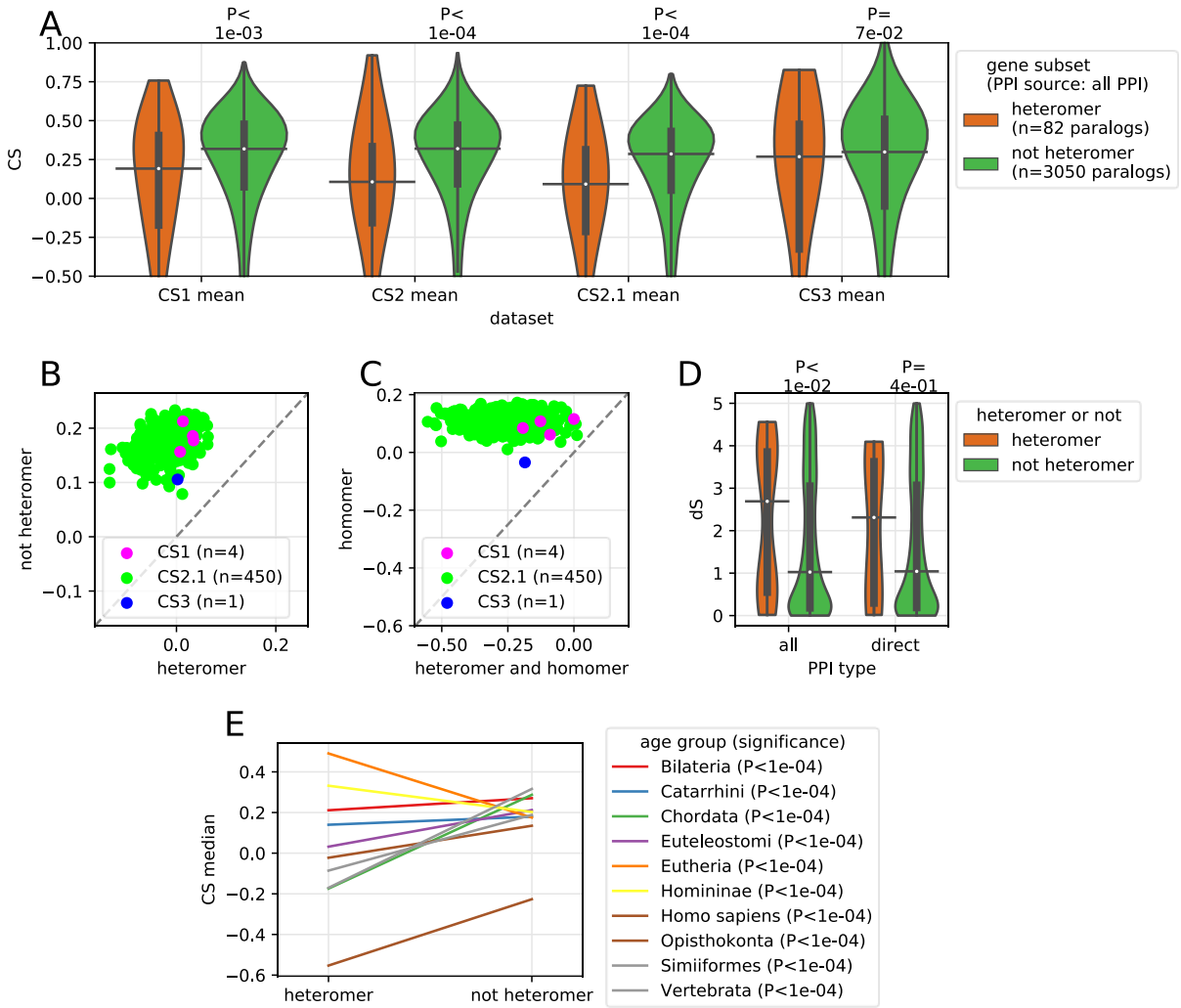


Fig 2. The LOF of paralogs that form heteromers is more deleterious than the LOF of non-heteromers.

A) The effect of LOF on cell proliferation (CS values) is relatively more deleterious for paralogous heteromers than non-heteromers, across all 4 CS datasets. P-values of two-sided Mann-Whitney U tests are shown. Similar plot for heteromers defined with direct PPI only is shown in Appendix Fig S3.

B) Mean CS values of paralogous heteromers and non-heteromers (defined by 'all PPI's from BioGRID source) are shown across cell lines. Each point represents the mean CS value for a class in an individual cell line. All the points are above the diagonal (dashed gray line), showing that the effect is systematic and largely independent of cell-line. Similar plots for both PPI sources and CS2 dataset are shown in Appendix Fig S4.

C) Similar to panel **B,** but comparing paralogs that form heteromers and homomers to those that form homomers only (defined by 'all PPI's from BioGRID source). This result shows that the difference between heteromers and non-heteromers is not caused by the fact that heteromers are also enriched for homomers. Similar plots for both PPI sources and CS2 dataset are shown in Appendix Fig S4.

D) Paralogs that form heteromers tend to have been duplicated earlier in evolution. The age of the paralog pairs is shown in terms of synonymous substitutions per site (dS) (see methods), a proxy for age. Data is shown for interactions derived from 'all PPI', and those that are more likely to detect 'direct PPI'.

E) Paralogs that form heteromers tend to be more deleterious upon LOF than other paralogs. Data from CS2.1 is shown, largely independent of the age of the paralog. In the legends, paralogs are ordered by their age. The CS values per class of paralogs (heteromer or not) and their age group are aggregated by taking median across cell lines. Note that while heteromers are more deleterious in most of the age groups, in the case of 2 out of 10 age groups a reverse trend is observed. Distributions of the CS values per class of paralogs (heteromer or not) and their age group for this analysis are shown in Appendix Fig S5A. Similar analysis with dataset CS2 and for heteromers detected with 'direct PPI's only is shown in Appendix Fig S5 panels B to D.

On the violin plots (panel **A** and **D**), the medians of the distributions are denoted by a horizontal black line, while the quartiles of the distributions from the median value are indicated by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown in panel A.

Molecular functions enriched for heteromeric paralogs tend to be more critical for cell proliferation

It is possible that the effects detected are due to specific gene functions that would be particularly associated with heteromeric paralogs. We first examined whether heteromers of paralogs are enriched for particular function among all paralogs. We found that heteromers of paralogs are enriched for gene sets containing proteins that have catalytic activity and known to directly interact/regulate with each other such as kinase binding (Breitkreutz et al. 2010)as well as DNA binding proteins from histone deacetylase binding gene set (see Dataset EV4 gene sets and GO terms used in the analysis, Dataset EV5 for enrichment analysis).

From this gene set enrichment analysis, we find that the proportion of heteromerization of paralogs in a gene set, in general, is negatively correlated (Spearman correlation coefficient=-0.26, P-value=0.086) with the average CS value of paralogs per gene set. This shows that the stronger effects of LOF for heteromeric paralogs is caused at least partially by heteromerization even when controlling for molecular functions. This is the case in the analysis while considering both PPI methods (see Fig 3 for 'all PPI' and Appendix Fig S6A for 'direct PPI'). The negative correlations also hold true in the case of GO biological process and GO cell component gene sets (see Fig EV2 for 'all PPI' and Appendix Fig S6B and C for 'direct PPI').

Some molecular functions enriched among heteromers also show a significant difference in the average CS values of the heteromers and non-heteromers in that particular gene set (Fig 3, Dataset EV5). Such gene sets include, for instance, RNA polymerase II and transcription and DNA and nucleic acids binding genes, which are frequent among paralogs known to be in large families of dimeric transcription factors that evolve through duplication (Amoutzias et al. 2008) and that has been shown to have co-dependent

evolution (Baker, Hanson-Smith, and Johnson 2013). Among other such gene sets, phophatase activity related genes were identified in the case of ‘all PPI’ dataset (Fig 3) are also. Gene sets corresponding to protein homodimerisation activity was commonly found in both the analyses with all PPI and with direct PPI (Fig 3 and Appendix Fig S6A) as showing significant lower CS values for heteromeric pairs, consistent with the correlation observed above between heteromerization and homomerization.

In terms of biological processes, protein dephosphorylation process related genes, regulation of cell proliferation, apoptotic, transcription process, and cell junction assembly process are enriched among heteromers and also show significant deleteriousness (i.e. depletion in CS values of the heteromers, see Fig EV2A for analysis with ‘all PPI’, Appendix Fig S6B for analysis with ‘direct PPI’). In terms of cellular components, essential genes related actin cytoskeleton and chromatin are enriched among heteromers and are significantly more deleterious (see Fig EV2B for analysis with ‘all PPI’ and Appendix Fig S6C for analysis with ‘direct PPI’). These genes are therefore interesting candidates for future functional analysis on the consequences of heteromerization for protein function and robustness.

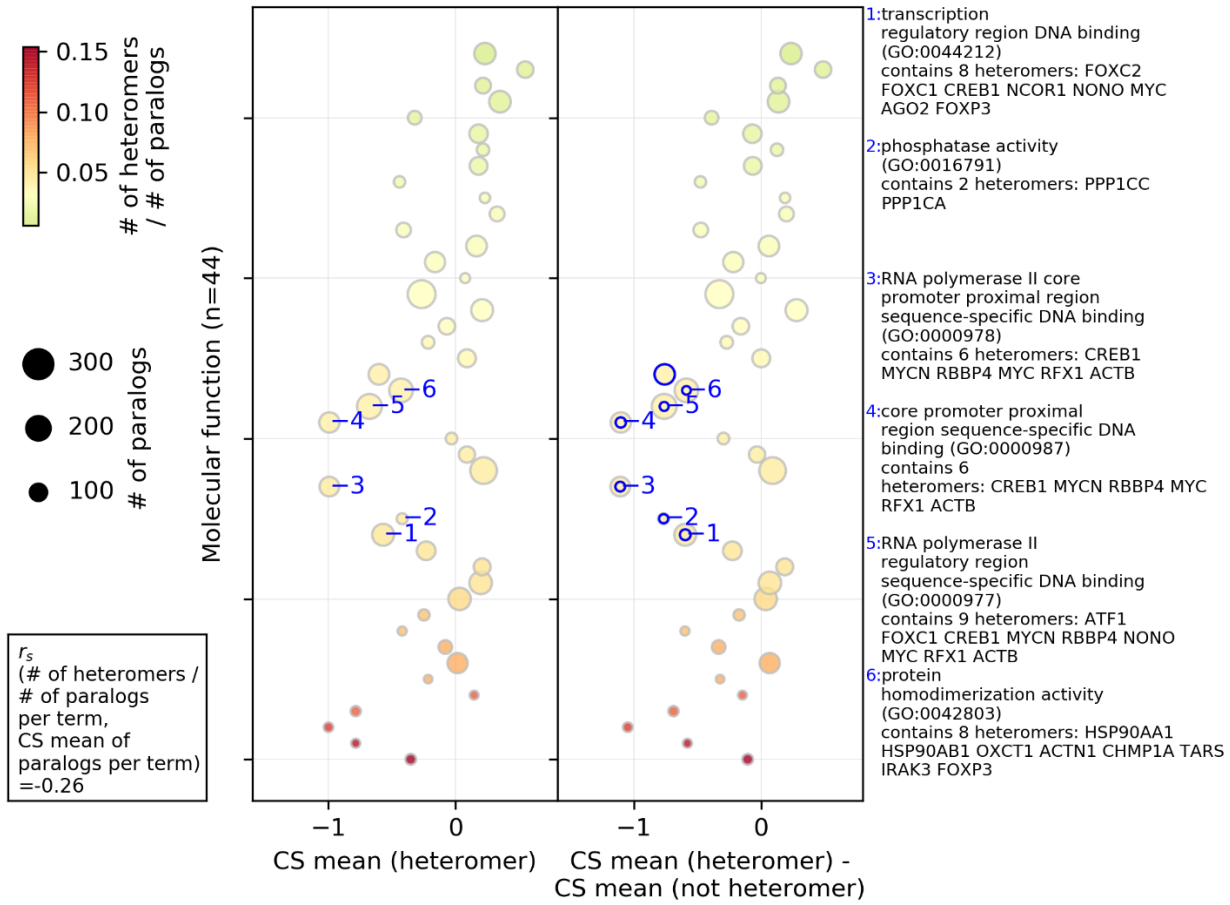


Fig 3. Association between the molecular functions of paralogs, their probability of heteromerization and the effect of gene LOF on cell proliferation.
Average CS values (x-axis) of paralogs (heteromer or not heteromer) belonging to a gene set were used in the analysis. GO molecular functions are sorted according to their proportion of heteromeric paralogs (i.e. # of heteromers/ # of paralogs, heteromers defined by 'all PPI'). The size of the circles represent the number of paralog pairs in a category and the colors represent the proportion of heteromers in that category. The difference between the average CS value of the heteromers and average CS value of the non-heteromers is plotted in the right panel. The terms with significant difference between the average CS value of the heteromers and average CS value of the non-heteromers (estimated by two-sided t-test) are annotated with the blue edges. Descriptions of the representative significant GO terms with the highest difference are shown in the right side-panel. Spearman rank correlation between the proportion of the heteromers in the GO terms and the average CS value of paralogs in the term (r_s(# of heteromers / # of paralogs per term, CS mean of paralogs per term)) is shown in left right corner. Only GO molecular functions with more than 10% of the number of paralogs in all the gene sets are shown. Similar analysis for the GO biological process and GO cellular component aspect, for the 'all PPI' based data are shown in Fig EV2. Similar analysis with the 'direct PPI' data is shown in Appendix Fig S6. See Dataset EV5 for GO terms and annotations shown on this

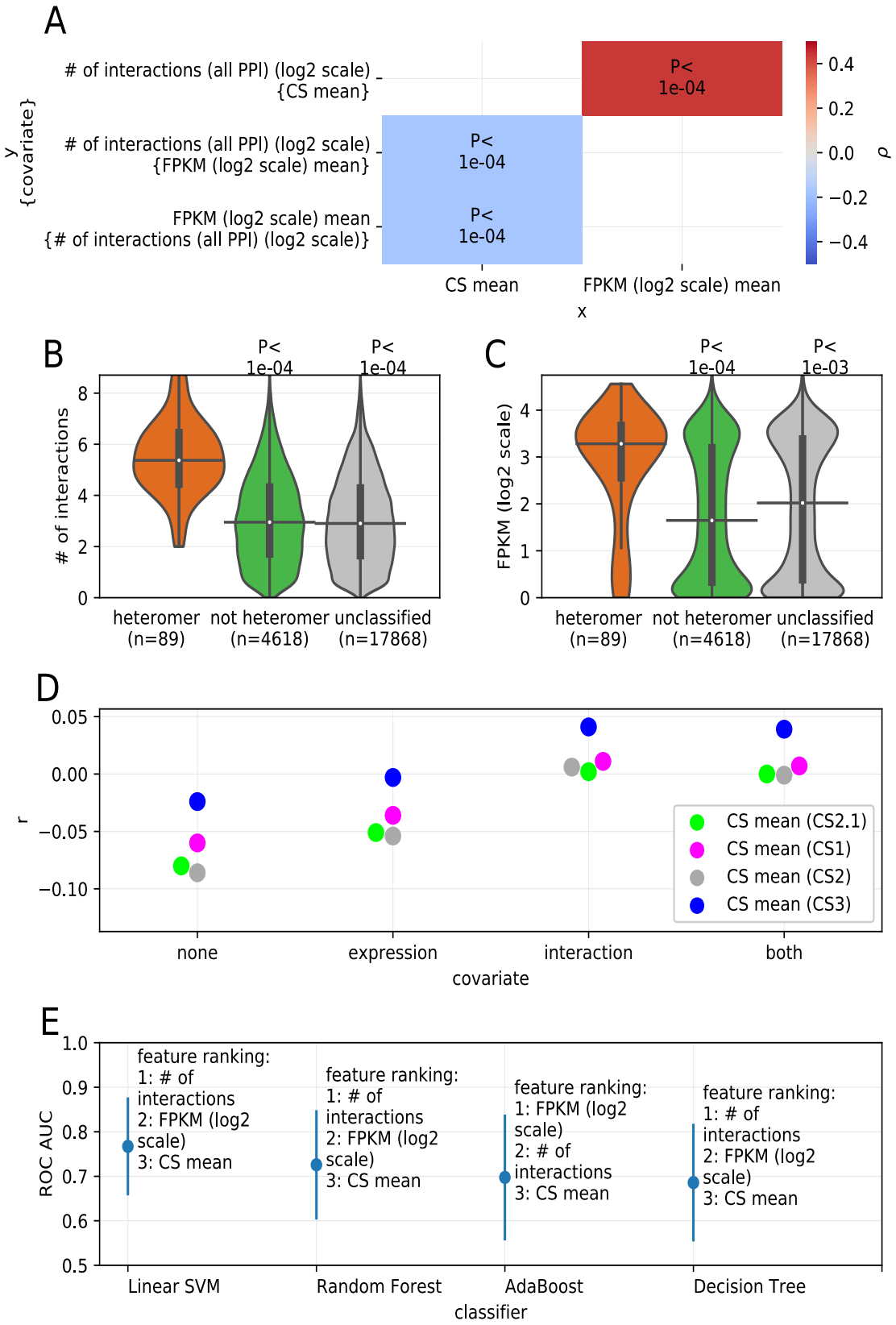
figure. Note that not all gene sets are independent because some genes are in several categories.

Heteromeric paralogs are more highly expressed and have more protein interaction partners

Previous reports have shown that homomeric proteins tend to have a larger number of interaction partners (Ispolatov 2005). If heteromers of paralogs inherit their interactions from the homomeric ancestor, they could also have a larger number of interaction partners, which could explain their relatively lower CS values (Fig 2A). We first find that there is a significant negative correlation between the average CS values in the three biologically independent datasets and the number of PPI partners and its mRNA expression level (measured in terms of Fragments Per Kilobase of transcript per Million mapped reads i,e, FPKM) (Fig 4A, see Appendix Fig S7 for analysis with each CS dataset). Considering the interdependence between the three related factors, partial correlation between each two factors was calculated while controlling for the third factor. As hypothesized, we find that heteromers of paralogs have a larger number of PPI partners, both considering 'all PPI' (Fig 4B) and 'direct PPI' only (Appendix Fig S8A) are more highly expressed (see Fig 4C for analysis with the heteromers defined by 'all PPI' and Appendix Fig S8B for the ones defined by 'direct PPI's). The trend with mRNA expression is also true in the case of most of the cell lines (see Appendix Fig S8C for analysis with 'all PPI' and Appendix Fig S8D for analysis with 'direct PPI' only). Collectively, the number of PPI and mRNA expression seem to explain the greater deleteriousness of the heteromers compared to non-heteromeric paralogs.

Further, we therefore tested which molecular feature among mRNA expression and number of interaction partners act as a stronger predictor of the paralog being heteromer or not. Considering that the molecular features are interdependent, we relied on joint modelling using partial correlations and ensemble modelling approach to estimate the predictiveness of the molecular features. The partial correlations were carried out between the status of the paralog being either heteromer or not (binary variable) and CS values (i.e. deleteriousness of the paralog), while controlling for either mRNA expression or the number of protein interaction partners. From this analysis, the number of protein interaction partners emerges as a better predictor of the deleteriousness of the heteromers than mRNA expression (Fig 4D, for analysis with 'direct PPI's see Appendix Fig S8E), because controlling for the former diminishes the correlation more than controlling for the later. Secondly, we used a set of 4 ensembl machine learning models to predict the status of the paralog, i.e. heteromer or not, using the three datasets (i.e. CS values, mRNA expression and the number of protein interaction partners) (see Methods). From the feature importance obtained from the classification models, it is again apparent that the number of interactions of a protein is a better predictor of the status of the paralog (Fig 4E, for analysis with individual CS datasets, see Appendix Fig S9), and thereby their relative deleteriousness. These two analyses thus confirm that indirect effects owing mostly to the number of PPI partners, and to mRNA expression to a lesser extent, seem to explain the stronger impact of LOF on heteromers.

As an addendum, in general, the protective effect of paralogs compared to singleton is partially caused by their lower expression and smaller number of protein interaction partners than that of the singletons (Fig EV3A and B, cell-line wise comparison in Fig EV3C). Partial correlation analysis (Fig EV3D) indicates that the protective effect of the paralogs is more attributable to the expression of the genes but this does not completely explain the results. In the case of all the CS datasets, at lower expression values, the difference in CS between paralogs and singletons is non-significant (Fig EV3E). The reason for this could be attributed to the small counts of mRNA expression generally being relatively more noisy. Also, sequence similarity between the paralogs leads to removing reads from the analysis and may thus act as one of the confounding factors in this analysis (see methods) by underestimating mRNA expression of paralogs.


A
# of interactions (all PPI) (log2 scale) {CS mean}
# of interactions (all PPI) (log2 scale) {FPKM (log2 scale) mean}
FPKM (log2 scale) mean {# of interactions (all PPI) (log2 scale)}
y {covariate}
P< 1e-04
P< 1e-04
P< 1e-04
CS mean
FPKM (log2 scale) mean
x
ρ
B
# of interactions
P< 1e-04
P< 1e-04
heteromer (n=89)
not heteromer (n=4618)
unclassified (n=17868)
C
FPKM (log2 scale)
P< 1e-04
P< 1e-03
heteromer (n=89)
not heteromer (n=4618)
unclassified (n=17868)
D
r
CS mean (CS2.1)
CS mean (CS1)
CS mean (CS2)
CS mean (CS3)
none
expression
interaction
both
covariate
E
ROC AUC
feature ranking: 1: # of interactions 2: FPKM (log2 scale) 3: CS mean
feature ranking: 1: # of interactions 2: FPKM (log2 scale) 3: CS mean
feature ranking: 1: FPKM (log2 scale) 2: # of interactions 3: CS mean
feature ranking: 1: # of interactions 2: FPKM (log2 scale) 3: CS mean
Linear SVM
Random Forest
AdaBoost
Decision Tree
classifier

1

Fig 4: Relationship between the effect of LOF of a gene on cell proliferation, mRNA expression and number of protein-protein interaction partners.

A) The effect of gene LOF on cell proliferation as measured in terms of CS values is correlated with mRNA expression and number of protein-protein interaction partners. Partial correlations were estimated in terms of Spearman correlation coefficients (ρ) between each pair of factors while controlling for the third factor (covariate, indicated in the curly brackets). The associated P-values are denoted on the heatmap. Average CS values across CS datasets was used. See Appendix Fig S7 for correlations in case of individual CS datasets and direct PPI.

B) Paralogs that form heteromers have more interacting partners compared to non-heteromers. Number of interactions are in log2 scale. Similar plot with paralogous heteromers detected with only direct physical interactions is shown in Appendix Fig S8A.

C) Paralogs that form heteromers show higher expression than non-heteromers. Similar plot with heteromers of paralogs detected with only direct PPI is shown in Appendix Fig S8B. Cell-line wise comparisons with heteromers defined by 'all PPI' and 'direct PPI' is shown in Appendix Fig S8C and Appendix Fig S8D respectively.

Contribution of the interacting factors in determining the paralog status is determined by jointly modeling through two approaches: partial correlations (panel **D**) and ensemble classification models (panel **E**).

D) Partial Spearman correlation coefficients (r, shown on the y axis), between CS values and a paralog status (heteromer or not, binary variable, 1 : heteromer, 0 : not heteromer). The correlations were determined while controlling for none of mRNA expression and number of interactions ("none"), only mRNA expression ("expression"), only number of interactions ("interaction") or both ("both") (as shown on the x axis). Controlling for the number of interactions leads to the greater loss of negative correlation, indicating that it contributes to the correlation more than mRNA expression. Similar analysis with heteromers defined by 'direct PPI' is shown in Appendix Fig S8E.

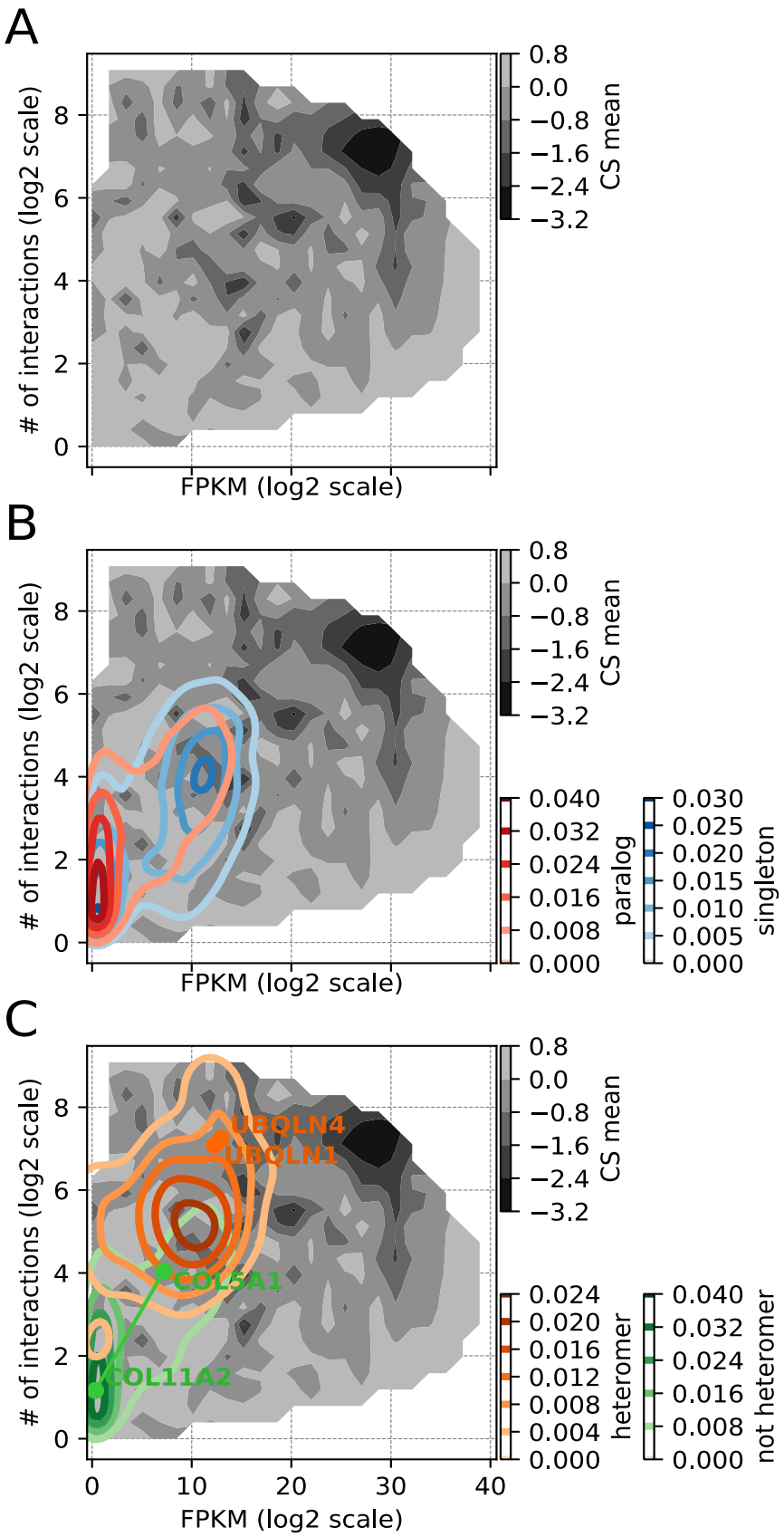
E) Feature importance (shown on the y axis) of the three factors as determined through four different ensemble classification models (shown on the x axis). Mean and standard deviation of the ROC AUC values across cross validations and bootstraping runs (see methods) is plotted for each of the 4 classifiers. The CS values used for this analysis are mean of the CS values across all the CS datasets. For similar analysis with the 4 individual CS datasets, see Appendix Fig S9.

In panels **B** and **C,** P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

Heteromerizing paralogs occupy a space of the robustness landscape where gene LOF is more deleterious

Overall, these results can be summarized in a robustness landscape in which the robustness against LOF is shown as a function of the number of PPIs and mRNA expression level. Following the pattern of correlations between the three factors (as

shown in Fig 4A), the landscape clearly shows that strong deleteriousness is localized in the upper right corner, where expression as well as number of interaction partners are high (Fig 5A for analysis with ‘all PPI’, see Fig EV4A for analysis with ‘direct PPI’ only). The overlay on this landscape of the density of singletons and paralogs shows that they occupy a similar space (Fig 5B analysis with ‘all PPI’, see Fig EV4B for analysis with ‘direct PPI’ only), alhought singletons are in a space in which genes are slightly more expressed and have a slightly larger number of interactions. Paralogous genes that form heteromers clearly occupy a space that is distinct from non-heteromeric ones, which brings them closer to a more deleterious parameter space in which genes are relatively more expressed and have more protein interaction partners (Fig 5C for analysis with ‘all PPI’, see Fig EV4C for analysis with ‘direct PPI’ only). We also show representative pairs of heteromeric and non-heteromeric pairs of paralogs on the map (Fig 5C and Fig EV4C). For instance, Ubiqulin 1 and 4 (UBQLN1 and UBQLN4) form a heteromer, are highly expressed, have a large number of protein interaction partners and their LOF is highly deleterious, as seen by their location nearing the valley of the fitness landscape. On the other hand, a non-heteromeric pair, collagen type V alpha 1 chain (COL5A1) and collagen type XI alpha 2 chain (COL11A2) have lower mRNA expression and lower number of PPI partners, setting position at the peak of the landscape with relatively non-deleterious CS values. In terms of network features, paralogous pairs that heterodimerize are more similar to singletons and have correspondingly similar effects on proliferation when inactivated.


A
B
C
# of interactions (log2 scale)
FPKM (log2 scale)
CS mean
0.8
0.0
−0.8
−1.6
−2.4
−3.2
paralog
singleton
heteromer
not heteromer
UBQLN4
UBQLN1
COL5A1
COL11A2

1

Fig 5: Robustness landscape visualization showing regions of deleteriousness to LOF of a function of mRNA expression and number of interaction partners.
mRNA expression (lined on x axis) and number of PPI partners (y axis) are strong determinants of the deleteriousness of gene LOF (measured in terms of average CS across CS datasets, shown on z axis).

A) The landscape shows the effect of LOF of genes on cell proliferation (CS) as a function of the two parameters. Regime with high gene expression levels and large number of interactions clearly shows more relatively lower CS values, indicating deleteriousness upon LOF.

B) Kernel density estimates for paralogs and singletons are overlaid on the landscape to indicate their level of occupancy. The density of paralogs is located towards lower expression levels and small numbers of protein interaction partners, compared to singletons.

C) Similar to B, kernel densities of paralogous heteromers and paralogous non-heteromers are overlaid on the landscape. The location of heteromers is biased towards higher expression levels and larger number of protein interaction partners, compared to non-heteromers. Also, locations of representative heteromeric (UBQLN1 and UBQLN4) and non-heteromeric pairs (COL5A1 and COL11A2) is annotated on the landscape.

Similar plots with direct PPIs only are shown in Fig EV4.

Potential consequences of the heterodimerization of paralogs

We examined the results from the meta-analysis further to explore other potential features leading to the association between the heteromerization of paralogs and their deleteriousness upon LOF. Given that gene expression level appears to be one of the determinants of the fitness effect of LOF, it is possible that mechanistically, the ability of a paralog to buffer for the loss of its sister copy depends on their relative abundance. For instance, if highly asymmetrically expressed, the LOF of the most expressed gene of a pair is unlikely to be buffered by the least expressed one. However, if both are expressed at a comparable level, both would be expected to affect cell proliferation in a comparable manner. In addition, Diss et al. (Diss et al. 2017) showed that physical dependency between interacting paralogs is often asymmetrical, the lowly expressed copy being affected by the deletion of the highly expressed one more than in the reciprocal condition. They suggested that this asymmetry could derive from the fact that the lowly expressed one may be post translationally stabilized by the most expressed one. We therefore tested whether asymmetry in mRNA expression influences which gene in a pair is more deleterious upon LOF (Fig 6A).

We found that across cell lines, the paralog of a pair with the highest level of expression (P1) shows significantly lower CS values than the lowly expressed one (P2) (Fig 6B for analysis with heteromers defined by ‘all PPI’, see Appendix Fig S11 for similar analysis with heteromers defined by ‘direct PPI’). This means that the loss of a paralog is more deleterious when it is the most expressed one in a pair. Properties that determine the deleteriousness of LOF across genes therefore also apply within pairs of paralogous genes. The fraction of pairs across cell lines for which the LOF effect is greater than the

other paralog in the pair is strongly associated with their asymmetry of expression (Fig EV5A for analysis with CS2.1 dataset, see Appendix Fig S12 for similar analysis with CS2 dataset). Moreover, investigating the relationship between the difference of average CS of the paralog pair and asymmetry of mRNA expression levels, we find that these two factors are more negatively correlated in case of heteromers than non-heteromers (see Fig 6C for analysis with 'all PPi' and Appendix Fig S13A for analysis with 'direct PPI'. See Fig EV5B and Appendix Fig S13B for distributions of correlation coefficients in case of CS2.1 and CS2 datasets respectively), although the difference is statistically significant only in the case of heteromers defined by 'all PPI' and CS2.1 dataset. These correlations indicate that if the heteromeric pair of paralogs are asymmetrically expressed, then the difference in deleteriousness is larger upon LOF than for non heteromeric ones. This suggests that the lowly expressed gene of a pair is less able to buffer the loss of the highly expressed one in case of heteromers than non heteromers. This enhanced sensitivity appears to be counterbalanced at the systems level by the fact that heteromeric paralogs are more likely to have symmetrical expression (Fig 6D). Altogether, this suggests that symmetrical expression imposes a stricter contingency for heteromers than non-heteromers, arguably due to the need for their stoichiometric balance in their physical assembly. Comparing the mRNA expression of the heteromers across 374 cell lines (Fig EV5C) and protein expression across 49 cell lines (Fig EV5D), it appears that the heteromers are indeed on average more dosage balanced than non-heteromers. This trend is observed in both the comparisons and data from either PPI source, although it is statistically significant only in the case of comparison of mRNA expression for heteromers defined by 'all PPI'. This is potentially because it covers more pairs or because those include more pairs in large complexes, which are submitted to dosage balance constraints (Papp, Pál, and Hurst 2003; Teichmann and Veitia 2004), and because the proteomics data covers a smaller number of gene pairs and cell lines.

Finally, we examined whether the enhanced deleteriousness of LOF for heteromers could be due to their physical dependency, which would be manifested as the alteration of one member of a pair when the other member is absent as previously observed by Diss et al (Diss et al. 2014; Pickett and Meeks-Wagner 1995). This could offer a mechanistic insight into some of our observations. It is difficult to predict the physical dependency of paralogs but one could hypothesize that it is more likely to occur for strongly interacting pairs. We therefore used the size of the interaction interface of heteromers was a proxy for the strength of interaction (as in the case of Sousa et al. study (Sousa et al. 2019), see methods). Using the data for 25 heteromers of paralogs, we indeed observed a marginally significant negative correlation with the average CS values (Fig 6E) of paralog pairs with the strength of interactions, suggesting that codependency indeed could be a mechanism that contributes to the enhanced deleteriousness of paralogs pairs that interact with each other.

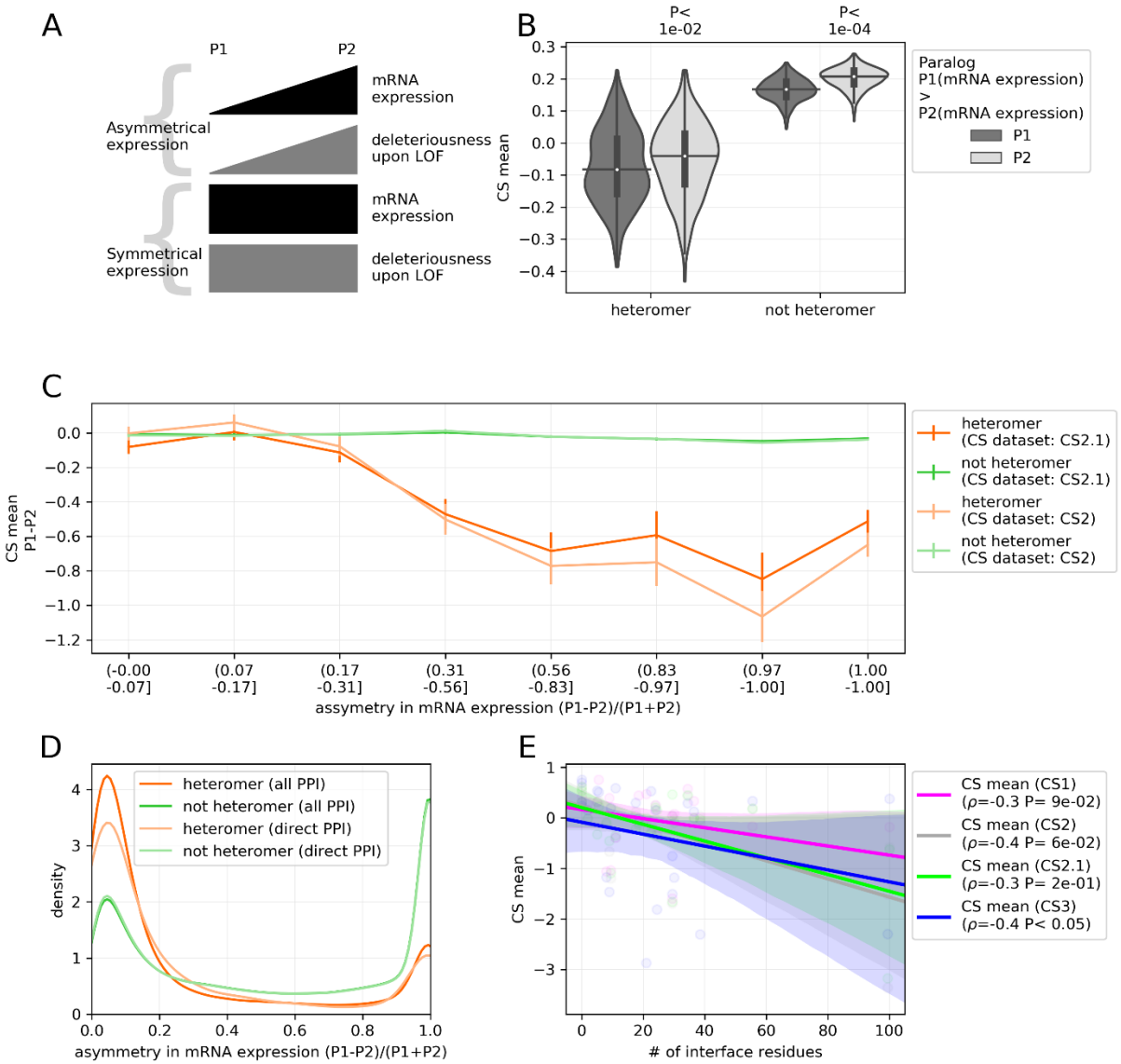


Fig 6: Asymmetric expression of paralogs and mechanistic insights into the relatively greater deleteriousness of the heteromeric paralogs.

A) Schematic representing likely scenarios pertaining to the relationship between the asymmetry in mRNA expression of a pair of paralogs (P1 and P2) and their relative deleteriousness upon LOF, as discussed in the text.

B) The most expressed paralog (P1) of a pair is more likely to be deleterious than the least expressed (P2), across 374 cell lines. Each point represents CS value of an individual cell line. P-value is from two-sided Mann-Whitney U test. On the violin plots, the medians of the distributions are denoted by a horizontal black line, quartiles of the distributions from the medians are indicated by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown. Heteromers in this analysis are defined from the 'all PPI's. For similar analysis with 'direct PPI's, see Appendix Fig S11.

C) Relationship between the difference in CS of the paralog pair (P1-P2) and the asymmetry of mRNA expression levels i.e. (P1-P2)/(P1+P2), where mRNA expression of P1 is higher than P2. Values near 0 are cases in which the mRNA expression is symmetrical and asymmetrical for values near 1. The heteromers are defined by 'all PPI'. Similar analysis with heteromers defined by 'direct PPI' is shown in Appendix Fig S13A. The relationship between the two factors in case of representative pairs of heteromeric and non-heteromeric paralogs is shown in Appendix Fig S14. Comparison of distributions of the correlation scores between heteromers and non-heteromers is shown in Fig EV5B.

D) Heteromeric paralogs tend to have more symmetric mRNA expression as compared to non-heteromers. Distribution of the asymmetry in the mRNA expression i.e. (P1-P2)/(P1+P2), where mRNA expression of P1 is higher than P2. Values near 0 are cases in which the mRNA expression is symmetrical and asymmetrical for values near 1.

E)The deleteriousness of the heteromers upon LOF (lined on the y axis) is negatively correlated with the number of residues at the interaction interface (x axis). ρ is Spearman's correlation coefficient. P-value associated with the Spearman's correlation coefficient is shown in the legend. Structures of representative heteromers are shown in Appendix Fig S15.

Discussion

The contribution of gene duplicates to cellular robustness has been established for several individual genes prior to the era of large-scale screening (Melton 1994; Gibson and Spring 1998; Thomas 1993; Pickett and Meeks-Wagner 1995). It was well established for model organisms such as yeast for which systematic gene deletion experiments have been performed (Zhenglong Gu et al. 2003). Systematically investigating the extent of the contribution of gene duplication to cellular robustness in the context of human cells was only recently made possible owing to large-scale CRISPR-cas9 screening (Wang et al. 2015). Here, using 3 biologically independent datasets of gene LOF that represent a large number of diverse cell lines and different experimental approaches (Table EV1), we find that paralogs systematically contribute to cellular robustness across all cell lines (Fig 1A).

Although the signal for the contribution of gene duplicates to robustness is significant and reproducible across datasets, some factors could limit the effects measured. The type of the cell lines used, i.e. cancer cell lines (in case of CS1, CS2/2.1) and primary (in case of CS3), clearly shows a difference in terms of the correlations (Appendix Fig S1). A second factor that is particular to the LOF screens in mammalian cell lines is the robustness of these cells to gene LOF. As seen from the distributions of the CS values across CS datasets, most of the genes are robust to LOF (Fig EV1). The effective range of deleteriousness is thereby very narrow, allowing the assessment of deleteriousness on a relative basis, rather than dependent on the absolute scores. Another limiting factors is that not all of the paralogs could be present as pairs in all cell lines, in particular in cancer cell lines that may have had additional duplications and deletions, which may have altered the copy number of paralogous genes. Although this effect may have biased our analyses, the use of the dataset CS2, which has been corrected for copy number variation across the cell line genomes (Meyers et al. 2017), shows that the results are likely robust to these effects. Another factor that may affect the results is that guide RNAs could potentially inactivate both paralogs of a pair, thereby leading to double gene LOF rather than a single one (Fortin et al. 2019). For instance, the CS1 original dataset (Wang et al. 2015) contained guides that targeted multiple positions in the genomes, many of which could be positions that correspond to duplicated genes. This could lead to double-gene-LOF by Cas9 cutting and DNA repair but could also lead to chromosomal rearrangements, leading to even stronger effects (Després et al. 2018; Kosicki, Tomberg, and Bradley 2018) than double gene LOF (Fortin et al. 2019). For this reason, we re-analyzed all data and considered only uniquely aligned gene-specific gRNAs (see methods). Nevertheless, it is not clear how many mismatches could be tolerated for efficient mutagenesis by Cas9 activity to occur, so eliminating all gRNAs that could lead to more than one gene LOF remains a difficult task.

We examined whether specific features of paralogous genes could affect their ability to buffer each other's LOF effect. We focused on their heteromerization because recent reports have shown that paralogous proteins often physically associate and that these physical associations could reduce their ability to buffer each other's LOF (DeLuna et al. 2010; Diss et al. 2017). This observation led to the prediction that due to their physical

and thus potential functional dependency, paralogs that form heteromers could contribute less to cellular robustness than non-heteromers, essentially behaving like singletons. We found that these paralogs indeed lead on average to larger effects on cell proliferation when inactivated by LOF mutations (Fig 2A). However, this is largely explained by larger number of protein interaction partners and higher expression levels for this class of paralogs (Fig 4). On the robustness landscape outlined by the two factors (Fig 5A), expression levels and number of protein interaction partners clearly separates genes based on their deleteriousness. It also helps in understanding the major determinants of buffering effect of paralogs in general (Fig 5B) and the greater deleteriousness of heteromers (Fig 5C).

One limitation of our analysis is the use physical interactions between paralogs as a proxy for dependency. Indeed, physical interactions may not be necessary nor sufficient for paralogs to be dependent (Kaltenegger and Ober 2015). It is possible that dependency concerns only obligate heterocomplexes, which are difficult to distinguish in large-scale data. However, our analyses using protein interaction interface size as a proxy for interaction strength suggest that this could be a potential mechanism. Dependency between paralogs could also evolve by other means than physical interactions. Additionally, it is also difficult to determine from large-scale proteomics data if two paralogs are part of the same complex simultaneously or if they occupy the same position but switch according to cellular compartments of expression timing (Ori et al. 2016). In this latter case, it would be unlikely that paralogs are dependent on each other, although the proteomics data would inaccurately suggest that they physically interact by being in the same complex. So far, there is also an issue of the sparseness of the known interactome. Especially, we still lack evidence for direct physical interactions (i.e. direct PPI) for most cases. This eventually obscures analysis because of lack of statistical power (as in the case of Appendix Fig S5).

Our results show that the association between paralog heteromerization and strong fitness effects is largely if not completely driven by the fact that it is also associated with expression levels and number of protein interaction partners. This is in line with the observation made by (Wang et al. 2015) who showed that essential genes tend to be more expressed and have more protein interaction partners. Recent observations supporting this trend were made by showing that LOF variants are rarer in humans for proteins with large number of protein interaction partners (Karczewski et al. 2019). Here we observe a similar result and identified that such features are enriched among paralogs that form heteromers. It is therefore difficult to determine if heteromerization indeed prevents buffering directly because of cross-dependency, or if all of the effects measured are caused by abundance itself. Our analysis showed that heteromeric paralogs have a tendency to be often associated with particular molecular functions (Fig 3) and these functions appear to lead to stronger effects on cell proliferation when inactivated. Heteromers of paralogs could therefore also have a lower buffering capacity overall because they associate with specific functions, including for instance transcription factors and protein kinases.

The capacity of paralogs to buffer each other's LOF also likely depends on their mechanisms of maintenance, which can be subfunctionalization and neofunctionalization (Force et al. 1999; Lynch and Force 2000; Lynch et al. 2001; Innan and Kondrashov 2010). Which one applies here for each paralog pair is difficult to determine without knowledge of the ancestral functions of the genes prior to duplication. Paralogs could fall into three categories. First, the duplication could be mostly neutral and has not been maintained by natural selection. Under this scenario, and in the absence of other changes, gene duplicates should be able to compensate each other's loss as long as they persist. Their function should not depend on each other's. The second possibility is that one copy or the other or both have neofunctionalized (reviewed in (Innan and Kondrashov 2010)). In this case, the novel function acquired by a paralog could not be compensated by the other copy but all ancestral functions could. Dependency in this case could arise from the acquisition of new functions by both paralogs at the same time. This has been seen for instance by Boncoeur et al (Boncoeur et al. 2012) who showed that the drug-pumping specificity of some ABC transporters are specific to heterodimers of paralogs and cannot be performed by individual homodimers. Buffering of this function would be possible by neither paralogs. For transcription factors, the neofunctionalization of one copy could be to become a repressor of the second copy, essentially given the heteromer a new functionality (Bridgham et al. 2008). The heteromer would now have a new regulatory mode that depends on the presence of both paralogs.

The final scenario is the most supported for the maintenance of paralogs and involve the accumulation of complementary degenerative mutations that lead to subfunctionalization (Force et al. 1999; Lynch et al. 2001). In this case, paralogs are maintained but without a net gain of function. This degeneracy would prevent compensation for the functions that have been lost in one or the other paralog. However, any function that did not subfunctionalize could be compensated for the second paralog upon the deletion of the first one. One way subfunctionalization could lead to dependency would be by complementary degenerative mutations (Kaltenegger and Ober 2015) that maintain the heterodimer but lead to the loss of the homomeric ones when proteins need to act as multimers (Pereira-Leal et al. 2007). In this case, the heteromeric form could replace the homomeric ones while making the presence of both paralogs necessary and preventing their mutual compensation. Dependency is therefore compatible with both modes of paralog maintenance but how frequent it is in each case remains to be examined and may require detailed functional characterization of paralog pairs.

We found that the relative expression level of paralogs is significantly associated with which one would be the most deleterious upon LOF (Fig 6A and B), revealing that buffering capacity is dependent on relative expression levels. Consistent with our observation, Barshir et al. 2018 (Barshir et al. 2018) recently showed that diseases that are tissue specific and that affect paralogous genes tend to affect tissues in which the second copy of a pair is generally lowly expressed, reducing its buffering capacity. These observations and ours have important consequences regarding the buffering effects of paralogs and their evolution. Qian and Zhang (Qian and Zhang 2008) and Gout and Lynch (Jean-Francois Gout and Lynch 2015; Jean-François Gout et al. 2010) showed that expression levels alone could be a strong determinant for the maintenance of paralogous

genes. According to their model, paralogs would drift from one another in terms of expression levels (Z. Gu et al. 2002) because only their cumulative abundance is gauged by natural selection. Functional divergence at the protein level would therefore not be necessary for paralogs to lose their buffering ability, divergence of expression would be enough. Once a paralog is dominating expression level, the loss of the second copy becomes almost inconsequential, rendering its loss effectively neutral. Our results support this model by showing that the loss of the least expressed paralog of a pair is generally less consequential than the loss of the most expressed ones (Fig 6B). If gene expression levels evolve at a faster rate than protein functions, this type of sub-functionalization could be the dominating cause of paralog maintenance and at the same time contribute largely to shape the robustness landscape of cells to LOF mutation. Interestingly, we found that in general, heteromerizing pairs of paralogs have more symmetrical expression levels than non-heteromerizing ones (Fig 6D). Heteromerization could slow down gene expression drifting and contribute to paralog maintenance, which could explain the relatively older age (higher dS) of heteromers (Fig 2D). Interestingly, for heteromerizing ones, difference of deleteriousness between paralogs is strongly correlated with the asymmetry in mRNA expression (Fig 6C and Fig EV5B), indicating that there are larger effects in terms of the deleteriousness effects when the paralogs are dosage balanced and vice versa. In addition, maintenance of better dosage balance through regulation at transcriptional and post-transcriptional level (Fig EV5C and D), indicate a strict contingency to on the stoichiometry, most likely imposed by the requirement for the assembly of the heteromers. Finally, we observed that heterodimers of paralogs could be more dependent on each other if their interaction is stronger. Our results concern a very small set of proteins and uses a proxy for binding strength. They will therefore need to be investigated further.

Overall, our analyses show that not all paralogs are equally likely to buffer each other's LOF in human cells. The underlying mechanisms for this ability remain to be fully understood beyond gene expression and protein-protein interactions, and may depend on the specific function of paralogs. Overall, considering the frequent occurrence of copy number variations in cancer cells, the insights obtained from this study regarding the mechanism of robustness of duplicates, could be relevant in the development of cancer therapies. However, more detailed functional analyses will be required to fully determine what is the role of paralog dependency and how dependency could be driven by the physical assembly of paralogs. Since previous studies have shown that this dependency could take place through post-translational regulation, a systematic combination of gene LOF and protein abundance measurements would be the next important step in this investigation to efficiently identify potentially dependent pairs of paralogous genes.

Materials and Methods section

Protein-protein interactions

The human protein-protein interaction data is obtained from BioGRID (Chatr-Aryamontri et al. 2015; Livstone et al. 2011) (Data ref: BioGRID, 2018) and IntAct (Orchard et al. 2014)(Data ref: IntAct, 2019). While defining all methods of detection for protein-protein interactions ('all PPI'), co-fractionation, protein-RNA, co-localization, proximity Label-MS, and affinity capture-RNA were removed because they are not strictly speaking capturing PPIs. A subset of these methods capturing 'all PPI', defined as two-hybrid, biochemical activity, protein-peptide, PCA and Far Western were considered as methods detecting 'direct PPIs'. The number of PPI partners per gene are provided in Dataset EV6.

Gene sets: paralogs and singletons

The set of human paralogous was obtained from Lan et al. study (Lan and Pritchard 2016) and is enriched for small-scale duplication events (1436 pairs). This set of paralogs was completed with a set of paralogs from whole-genome duplication events, obtained from Ohnologs-2 database (Singh, Arora, and Isambert 2015) using the strictest set (Data ref: Ohnolog 2018). As a complete set, 3132 non-redundant pairs of paralogs were used in the study (Dataset EV1). Only the paralogs for which annotations exist in the Ensembl Compara database (Herrero et al. 2016) were used in the analysis. For the merging of the datasets, gene ids of the paralogs were obtained from both Ensembl release 75 and 95 (Zerbino et al. 2018). Protein ids of the paralogs were retrieved from Ensembl Compara (Herrero et al. 2016).

Pure singletons were identified using BLASTP (Altschul et al. 1990) searches of the unique sequences from human proteome (Data ref: Human proteome sequences, 2018) against itself. Any protein that had no hits with E-value smaller than 0.001 over a segment longer than 0.6 times the smaller protein was considered as singleton. Gene symbols were used to merge the data from paralogs and protein-interaction data.

List of paralogs and singletons is included as Dataset EV1.

Gene sets: heteromers and homomers

From the protein-protein interactions, heteromer of paralogs were identified as the pairs of paralogs that physically interact with each other. The rest of the paralog pairs were classified as 'not heteromer'. Homomers are the proteins that interact with themselves. This classification was carried out considering both 'all PPI' and 'direct PPI'. The number of homomers and heteromers identified by each method are indicated in Appendix Table S1.

The gene sets (i.e. heteromers and homomers) identified through PPI from BioGRID and IntAct datasets were merged by taking intersections. For instance, heteromers identified in both datasets, were considered in the merged dataset. If the dataset (BioGRID or IntAct) is not mentioned, the merged dataset is used in the given analysis.

List of the heteromers and homomers is included as Dataset EV1.

Gene sets: essential and non-essential genes

Sets of essential and non-essential genes were derived from the union of gene sets reported by DepMap (DepMap 2018) and BAGEL (Hart and Moffat 2016).

List of the heteromers and homomers is included as Dataset EV1.

Sequence divergence scores and age groups of paralogs

dS scores were determined though codeml (Yang 2007). For the dS score estimations, protein sequences and coding sequences (CDS) of the paralogs were obtained from GRCh38 assembly of human genome (Ensembl genome version 95), using pyEnsembl (Rubinsteyn et al. 2017). dS value greater than 5 were not considered in the analysis (eg. Fig 2D), because larger values are likely saturated and non reliable.

The age groups of the paralogs i.e. evolutionary distances in terms of the taxonomy levels were retrieved from Ensembl Compara (Herrero et al. 2016). The evolutionary distances of taxonomy levels were obtained from the Ensembl species tree (Data ref: Ensembl species tree, 2019).

dS values, age groups and evolutionary distances of the age groups are included in Dataset EV1.

CRISPR score dataset CS1 (Wang et al. 2015)

The CS values of set CS1 were derived from data from genome-wide CRISPR-Cas9 screening experiment in Wang et al study (Wang et al. 2015). The raw sequencing read counts were reanalyzed to remove all gRNA that hit more than one locus in the genome (multi-hit gRNAs) as these could possibly lead to double gene knockouts, particularly for young paralogs. For the cell lines with replicated experiments, replicates of the read count data were averaged. The resulting raw read counts were used as input of BAGEL (Hart and Moffat 2016)to calculate the fold changes. The fold changes calculated by BAGEL are then multiplied by -1 (in order to scale them according to the gene essentiality), so that lower values indicate relative deleteriousness. Z-score normalised fold-change values are used as CS values per gene. Gene-wise CS values from CS1 dataset are included in Dataset EV3.

CRISPR score dataset CS2 (DepMap 2018)

We used the published data from DepMap consortium (DepMap 2018) (18Q3 release), that corresponds to genome-wide CRISPR knock-out screen in cancer cell lines. The CS values in this case are corrected for copy-number variation by CERES method (Meyers et al. 2017). A total of 450 cell lines with replicated experiments were considered in this dataset (Table EV1). CS values obtained from the DepMap repository (DepMap 2018) (file name: gene_dependency.csv) were z-score normalised and integrated in the overall CS dataset. Gene-wise CS values are included in Dataset EV3.

CRISPR score dataset CS2.1 (DepMap 2018)

The CS2.1 dataset was generated by analysing data for the same experimental system as CS2 (DepMap 2018) (18Q3 release) but with removal of 'multi-hit' gRNA that may lead to double paralog knockouts. gRNA-wise fold change values (file name: logfold_change.csv) were used in the analysis. The associated gRNA to gene map (file name: guide_gene_map.csv) was used to obtain gene-wise CS values. The fold change values per gene were calculated using BAGEL tool (Hart and Moffat 2016). The fold changes calculated by BAGEL are then multiplied by -1 (in order to scale them according to the gene essentiality), so that lower values indicate relative deleteriousness. . Z-score normalised fold-change values are used as CS values per gene. Gene-wise CS values are included in Dataset EV3.

CRISPR score dataset CS3 (Shifrut et al. 2018)

CS values for the CS3 dataset were obtained from a genome-wide CRISPR-Cas9 screening experiment in primary T-cells (Shifrut et al. 2018). This dataset serves as an independent reference to the cancer or immortalized cell lines used in the other datasets. gRNAs obtained from the study were first filtered to remove all the multi-hit gRNAs. CS values per gene were obtained processing the gRNA counts through BAGEL (Hart and Moffat 2016). The fold changes calculated by BAGEL are then multiplied by -1 (in order to scale them according to the gene essentiality), so that lower values indicate relative deleteriousness. Z-score normalised fold-change values are used as CS values per gene. The CS3 dataset is included in Dataset EV3.

Merging of CRISPR score datasets

For the comparative analysis of the 4 datasets, CS values in each dataset were first quantile normalized and individual datasets were merged by gene symbols (Ensembl release version 75). Cell-line wise merged CS datasets are available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-BSST233 and aggregated CS values from all datasets are provided in Dataset EV3.

In case of CS datasets CS2 and CS 2.1, as an aggregated CS value per gene, the average CS value per gene over cell lines was computed. Mean and median aggregation was found to correlate very strongly (pearson's r~0.99), therefore mean aggregation was used as a method of choice. Unless mentioned, in the analysis, the average CS values across datasets are used as a vector of gene-wise CS values, for instance in case of Fig 3 and Fig EV2.

GO enrichment analysis

The GO Molecular Function enrichment analysis was performed using GSEA (Subramanian et al. 2005) and Enrichr (E. Y. Chen et al. 2013; Kuleshov et al. 2016) through gseapy (https://github.com/zqfang/GSEApy). Note that GO gene sets may originate from evidences which may not entirely be independent from the rest of the data used in the metaanalysis. Therefore, potentially confounding sources of evidence pertaining to the sequence orthology (Inferred from Sequence Orthology (ISO)) and Inferred from Physical Interaction (IPI)) were removed from the gene set annotation file.

List of all the paralogs was used as the reference set and the list of heteromeric paralogs from the 'all PPI' data (analysis shown in Fig 3 and EV2) and from the 'direct PPI' (analysis shown in Appendix Fig S6) were used as the test sets. GO term annotations used in the analysis are included as Dataset EV4 and P-values are available in Dataset EV5.

mRNA expression

In order to obtain gene expression levels of paralogs, we used transcriptomics data from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al. 2012). We considered data of the 374 cell lines that had complementary CS data in the CS2 and CS2.1 datasets (see Dataset EV2 for cell lines used). Raw RNAseq alignment files (BAM format) were obtained from Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/). Expression of paralogous genes may be underestimated due to their sequence similarity. In order to address this confounding factor, we only considered uniquely aligned reads. Such reads were obtained by filtering the raw BAM files using SAMtools (H. Li et al. 2009) command: 'samtools view -bq 254 -F 512 $bamp | samtools rmdup -sS - $bamp.unique.bam'. Here $bamp is the path to the raw BAM file. Next, the FPKM values were estimated using Cufflinks (Trapnell et al. 2012): 'cufflinks -p 1 --max-frag-multihits 1 -g $gtfp -o output_folder $bamp.unique.bam'. Here $gtfp is path to the annotation file (*Homo sapiens,* assembly:GRCh37, Ensembl release:75) and $bamp.unique.bam is path to the BAM file containing only unique reads (as made in the preceding step). Gene wise mRNA abundance is included as Dataset EV7. Cell line wise mRNA abundance is available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-BSST233.

Protein expression

Protein expression data for 49 cell lines that are also represented in the mRNA expression dataset and CS datasets was retrieved from Ensembl expression atlas (Papatheodorou et al. 2018). This dataset is available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-BSST233.

Classification models

CS values, mRNA expression, and number of interaction partners of the protein are used as feature set to predict whether the gene is heteromer or not (target). Four different classifiers that provide feature importance values were used: Linear SVM, Random Forest, AdaBoost and Decision Tree. Ensemble machine learning classification models were trained using scikit-learn (Pedregosa et al. 2011). For training, five fold cross validations were carried out. In each cross validation 40% of the data was used as a testing set. For each classifier, default parameters were used to train the models. In order to balance the unbalanced classes, equal sized data was bootstrapped from the bigger class. ROC-AUC value of a classifier was calculated as an average of the all the cross-validation and bootstrapped runs.

Protein interaction interfaces

Length of the interaction interface between heteromeric paralogs was obtained from Interactome INSIDER (Meyer et al. 2018). Structures of the interacting paralogs

(Appendix Fig S15) were obtained from Interactome3D database (Mosca, Céol, and Aloy 2013).

Data analysis and visualization

For the retrieval of the CDS and protein sequences, PyEnsembl (Rubinsteyn et al. 2017) was used. For mapping of ids, uniprot REST API (UniProt Consortium 2019) was used. Protein structures were visualized using UCSF Chimera (Pettersen et al. 2004). For general statistical analysis, SciPy (Jones et al., n.d.) was used. Partial correlations were estimated using Pingouin (Vallat 2018). Plots were generated using matplotlib (Hunter 2007), seaborn (Waskom et al. 2018) and rohan (Dandage 2019) was used to generate figures. Ensemble machine learning modeling was carried out using scikit-learn (Pedregosa et al. 2011). Anaconda virtual environment was used to install external programs such as codeml (Yang 2007).

Data Availability

Cell-line wise CS values, mRNA expression values and protein expression is deposited at the BioStudies database (http://www.ebi.ac.uk/biostudies), under accession number S-BSST233. The codes used for the curation of the data and metaanalysis in the study are available at: https://github.com/Landrylab/human_paralogs.

Acknowledgements

We thank the members of the Landrylab for discussions. We thank Anna Fijarczyk, Philippe Després, Carla Bautista, Johan Hallin, Angel Cisneros, Diana Ascencio, Ugo Dionne and Axelle Marchant for insightful discussions and comments on the manuscript. RD is funded by Fonds de recherche du Québec-Santé (FRQS) Programme Postdoctoral. This research was supported by the Canadian Institute of Health Research (CIHR) Foundation grant (RN348479 - 387697) to CRL. CRL holds the Canada Research Chair in Evolutionary Cell and Systems Biology.

Author's contributions

RD & CRL designed and performed research and wrote the paper.

Declaration of interests

None to declare

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Amoutzias, Grigoris D., David L. Robertson, Yves Van de Peer, and Stephen G. Oliver. 2008. “Choose Your Partners: Dimerization in Eukaryotic Transcription Factors.” *Trends in Biochemical Sciences* 33 (5): 220–29.

Baker, Christopher R., Victor Hanson-Smith, and Alexander D. Johnson. 2013. “Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution.” *Science* 342 (6154): 104–8.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. “The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity.” *Nature* 483 (7391): 603–7.

Barshir, Ruth, Idan Hekselman, Netta Shemesh, Moran Sharon, Lena Novack, and Esti Yeger-Lotem. 2018. “Role of Duplicate Genes in Determining the Tissue-Selectivity of Hereditary Diseases.” *PLoS Genetics* 14 (5): e1007327.

Boncoeur, Emilie, Claire Durmort, Benoît Bernay, Christine Ebel, Anne Marie Di Guilmi, Jacques Croizé, Thierry Vernet, and Jean-Michel Jault. 2012. “PatA and PatB Form a Functional Heterodimeric ABC Multidrug Efflux Transporter Responsible for the Resistance of Streptococcus Pneumoniae to Fluoroquinolones.” *Biochemistry* 51 (39): 7755–65.

Breitkreutz, Ashton, Hyungwon Choi, Jeffrey R. Sharom, Lorrie Boucher, Victor Neduva, Brett Larsen, Zhen-Yuan Lin, et al. 2010. “A Global Protein Kinase and Phosphatase Interaction Network in Yeast.” *Science* 328 (5981): 1043–46.

Bridgham, Jamie T., Justine E. Brown, Adriana Rodríguez-Marí, Julian M. Catchen, and Joseph W. Thornton. 2008. “Evolution of a New Function by Degenerative Mutation in Cephalochordate Steroid Receptors.” *PLoS Genetics* 4 (9): e1000191.

Brookfield, J. F. 1997. “Genetic Redundancy.” *Advances in Genetics* 36: 137–55.

Chatr-Aryamontri, Andrew, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, et al. 2015. “The BioGRID Interaction Database: 2015 Update.” *Nucleic Acids Research* 43 (Database issue): D470–78.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma’ayan. 2013. “Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool.” *BMC Bioinformatics* 14 (April): 128.

Chen, Wei-Hua, Xing-Ming Zhao, Vera van Noort, and Peer Bork. 2013. “Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs.” *PLoS Computational Biology* 9 (5): e1003073.

Dandage, Rohan. 2019. *rraadd88/rohan v0.1.0*. https://doi.org/10.5281/zenodo.2682671.

Dean, E. Jedediah, Jerel C. Davis, Ronald W. Davis, and Dmitri A. Petrov. 2008. “Pervasive and Persistent Redundancy among Duplicated Genes in Yeast.” *PLoS Genetics* 4 (7): e1000113.

DeLuna, Alexander, Michael Springer, Marc W. Kirschner, and Roy Kishony. 2010. “Need-Based up-Regulation of Protein Levels in Response to Deletion of Their Duplicate Genes.” *PLoS Biology* 8 (3): e1000347.

DepMap, Broad. 2018. “DepMap Achilles 18Q3 Public.” Figshare. https://doi.org/10.6084/M9.FIGSHARE.6931364.V1 [DATASET].

Després, Philippe C., Alexandre K. Dubé, Lou Nielly-Thibault, Nozomu Yachie, and Christian R. Landry. 2018. “Double Selection Enhances the Efficiency of Target-AID and Cas9-Based Genome Editing in Yeast.” *G3* 8 (10): 3163–71.

Diss, Guillaume, Diana Ascencio, Alexander DeLuna, and Christian R. Landry. 2014. “Molecular Mechanisms of Paralogous Compensation and the Robustness of Cellular Networks.”

Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution 322 (7): 488–99.

Diss, Guillaume, Alexandre K. Dubé, Joël Boutin, Isabelle Gagnon-Arsenault, and Christian R. Landry. 2013. "A Systematic Approach for the Genetic Dissection of Protein Complexes in Living Cells." *Cell Reports* 3 (6): 2155–67.

Diss, Guillaume, Isabelle Gagnon-Arsenault, Anne-Marie Dion-Coté, Hélène Vignaud, Diana I. Ascencio, Caroline M. Berger, and Christian R. Landry. 2017. "Gene Duplication Can Impart Fragility, Not Robustness, in the Yeast Protein Interaction Network." *Science* 355 (6325): 630–34.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. "Preservation of Duplicate Genes by Complementary, Degenerative Mutations." *Genetics* 151 (4): 1531–45.

Fortin, Jean-Philippe, Jenille Tan, Karen E. Gascoigne, Peter M. Haverty, William F. Forrest, Michael R. Costa, and Scott E. Martin. 2019. "Multiple-Gene Targeting and Mismatch Tolerance Can Confound Analysis of Genome-Wide Pooled CRISPR Screens." *Genome Biology* 20 (1): 21.

Gibson, T. J., and J. Spring. 1998. "Genetic Redundancy in Vertebrates: Polyploidy and Persistence of Genes Encoding Multidomain Proteins." *Trends in Genetics: TIG* 14 (2): 46–49; discussion 49–50.

Gonçalves, Emanuel, Athanassios Fragoulis, Luz Garcia-Alonso, Thorsten Cramer, Julio Saez-Rodriguez, and Pedro Beltrao. 2017. "Widespread Post-Transcriptional Attenuation of Genomic Copy-Number Variation in Cancer." *Cell Systems* 5 (4): 386–98.e4.

Gout, Jean-François, Daniel Kahn, Laurent Duret, and Paramecium Post-Genomics Consortium. 2010. "The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution." *PLoS Genetics* 6 (5): e1000944.

Gout, Jean-Francois, and Michael Lynch. 2015. "Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization." *Molecular Biology and Evolution* 32 (8): 2141–48.

Gu, Zhenglong, Lars M. Steinmetz, Xun Gu, Curt Scharfe, Ronald W. Davis, and Wen-Hsiung Li. 2003. "Role of Duplicate Genes in Genetic Robustness against Null Mutations." *Nature* 421 (6918): 63–66.

Gu, Z., D. Nicolae, H. H. S. Lu, and W. H. Li. 2002. "Rapid Divergence in Expression between Duplicate Genes Inferred from Microarray Data." *Trends in Genetics: TIG*. https://www.sciencedirect.com/science/article/pii/S0168952502028378.

Hart, Traver, and Jason Moffat. 2016. "BAGEL: A Computational Framework for Identifying Essential Genes from Pooled Library Screens." *BMC Bioinformatics*. https://doi.org/10.1186/s12859-016-1015-8.

Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, et al. 2016. "Ensembl Comparative Genomics Resources." *Database: The Journal of Biological Databases and Curation* 2016 (May). https://doi.org/10.1093/database/baw053.

Hsiao, Tzu-Lin, and Dennis Vitkup. 2008. "Role of Duplicate Genes in Robustness against Deleterious Human Mutations." *PLoS Genetics* 4 (3): e1000014.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science Engineering* 9 (3): 90–95.

Ihmels, Jan, Sean R. Collins, Maya Schuldiner, Nevan J. Krogan, and Jonathan S. Weissman. 2007. "Backup without Redundancy: Genetic Interactions Reveal the Cost of Duplicate Gene Loss." *Molecular Systems Biology* 3 (March): 86.

Innan, Hideki, and Fyodor Kondrashov. 2010. "The Evolution of Gene Duplications: Classifying and Distinguishing between Models." *Nature Reviews. Genetics* 11 (2): 97–108.

Ishikawa, Koji, Koji Makanae, Shintaro Iwasaki, Nicholas T. Ingolia, and Hisao Moriya. 2017. "Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry

of Protein Complexes." *PLoS Genetics* 13 (1): e1006554.
Ispolatov, I. 2005. "Binding Properties and Evolution of Homodimers in Protein-Protein Interaction Networks." *Nucleic Acids Research* 33 (11): 3629–35.
Jones, Eric, Travis Oliphant, Pearu Peterson, and Others. n.d. "SciPy: Open Source Scientific Tools for Python." http://www.scipy.org/.
Kafri, Ran, Arren Bar-Even, and Yitzhak Pilpel. 2005. "Transcription Control Reprogramming in Genetic Backup Circuits." *Nature Genetics* 37 (3): 295–99.
Kaltenegger, Elisabeth, and Dietrich Ober. 2015. "Paralogue Interference Affects the Dynamics after Gene Duplication." *Trends in Plant Science* 20 (12): 814–21.
Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2019. "Variation across 141,456 Human Exomes and Genomes Reveals the Spectrum of Loss-of-Function Intolerance across Human Protein-Coding Genes." *bioRxiv*. https://doi.org/10.1101/531210.
Kosicki, Michael, Kärt Tomberg, and Allan Bradley. 2018. "Repair of Double-Strand Breaks Induced by CRISPR–Cas9 Leads to Large Deletions and Complex Rearrangements." *Nature Biotechnology*. https://doi.org/10.1038/nbt.4192.
Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. "Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44 (W1): W90–97.
Lan, Xun, and Jonathan K. Pritchard. 2016. "Coregulation of Tandem Duplicate Genes Slows Evolution of Subfunctionalization in Mammals." *Science* 352 (6288): 1009–13.
Lavi, Orit. 2015. "Redundancy: A Critical Obstacle to Improving Cancer Therapy." *Cancer Research* 75 (5): 808–12.
Lever, Jake, Eric Y. Zhao, Jasleen Grewal, Martin R. Jones, and Steven J. M. Jones. 2019. "CancerMine: A Literature-Mined Resource for Drivers, Oncogenes and Tumor Suppressors in Cancer." *Nature Methods* 16 (6): 505–7.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
Li, Jingjing, Zineng Yuan, and Zhaolei Zhang. 2010. "The Cellular Robustness by Genetic Redundancy in Budding Yeast." *PLoS Genetics* 6 (11): e1001187.
Livstone, Michael, Michael Livstone, Bobby-Joe Breitkreutz, Chris Stark, Lorrie Boucher, Andrew Chatr-Aryamontri, Rose Oughtred, et al. 2011. "The BioGRID Interaction Database." *Nature Precedings*. https://doi.org/10.1038/npre.2011.5627.1.
Lynch, M., and A. Force. 2000. "The Probability of Duplicate Gene Preservation by Subfunctionalization." *Genetics* 154 (1): 459–73.
Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. "The Probability of Preservation of a Newly Arisen Gene Duplicate." *Genetics* 159 (4): 1789–1804.
Melton, D. W. 1994. "Gene Targeting in the Mouse." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 16 (9): 633–38.
Meyer, Michael J., Juan Felipe Beltrán, Siqi Liang, Robert Fragoza, Aaron Rumack, Jin Liang, Xiaomu Wei, and Haiyuan Yu. 2018. "Interactome INSIDER: A Structural Interactome Browser for Genomic Studies." *Nature Methods* 15 (2): 107–14.
Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. "Computational Correction of Copy Number Effect Improves Specificity of CRISPR–Cas9 Essentiality Screens in Cancer Cells." *Nature Genetics* 49 (October): 1779.
Mosca, Roberto, Arnaud Céol, and Patrick Aloy. 2013. "Interactome3D: Adding Structural Details to Protein Networks." *Nature Methods* 10 (1): 47–53.
Nimwegen, E. van, J. P. Crutchfield, and M. Huynen. 1999. "Neutral Evolution of Mutational

Robustness." *Proceedings of the National Academy of Sciences* 96 (17): 9716–20.
Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, et al. 2014. "The MIntAct Project--IntAct as a Common Curation Platform for 11 Molecular Interaction Databases." *Nucleic Acids Research* 42 (Database issue): D358–63.
Ori, Alessandro, Murat Iskar, Katarzyna Buczak, Panagiotis Kastritis, Luca Parca, Amparo Andrés-Pons, Stephan Singer, Peer Bork, and Martin Beck. 2016. "Spatiotemporal Variation of Mammalian Protein Complex Stoichiometries." *Genome Biology* 17 (March): 47.
Papatheodorou, Irene, Nuno A. Fonseca, Maria Keays, Y. Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, et al. 2018. "Expression Atlas: Gene and Protein Expression across Multiple Studies and Organisms." *Nucleic Acids Research* 46 (D1): D246–51.
Papp, Balázs, Csaba Pál, and Laurence D. Hurst. 2003. "Dosage Sensitivity and the Evolution of Gene Families in Yeast." *Nature* 424 (6945): 194–97.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12: 2825–30.
Pereira-Leal, Jose B., Emmanuel D. Levy, Christel Kamp, and Sarah A. Teichmann. 2007. "Evolution of Protein Complexes by Duplication of Homomeric Interactions." *Genome Biology* 8 (4): R51.
Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. "UCSF Chimera--a Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605–12.
Pickett, F. B., and D. R. Meeks-Wagner. 1995. "Seeing Double: Appreciating Genetic Redundancy." *The Plant Cell* 7 (9): 1347–56.
Plata, Germán, and Dennis Vitkup. 2014. "Genetic Robustness and Functional Evolution of Gene Duplicates." *Nucleic Acids Research* 42 (4): 2405–14.
Qian, Wenfeng, and Jianzhi Zhang. 2008. "Gene Dosage and Gene Duplicability." *Genetics* 179 (4): 2319–24.
Rajoo, Sasikumar, Pascal Vallotton, Evgeny Onischenko, and Karsten Weis. 2018. "Stoichiometry and Compositional Plasticity of the Yeast Nuclear Pore Complex Revealed by Quantitative Fluorescence Microscopy." *Proceedings of the National Academy of Sciences of the United States of America* 115 (17): E3969–77.
Rubinsteyn, Alex, Tavi Nathanson, Julia Kodysh, Tim O'Donnell, Arun Ahuja, Jeff Hammerbacher, B. Arman Aksoy, Brent Pedersen-Bioinformatics, Valentin Grouès, and Isaac Hodes. 2017. *Hammerlab/pyensembl: Version 1.1.0*. https://doi.org/10.5281/zenodo.822502.
Shifrut, Eric, Julia Carnevale, Victoria Tobin, Theodore L. Roth, Jonathan M. Woo, Christina T. Bui, P. Jonathan Li, Morgan E. Diolaiti, Alan Ashworth, and Alexander Marson. 2018. "Genome-Wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function." *Cell* 175 (7): 1958–71.e15.
Singh, Param Priya, Séverine Affeldt, Ilaria Cascone, Rasim Selimoglu, Jacques Camonis, and Hervé Isambert. 2012. "On the Expansion of 'Dangerous' Gene Repertoires by Whole-Genome Duplications in Early Vertebrates." *Cell Reports* 2 (5): 1387–98.
Singh, Param Priya, Jatin Arora, and Hervé Isambert. 2015. "Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes." *PLoS Computational Biology* 11 (7): e1004394.
Sousa, Abel, Emanuel Gonçalves, Bogdan Mirauta, David Ochoa, Oliver Stegle, and Pedro Beltrao. 2019. "Multi-Omics Characterization of Interaction-Mediated Control of Human

Protein Abundance Levels." *Molecular & Cellular Proteomics: MCP*, June. https://doi.org/10.1074/mcp.RA118.001280.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Taggart, James C., and Gene-Wei Li. 2018. "Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes." *Cell Systems* 7 (6): 580–89.e4.

Teichmann, Sarah Amalia, and Reiner Albert Veitia. 2004. "Genes Encoding Subunits of Stable Complexes Are Clustered on the Yeast Chromosomes: An Interpretation from a Dosage Balance Perspective." *Genetics* 167 (4): 2121–25.

Thomas, J. H. 1993. "Thinking about Genetic Redundancy." *Trends in Genetics: TIG* 9 (11): 395–99.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-Seq Experiments with TopHat and Cufflinks." *Nature Protocols* 7 (3): 562–78.

UniProt Consortium. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47 (D1): D506–15.

Vallat, Raphael. 2018. "Pingouin: Statistics in Python." *Journal of Open Source Software* 3 (31): 1026.

Wagner, Andreas. 2003. "How the Global Structure of Protein Interaction Networks Evolves." *Proceedings. Biological Sciences / The Royal Society* 270 (1514): 457–66.

Wang, Tim, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. 2015. "Identification and Characterization of Essential Genes in the Human Genome." *Science* 350 (6264): 1096–1101.

Waskom, Michael, Olga Botvinnik, Drew O'Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C. Gemperline, et al. 2018. *Mwaskom/seaborn: v0.9.0 (July 2018)*. https://doi.org/10.5281/zenodo.1313201.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, and C. G. Giro'n. 2018. "Ensembl 2018. Nuclic Acids Res."

Data citations

BioGRID PPI dataset (2018) BioGRID (https://downloads.thebiogrid.org/Download/BioGRID/Release-Archive/BIOGRID-3.5.167/BIOGRID-ORGANISM-3.5.167.tab2.zip) [DATASET]

IntAct PPI dataset (2019) IntAct (ftp://ftp.ebi.ac.uk/pub/databases/intact/2019-03-22/) [DATASET]

Ensembl Species Tree (2019) Ensembl Species Tree (https://useast.ensembl.org/info/about/species.html) [DATASET]

Ohnolog-2 WGD paralogs (2018) Ohnolog (http://ohnologs.curie.fr/cgi-bin/BrowsePage.cgi?org=hsapiens) [DATASET]

Human proteome sequences (2018) Human proteome sequences (ftp://ftp.ensembl.org/pub/release-95/fasta/homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz) [DATASET]

Expanded view figures

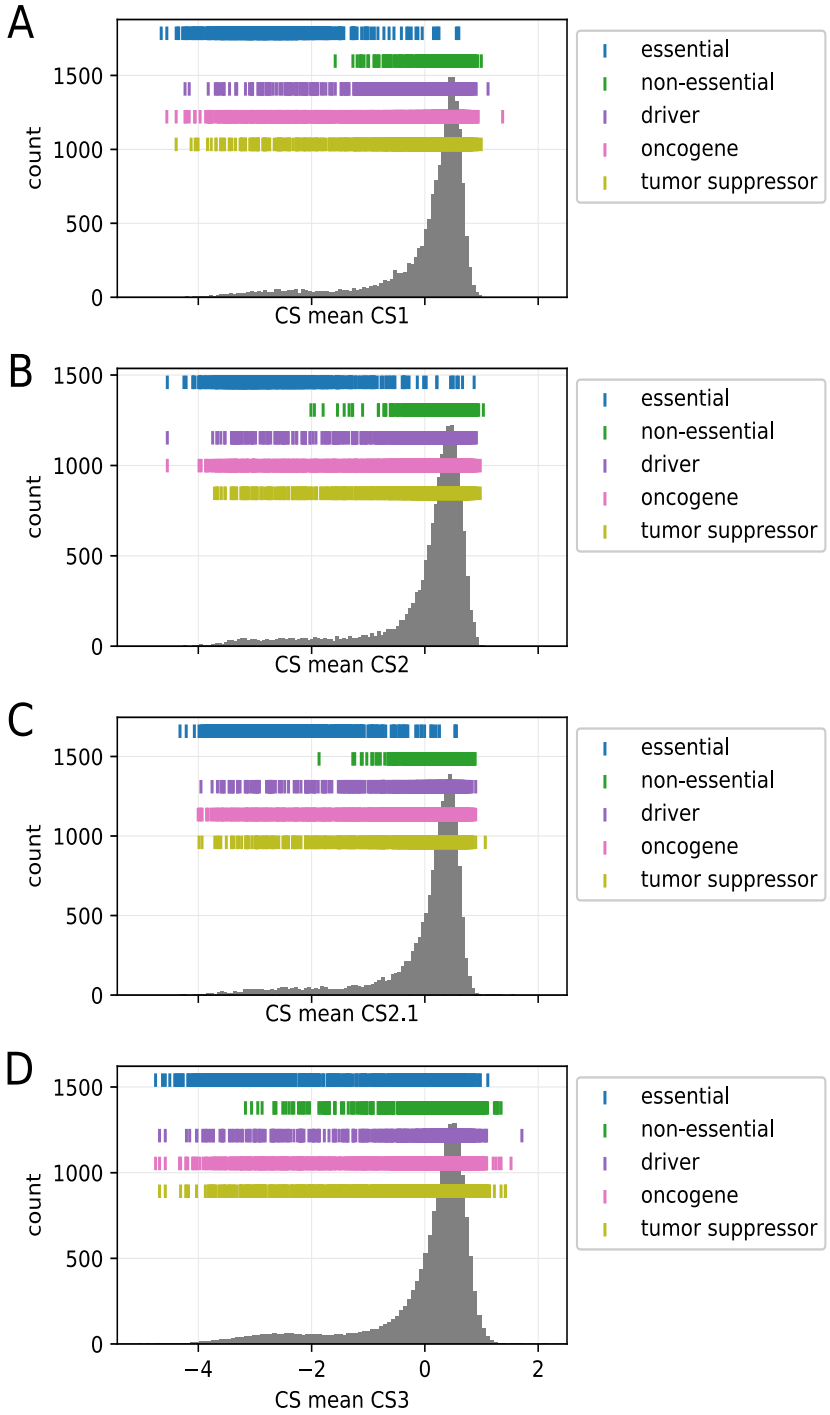
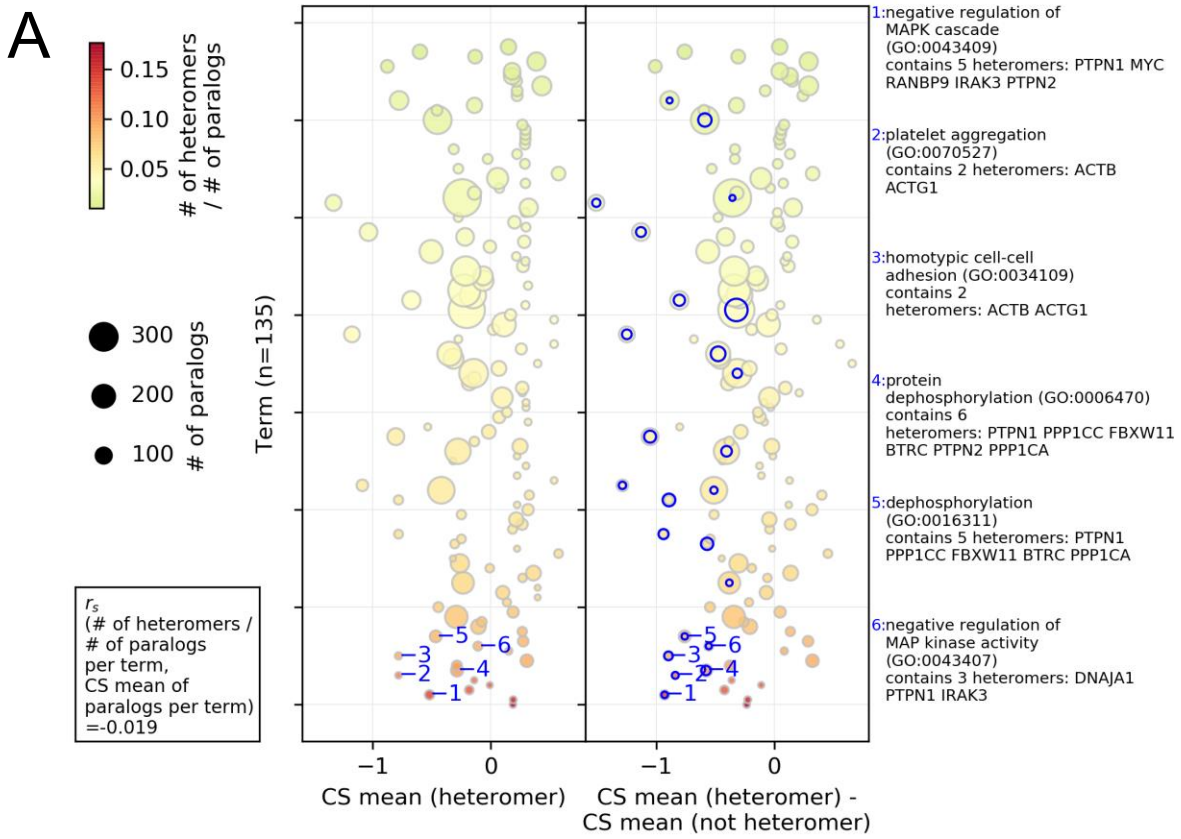
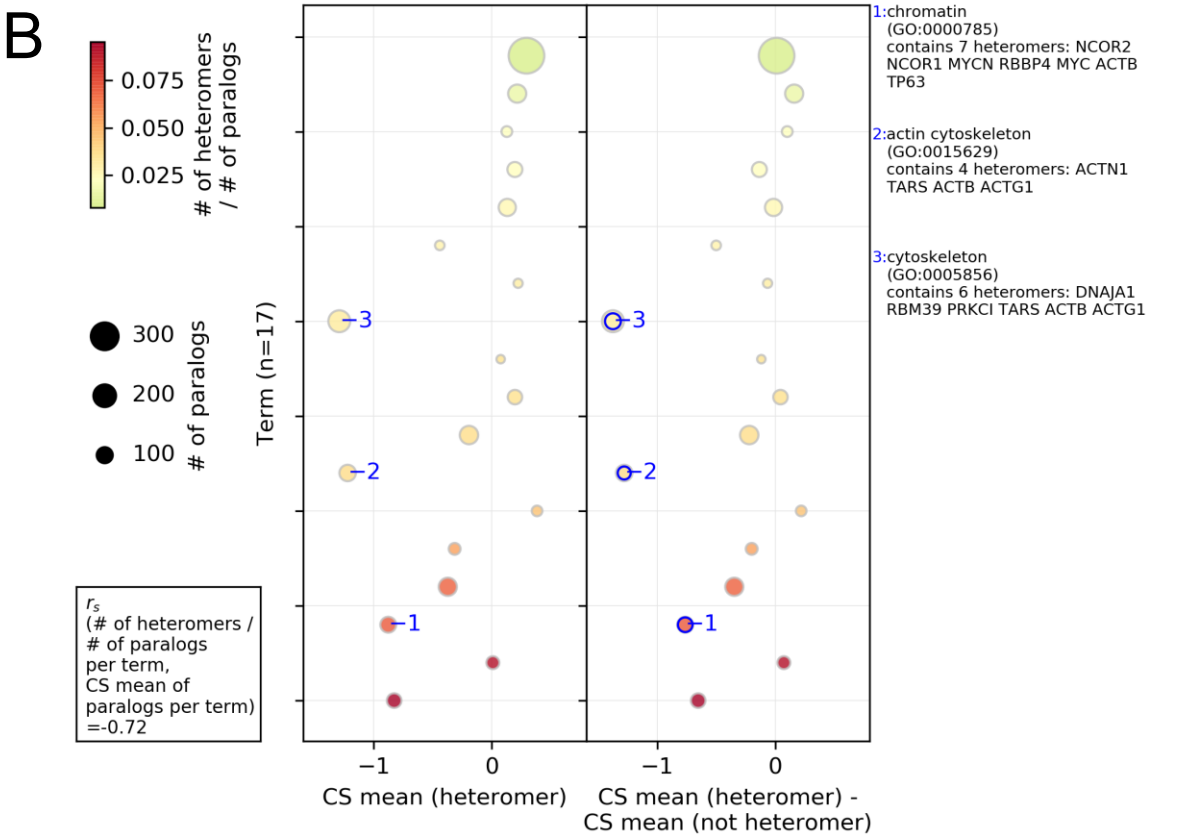


Fig EV1: Distribution of CS values in the 4 CS datasets.
The locations of essential and non-essential genes (taken as a union set of genes reported by DepMap (DepMap 2018) and BAGEL (Hart and Moffat 2016) are denoted on the distributions. he locations of the cancer drivers, oncogenes and tumor suppressors are also denoted on the distribution (derived from (Lever et al. 2019)).


A
0.15
0.10
0.05
# of heteromers / # of paralogs
300
200
100
# of paralogs
Term (n=135)
r_s (# of heteromers / # of paralogs per term, CS mean of paralogs per term) =-0.019
1:negative regulation of MAPK cascade (GO:0043409) contains 5 heteromers: PTPN1 MYC RANBP9 IRAK3 PTPN2
2:platelet aggregation (GO:0070527) contains 2 heteromers: ACTB ACTG1
3:homotypic cell-cell adhesion (GO:0034109) contains 2 heteromers: ACTB ACTG1
4:protein dephosphorylation (GO:0006470) contains 6 heteromers: PTPN1 PPP1CC FBXW11 BTRC PTPN2 PPP1CA
5:dephosphorylation (GO:0016311) contains 5 heteromers: PTPN1 PPP1CC FBXW11 BTRC PPP1CA
6:negative regulation of MAP kinase activity (GO:0043407) contains 3 heteromers: DNAJA1 PTPN1 IRAK3
−1
0
CS mean (heteromer)
CS mean (heteromer) - CS mean (not heteromer)


B
0.075
0.050
0.025
# of heteromers / # of paralogs
300
200
100
# of paralogs
Term (n=17)
r_s (# of heteromers / # of paralogs per term, CS mean of paralogs per term) =-0.72
1:chromatin (GO:0000785) contains 7 heteromers: NCOR2 NCOR1 MYCN RBBP4 MYC ACTB TP63
2:actin cytoskeleton (GO:0015629) contains 4 heteromers: ACTN1 TARS ACTB ACTG1
3:cytoskeleton (GO:0005856) contains 6 heteromers: DNAJA1 RBM39 PRKCI TARS ACTB ACTG1
−1
0
CS mean (heteromer)
CS mean (heteromer) - CS mean (not heteromer)

1

Fig EV2. Association between the biological processes and cellular components of paralogs, their probability of heteromerization and the effect of gene LOF on cell proliferation, in case of the heteromers defined by the 'all PPI' only.

Gene set analysis for the Biological Processes and Cellular Components aspect are shown in the panel A and B respectively.

In each panel, average CS values (x-axis) of paralogs (heteromer or not heteromer) belonging to a gene set were used in the analysis. GO terms are sorted according to their proportion of heteromeric paralogs (i.e. # of heteromers/ # of paralogs). The size of the circles represents the number of paralog pairs in a category and the colors represent the proportion of heteromers in the category. The difference between the average CS value of the heteromers and average CS value of the non-heteromers is plotted in the right panel. The terms with significant difference between the average CS value of the heteromers and average CS value of the non-heteromers (estimated by two-sided t-test) are annotated with the blue edges. Descriptions of the representative significant GO terms with the highest difference are shown in the right side-panel. Spearman rank correlation between the proportion of the heteromers in the GO terms and the average CS value of paralogs in the term (r_s(# of heteromers / # of paralogs per term, CS mean of paralogs per term)) is shown in left right corner. Only GO molecular functions with more than 10% of the number of paralogs in all the gene sets are shown.

See Dataset EV4 for GO term annotations shown on this figure. Note that not all gene sets are independent because some genes are in several categories.

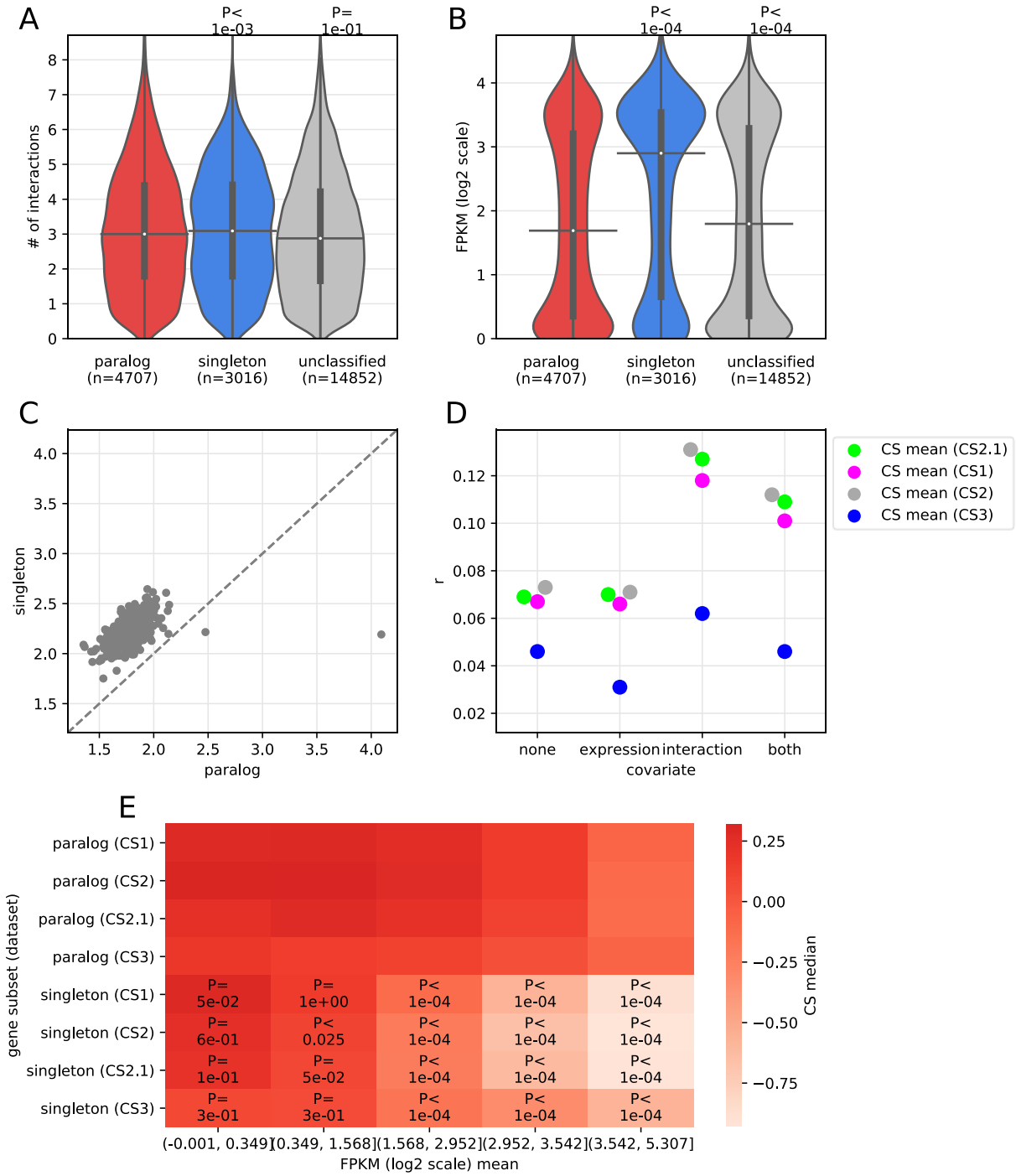


Fig EV3: Paralogs have less number of interactions and less mRNA expression compared to singletons.

A) Paralogs have less interaction partners than singletons. Number of interactions are in log2-transformed.

B) Paralogs have lower mRNA expression than singletons. mRNA expression of genes is shown in terms of log2 of FPKM.

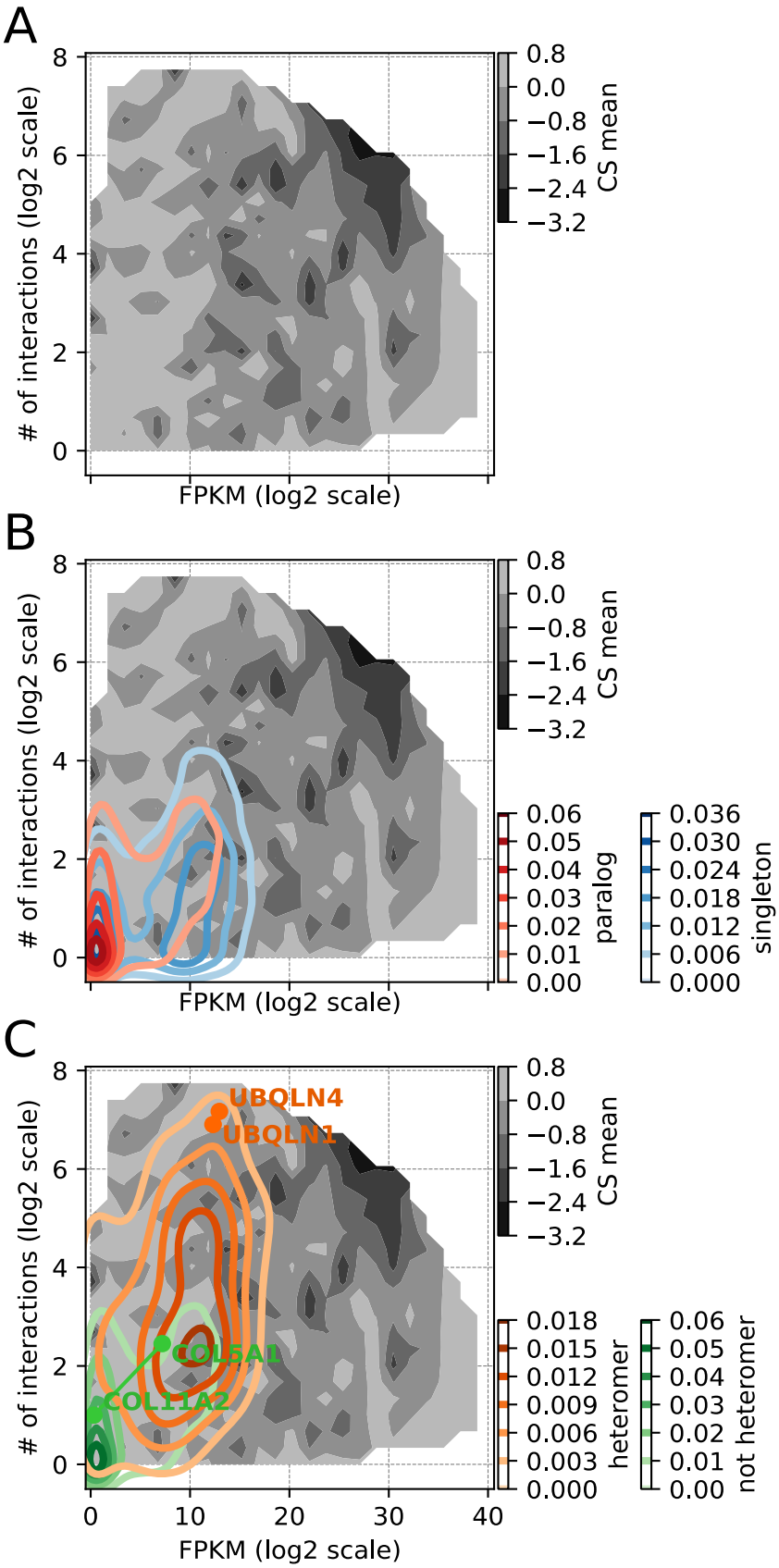
C) Across the majority of the cell-lines, the average mRNA expression of the paralogs is lower than that of singletons. Each point represents the average mRNA expression

(FPKM in log2 scale) for a class (paralog or singleton) in an individual cell line. All points are above the diagonal (dashed gray line), indicating that the effect is systematic and largely cell-line independent.

D) Partial Spearman correlation coefficients (r, shown on the y axis) between the CS value and a paralog status of a gene (paralog or singleton, binary variable, 1 : paralog, 0 : singleton). The correlations were calculated while controlling for none of mRNA expression and number of interactions (“none”), only mRNA expression (“expression”), only number of interactions (“interaction”) or both (“both”) (as shown on the x axis). Controlling for mRNA expression leads to the greater loss of correlation for interactions the mRNA expression of paralogs is a better contributor to correlation between the CS values and the status of the gene being paralog or singleton (binary variable), than the number of interaction partners.

E) Interdependence of the robustness of paralogs (shown in terms of CS score, y axis) on the mRNA expression (on y axis). Each of the four subpanels correspond to 4 CS datasets. mRNA expression of the genes was binned into 5 equal sized bins. Median of the CS values of the genes in each subset are shown on the heatmap. The P-values from two-sided Mann-Whitney U tests for the comparison of distributions of the CS values of the paralogs versus singletons, in each CS dataset and each bin of mRNA expression are denoted on the heatmap. Distributions across CS values in each case are shown in Appendix Fig S10.

In panels **A** and **B,** P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are denoted by a horizontal black line, whereas the quartiles from the median value are indicated by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.


A
B
C
# of interactions (log2 scale)
FPKM (log2 scale)
CS mean
0.8
0.0
−0.8
−1.6
−2.4
−3.2
paralog
singleton
heteromer
not heteromer
UBQLN4
UBQLN1
COL5A1
COL11A2

1

Fig EV4: Landscape of the robustness of human cell lines to LOF direct physical interactions only.

RNA expression level (log_2 scale FPKM scores) and number of direct protein-protein interaction partners (log_2 scale) are strong determinants of the deleteriousness of gene LOF.

A) The landscape shows the CS value as a function of these two parameters. Regions of the landscape with high mRNA expression and large number of interactions clearly show depletion in CS values.

B) Kernel density estimates for paralogs and singletons are overlaid on the landscape to indicate their level of occupancy. The density of paralogs is biased strongly towards lower expression levels compared to singletons.

C) Similar to **B,** paralogous heteromers and paralogous non-heteromers are overlaid on the landscape to indicate their level of occupancy. The density of heteromers is biased strongly towards higher number of protein interaction partners, compared to non-heteromers.

D) Locations of representative heteromeric and non-heteromeric pairs of paralogous genes is shown on the robustness landscape.

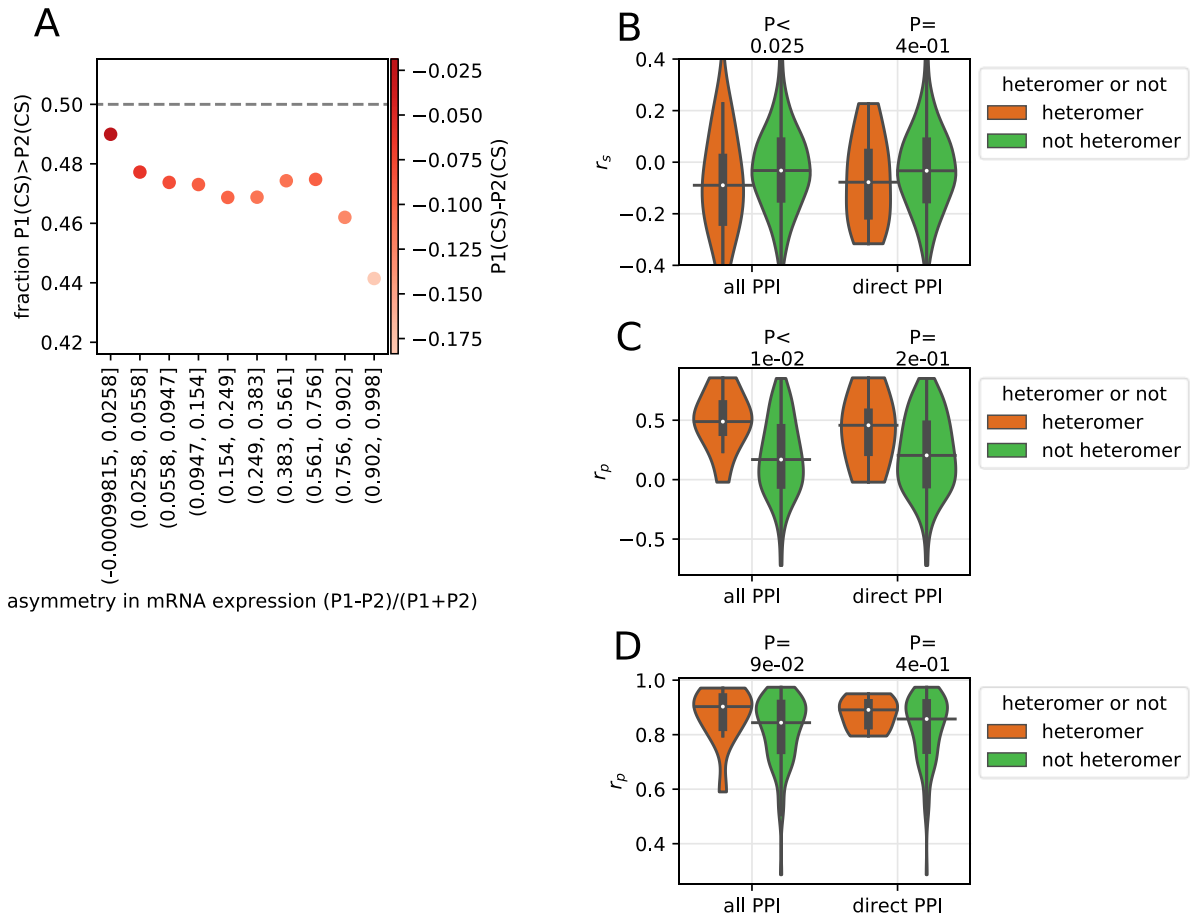


Fig EV5: Relationship between the asymmetry of expression and the relative deleteriousness of paralog.

A) The probability that a highly expressed paralog P1 has higher CS than comparatively weakly expressed paralog P2, as a function of its normalized relative mRNA expression to P2. Probabilities higher than 0.5 would indicate that paralog P1 is more likely to have a higher CS (less deleterious) value than P2. The scaled asymmetry of expression is shown on the x-axis. On the left, P1 is more likely to have higher CS value (less deleterious) and expression is symmetrical. On the right, P1 is more likely to have relatively lower CS value (more deleterious) and the expression is asymmetric. Asymmetry in mRNA expression (x axis) was binned into 10 equal size bins. The color of the points represents the average difference of CS value in the bin. Similar analysis with CS2 dataset is shown in Appendix Fig S11A.

B) Average difference of CS value between P1 and P2 (P1(CS) -P2(CS)) is correlated with the asymmetry of mRNA expression (i.e. (P1-P2)/(P1+P2), where mRNA expression of P1 is greater than that of the P2), across cell lines. Each point in the distribution corresponds to the correlation for a single pair of paralogs. r_s: Spearman correlation coefficient. Similar analysis with CS2 dataset is shown in Appendix Fig S11B. See Appendix Fig S12 for relationships between asymmetry of the mRNA expression and difference in CS values for representative pairs of heteromers and non-heteromeric paralogs.

C) Extent of the transcriptional dosage balance in heteromers versus non-heteromers. mRNA expression of the paralogs was correlated across 374 cell lines. r_p: Pearsons's correlation coefficient. mRNA expression values were z-score normalised before estimating correlations.

D) Extent of the post-transcriptional dosage balance in heteromers versus non-heteromers. protein expression of the paralogs was correlated across 49 cell lines. While estimating the partial correlation, the protein expression of the paralogs was controlled with the mRNA expression. r_p: Pearsons's correlation coefficient. Protein and mRNA expression values were z-score normalised before taking the correlations.

In panels **B, C** and **D,** P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

Expanded view tables

Table EV1: Scale of the datasets used in the study.

dataset	# of cell lines	# of genes
CS1	4	17344
CS2/CS2.1	450	17344
CS3	1	17344
total CS	455	17344
mRNA expression	374	17488
protein expression	49	2800
# of PPI	-	16013

Expanded view datasets

Dataset EV1: Gene-wise and paralog-wise annotations used in the study.

Dataset EV2: Cell lines used in the study.

Dataset EV3: CS values per gene.

Dataset EV4: Gene sets and GO terms used in the gene set enrichment analysis.

Dataset EV5: Results of the GO term enrichment analysis.

Dataset EV6: Number of PPIs per gene.

Dataset EV7: mRNA expression per gene.

Appendix

Appendix tables

Page

Appendix Table S1: Number of homomeric (P1P1 or P2P2) and heteromeric (P1P2) paralogs. 2

Appendix figures

Page

Appendix Fig S1: Correlation between CS values across datasets. 5

Appendix Fig S2: Effect of gene LOF for singletons and paralogous genes across cell lines. 6

Appendix Fig S3: The LOFs of paralogous heteromers, defined by 'direct PPI's only, are relatively more deleterious than the LOFs of non-heteromers. 7

Appendix Fig S4: The LOFs of paralogous heteromers, are relatively more deleterious than the LOFs of non-heteromers, across most of the cell lines in the CS datasets. 9

Appendix Fig S5: Paralogs that form heteromers tend to be more deleterious upon LOF than other paralogs, largely independently from the age of paralogs. 10

Appendix Fig S6: Association between the GO gene sets of paralogs, their probability of heteromerization and the effect of gene LOF on cell proliferation, in case of the heteromers defined by the 'direct PPI' only. 11

Appendix Fig S7: Correlations between the effect of LOF of a gene on cell proliferation, mRNA expression and number of protein-protein interaction partners. 13

Appendix Fig S8: Relationship between the effect of LOF of a gene on cell proliferation, mRNA expression and number of protein-protein interaction partners. Related to Fig 4. 14

Appendix Fig S9: Feature importance (shown on the y axis) determined through ensemble classification models (shown on the x axis), for 4 CS datasets. 17

Appendix Fig S10: Dependence of the robustness of paralogs (shown on y-axis in terms of CS values) on mRNA expression (y-axis), across 4 CS datasets. Related to Fig EV3E. 18

Appendix Fig S11: The most expressed paralog (P1) of a pair is more likely to be deleterious than the least expressed (P2), across 374 cell lines. 19

Appendix Fig S12: Relationship between the asymmetry of expression and relative deleteriousness of paralog for the CS2 dataset. 20

Appendix Fig S13: Relationships between asymmetry of the mRNA expression and difference in CS values. Related to Fig 6. 21

Appendix Fig S14: Relationships between the asymmetry of the mRNA expression and the difference in CS values is shown for representative heteromeric and non-heteromeric pairs of paralogs. 22

Appendix Fig S15: Structures of the representative heteromeric paralogs showing the number of interface residues. 23

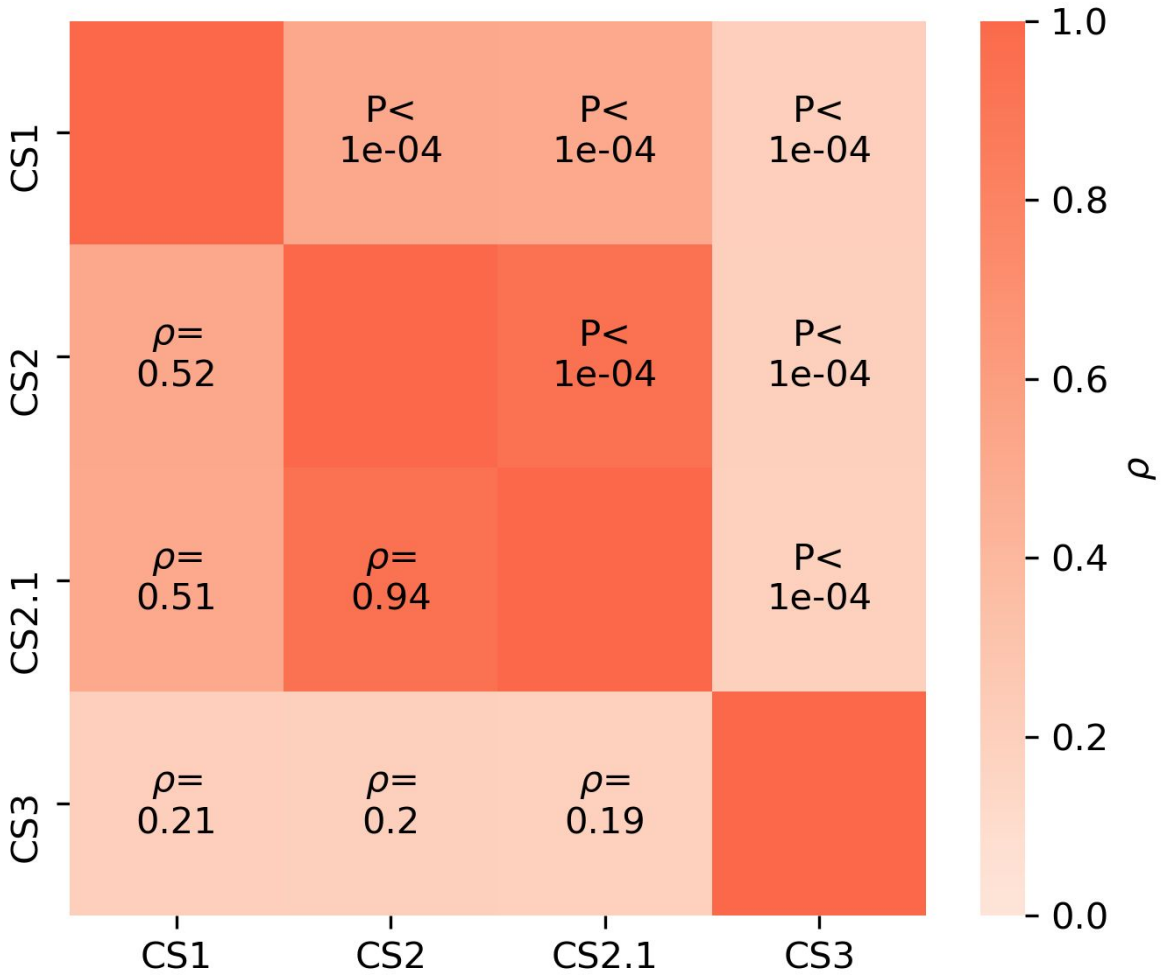
Appendix tables

Appendix Table S1: Number of homomeric (P1P1 or P2P2) and heteromeric (P1P2) paralogs.

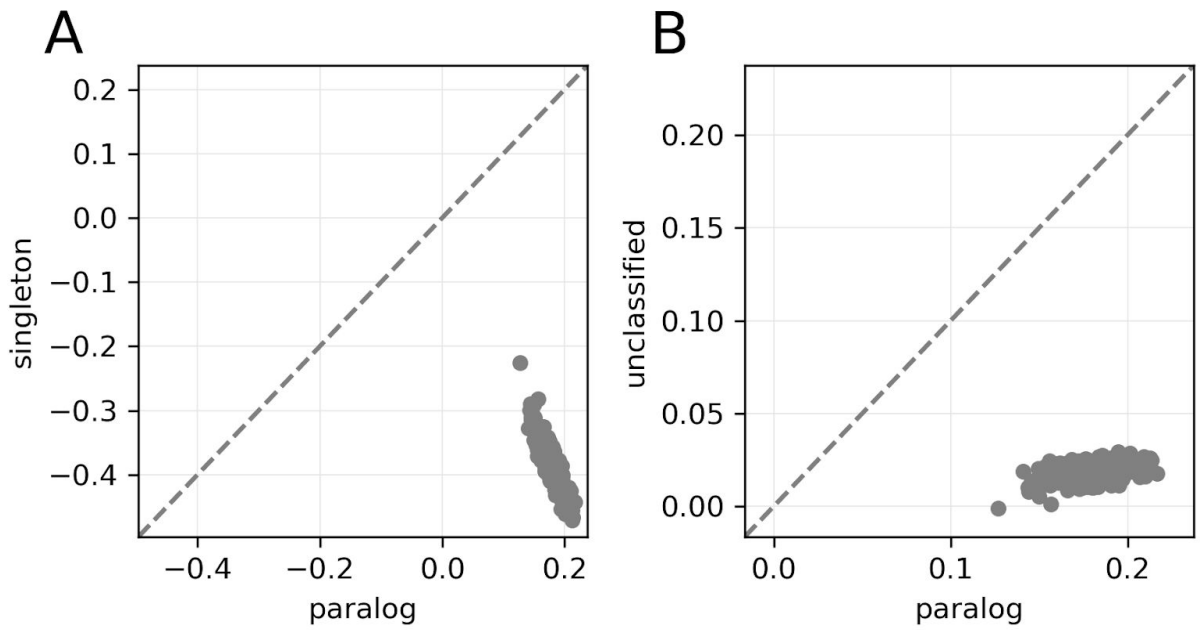
PPI type	gene1 homomer	gene2 homomer	heteromer	# of paralogs
all BioGRID				2231
all BioGRID			P1P2	145
all BioGRID		P2P2		248
all BioGRID		P2P2	P1P2	40
all BioGRID	P1P1			217
all BioGRID	P1P1		P1P2	58
all BioGRID	P1P1	P2P2		133
all BioGRID	P1P1	P2P2	P1P2	60
all IntAct				2570
all IntAct			P1P2	42
all IntAct		P2P2		189
all IntAct		P2P2	P1P2	6
all IntAct	P1P1			192
all IntAct	P1P1		P1P2	23
all IntAct	P1P1	P2P2		81
all IntAct	P1P1	P2P2	P1P2	29
direct BioGRID				2704
direct BioGRID			P1P2	11
direct BioGRID		P2P2		169
direct BioGRID		P2P2	P1P2	3
direct BioGRID	P1P1			152
direct BioGRID	P1P1		P1P2	7
direct BioGRID	P1P1	P2P2		65
direct BioGRID	P1P1	P2P2	P1P2	21
direct IntAct				2718
direct IntAct			P1P2	9
direct IntAct		P2P2		156
direct IntAct		P2P2	P1P2	3
direct IntAct	P1P1			172
direct IntAct	P1P1		P1P2	8

direct IntAct	P1P1	P2P2		49
direct IntAct	P1P1	P2P2	P1P2	17

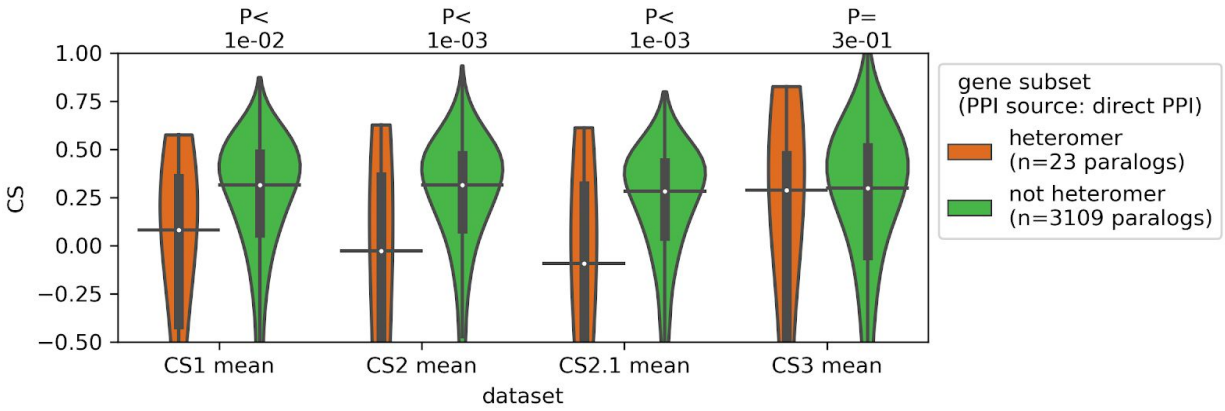
Appendix figures



Appendix Fig S1: Correlation between CS values across datasets.
Pairwise Spearman correlation coefficients (ρ) between CS values for the CRISPR screen datasets used in this study. Associated P-values are shown on the heatmap above the diagonal. The most significant correlation is for CS2 and CS2.1, which are CS values derived from the same experiments but with different sets of corrections. The significant correlation reflects that the corrections do not impact the relative ranking of genes.



Appendix Fig S2: Effect of gene LOF for singletons and paralogous genes across cell lines.
Unclassified genes are genes that are not in the paralog datasets but that were not identified as singletons in the stringent identification of singletons. Each point represents the mean CS for a class in an individual cell line (450 cell lines from CS2 dataset).

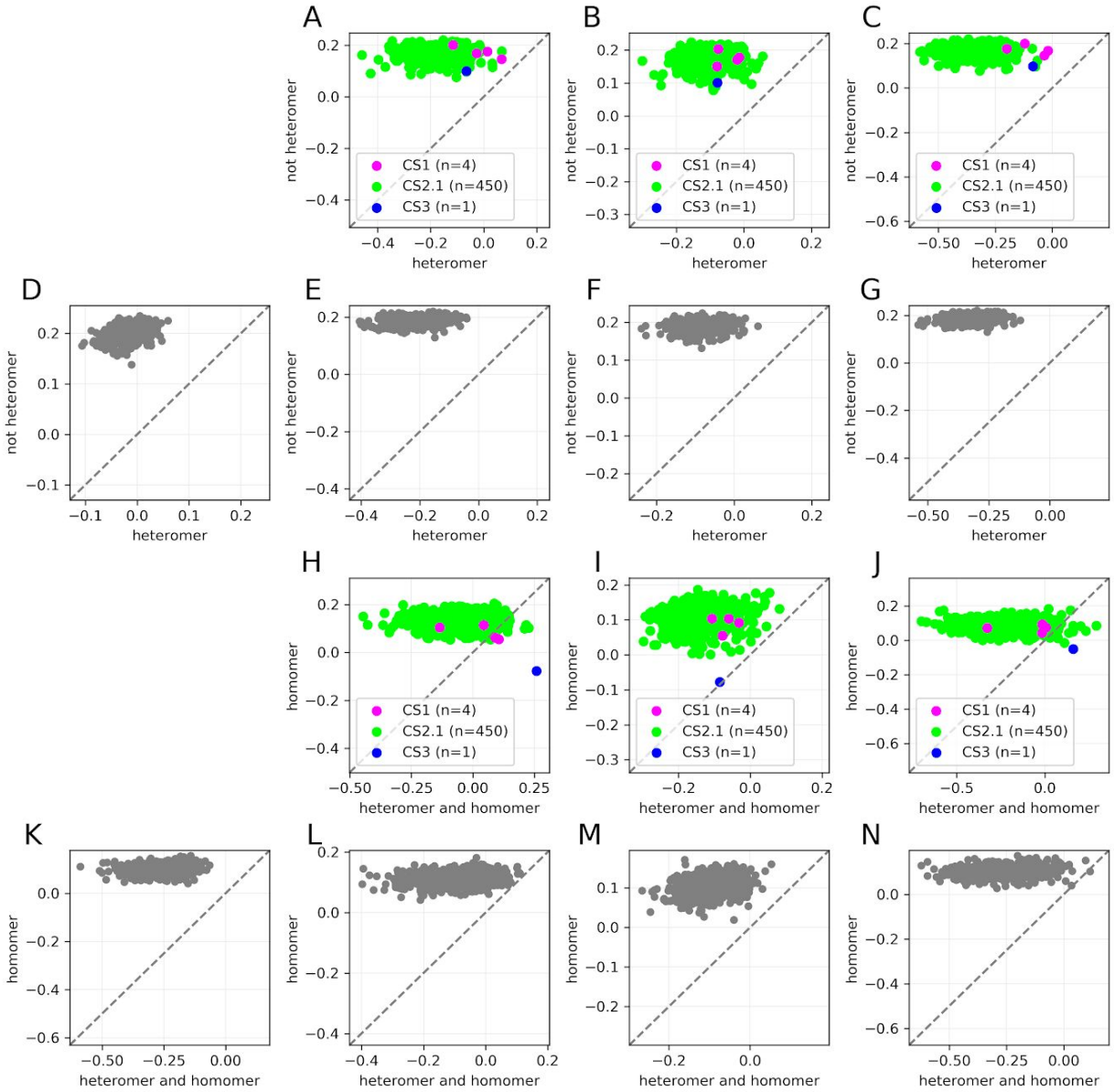


Appendix Fig S3: The LOFs of paralogous heteromers, defined by 'direct PPI's only, are relatively more deleterious than the LOFs of non-heteromers.

Similar analysis as that of Fig 2A.

LOF data derived from genome-wide CRISPR-Cas9 screening experiments. The effect of LOF is estimated by the depletion of gRNAs during the experiment, which reflects the deleteriousness of LOF on cell proliferation. The extent of depletion is measured as a CRISPR-score (CS). Relatively lower CS indicate relative more deleteriousness. CS values across cell lines from three biologically independent datasets — CS1 (Wang et al. 2015), CS2/CS2.1 (Meyers et al. 2017; DepMap 2018) and CS3 (Shifrut et al. 2018) are shown.

P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are by a horizontal black line and quartiles are indicated by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

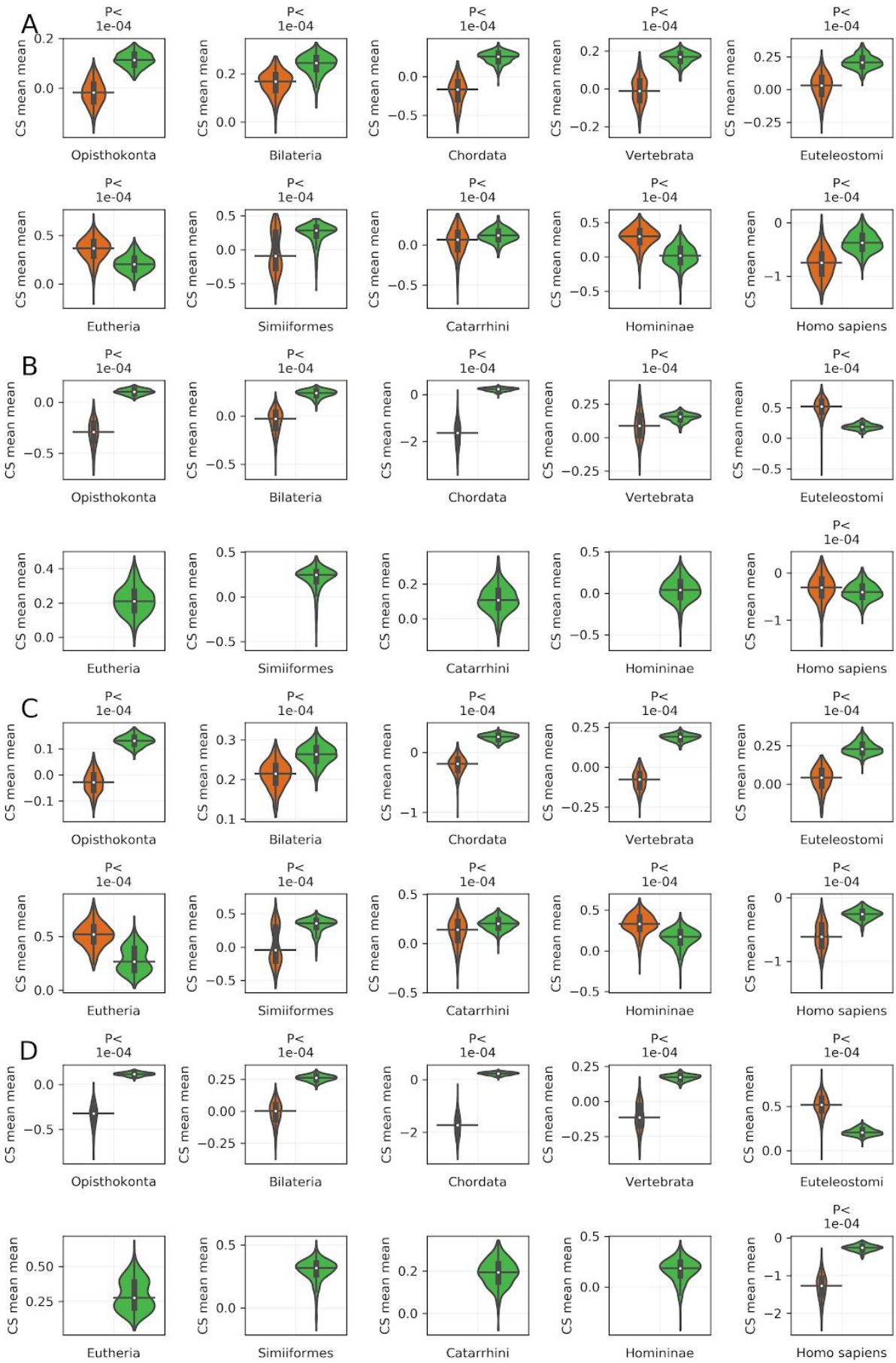


Appendix Fig S4: The LOFs of paralogous heteromers, are relatively more deleterious than the LOFs of non-heteromers, across most of the cell lines in the CS datasets.

Similar analysis as that of Fig 2B and 2C, but with 'direct PPI's and CS2 datasets.

The 1st and 2nd rows show the comparison between heteromers and non-heteromers, while the 2nd and 3rd rows show the comparison between heteromers that form homomers and homomers only. The 1st and and 2nd column show to the subsets of paralogs identified from BioGRID, while 3rd and 4th column show the subsets of paralogs identified from the IntAct. Analysis with the CS1, CS2.1 and CS3 is shown in the 1st and 3rd rows. Analysis with the CS2 is shown in 2nd and 4th rows. Analysis with subsets of paralogs (heteromers and homomers) defined by 'all PPI's is shown in the 1st and 3rd column while that with subsets of paralogs defined by 'direct PPI's is shown in 2nd and 4th column.

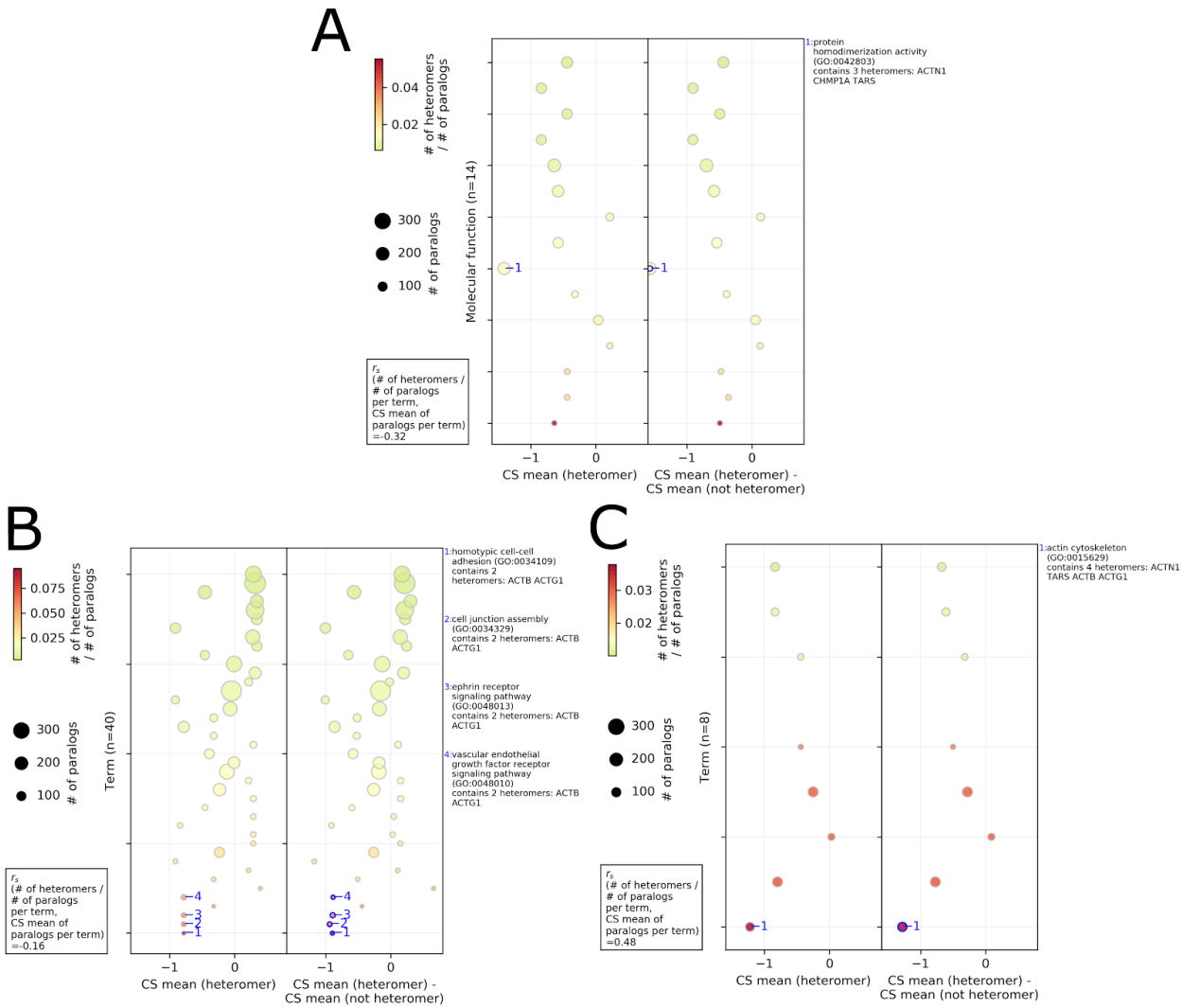
In all the cases, the mean CS values per cell line are well separated by the diagonal, indicating that the effects are systematic and independent of cell line.


A
B
C
D
P<
1e-04
CS mean mean
Opisthokonta
Bilateria
Chordata
Vertebrata
Euteleostomi
Eutheria
Simiiformes
Catarrhini
Homininae
Homo sapiens

Appendix Fig S5: Paralogs that form heteromers tend to be more deleterious upon LOF than other paralogs, largely independently from the age of paralogs.

The effect of LOF of heteromeric paralogs and non heteromeric paralogs on cell proliferation (CS) from CS2.1 dataset is shown. On the x axis, paralogs are ordered by the age in terms of dS bins (A and B) and age groups (C). The CS values per subset defined by class of paralogs (heteromer or not) and their age group is aggregated by taking the average across cell lines. Note that while heteromers are more deleterious in most of the age groups, the reverse trend is seen for a few cases.

The classification of paralogs in Panel A is based on all interactions while that in panel B and C is based on only direct interactions.
P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions is shown by a horizontal black line and quartiles are indicated by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

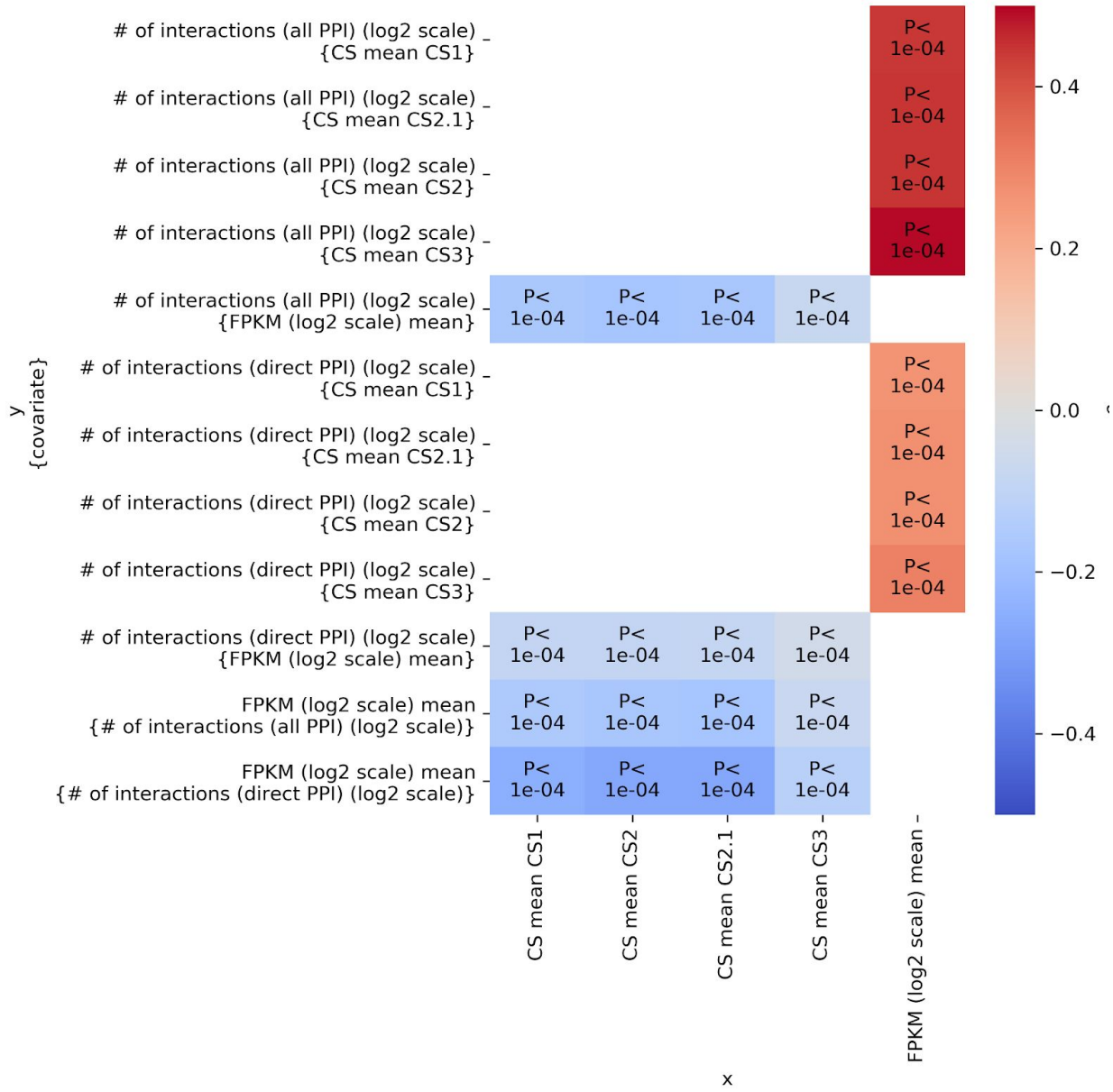


Appendix Fig S6: Association between the GO gene sets of paralogs, their probability of heteromerization and the effect of gene LOF on cell proliferation, in case of the heteromers defined by the 'direct PPI' only.

Gene set analysis for the Molecular Functions, Biological Processes and Cellular Components aspect are shown in the panel A to C respectively.

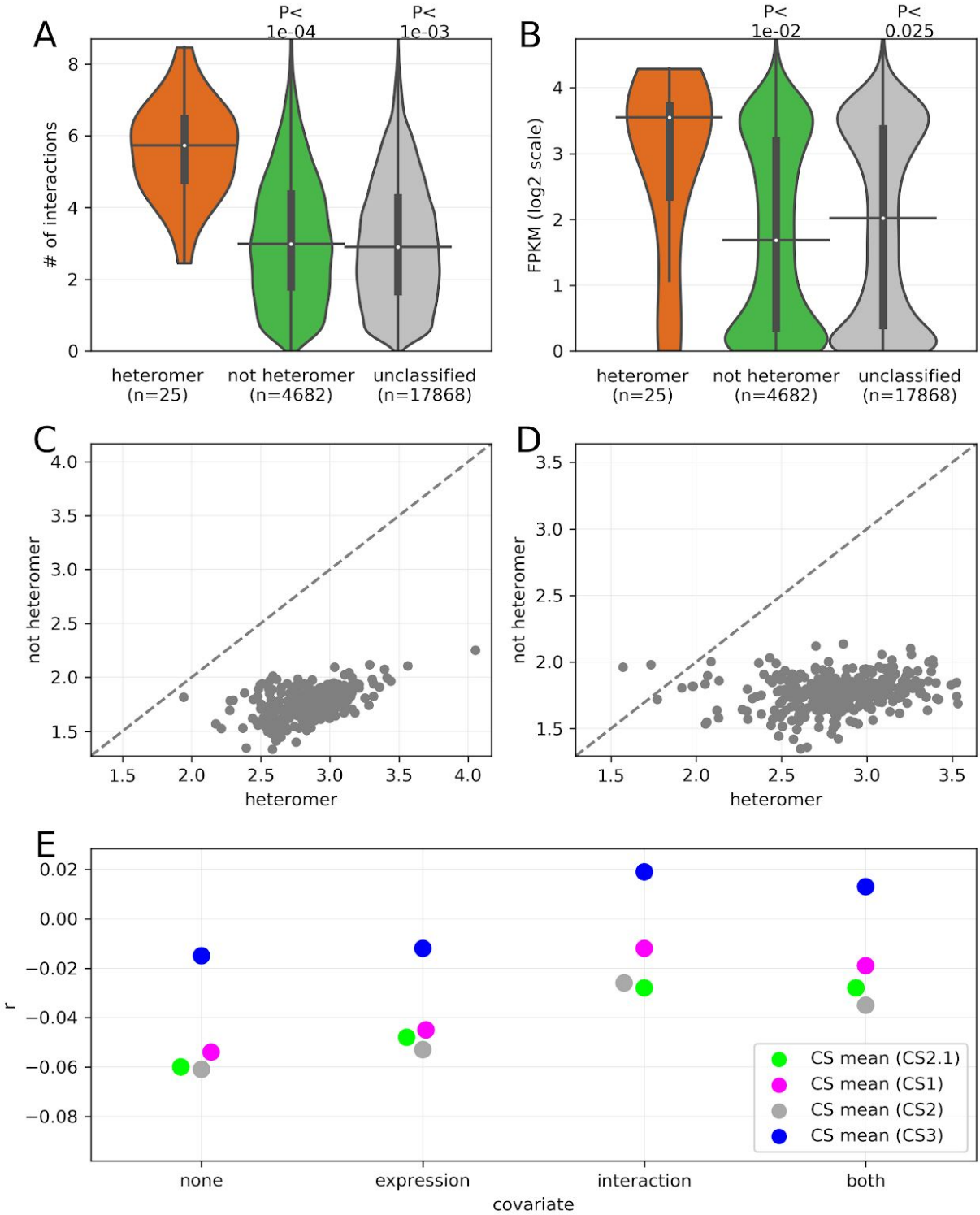
In each panel, average CS values (x-axis) of paralogs (heteromer or not heteromer) belonging to a gene set were used in the analysis. GO terms are sorted according to their proportion of heteromeric paralogs (i.e. # of heteromers/ # of paralogs). The size of the circles represent the number of paralog pairs in a category and the colors represent the proportion of heteromers in the category. The difference between the average CS value of the heteromers and average CS value of the non-heteromers is plotted in the right panel. The terms with significant difference between the average CS value of the heteromers and average CS value of the non-heteromers (estimated by two-sided t-test) are annotated with the blue edges. Descriptions of the representative significant GO terms with the highest difference are shown in the right side-panel. Spearman rank correlation between the proportion of the heteromers in the GO terms and the average CS value of paralogs in the term (r_s(# of heteromers / # of paralogs per term, CS mean of paralogs per term)) is shown in left right corner. Only GO molecular functions with more than 10% of the number of paralogs in all the gene sets are shown.

See Dataset EV4 for GO terms and annotations shown on this figure. Note that not all gene sets are independent because some genes are in several categories.



Appendix Fig S7: Correlations between the effect of LOF of a gene on cell proliferation, mRNA expression and number of protein-protein interaction partners.
Similar analysis as that of Fig 4A, but with individual CS datasets and direct PPI.

The effect of gene LOF on cell proliferation as measured in terms of CS values is correlated with mRNA expression and number of protein-protein interaction partners. Partial correlations were estimated in terms of Spearman correlation coefficients (ρ) between each pair of factors while controlling for the third factor (covariate, indicated in the curly brackets). The associated P-values are denoted on the heatmap.



Appendix Fig S8: Relationship between the effect of LOF of a gene on cell proliferation, mRNA expression and number of protein-protein interaction partners. Related to Fig 4.

A) Similar analysis as that of Fig 4B, but for heteromers identified with 'direct PPI' only. Paralogs that form heteromers have more interacting partners compared to

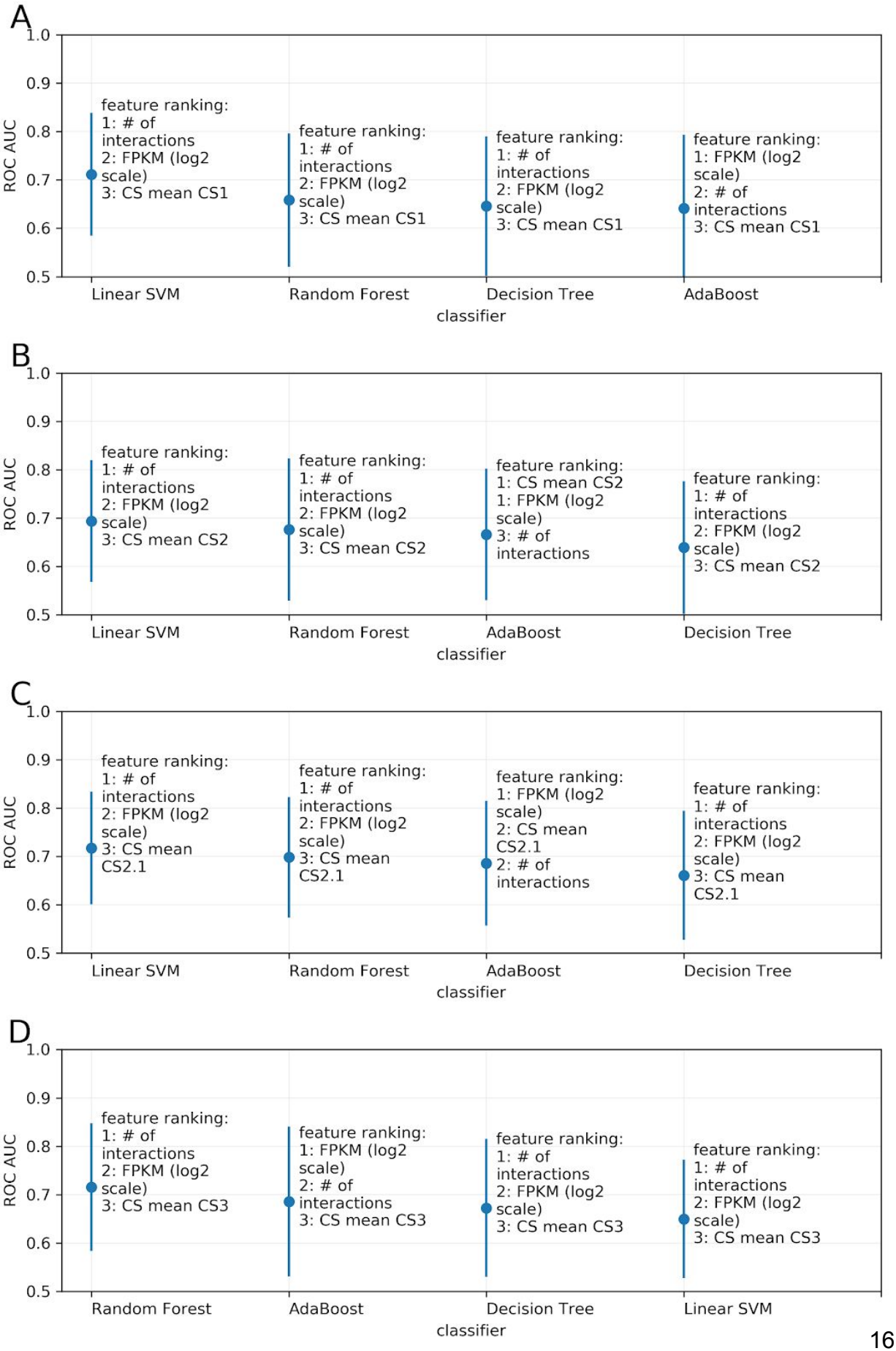
non-heteromers. Number of interactions are in log2 scale.

B) Similar analysis as that of Fig 4C, but in case of heteromers identified with 'direct PPI' only. Paralogs that form heteromers show higher expression than non-heteromers.

C) Cell-line wise comparison of mRNA expression between paralogous heteromers and paralogous non-heteromers identified from 'all PPI' (panel **C**) and 'direct PPI' (panel **D**). Similar analysis as that of Fig 2B, except here for mRNA expression values. Each point represents the mean FPKM (log2 scale) score for a class (heteromer or not heteromer) in an individual cell line (n=374). In each case the ~99% of the points are separated by the diagonal (dashed gray line) indicating cell-line independent systematic effects.

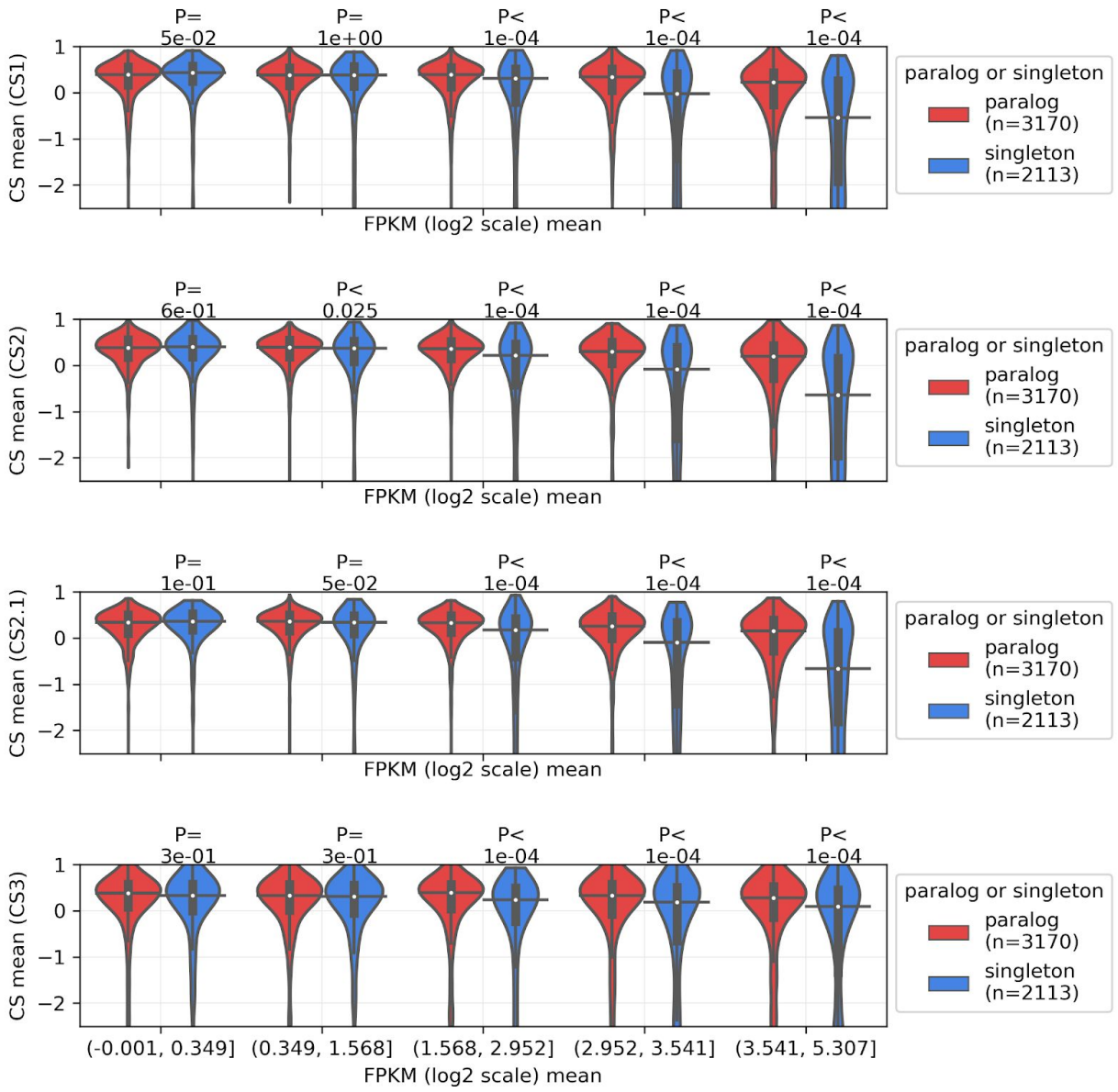
E) Similar analysis as that of Fig 4D, but in case of heteromers identified with 'direct PPI' only. Partial correlations were determined in terms Spearman correlation coefficients (r, shown on the y axis), between CS values and a paralog status (heteromer or not, binary variable, 1 : heteromer, 0 : not heteromer). The correlations were determined while controlling for none of mRNA expression and number of interactions, only mRNA expression, only number of interactions or both (as shown on the x axis).

In panels **A** and **B,** P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.


A
ROC AUC
1.0
0.9
0.8
0.7
0.6
0.5
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS1
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS1
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS1
feature ranking:
1: FPKM (log2 scale)
2: # of interactions
3: CS mean CS1
Linear SVM
Random Forest
Decision Tree
AdaBoost
classifier
B
ROC AUC
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2
feature ranking:
1: CS mean CS2
1: FPKM (log2 scale)
3: # of interactions
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2
Linear SVM
Random Forest
AdaBoost
Decision Tree
classifier
C
ROC AUC
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2.1
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2.1
feature ranking:
1: FPKM (log2 scale)
2: CS mean CS2.1
2: # of interactions
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS2.1
Linear SVM
Random Forest
AdaBoost
Decision Tree
classifier
D
ROC AUC
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS3
feature ranking:
1: FPKM (log2 scale)
2: # of interactions
3: CS mean CS3
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS3
feature ranking:
1: # of interactions
2: FPKM (log2 scale)
3: CS mean CS3
Random Forest
AdaBoost
Decision Tree
Linear SVM
classifier

Appendix Fig S9: Feature importance (shown on the y axis) determined through ensemble classification models (shown on the x axis), for 4 CS datasets.

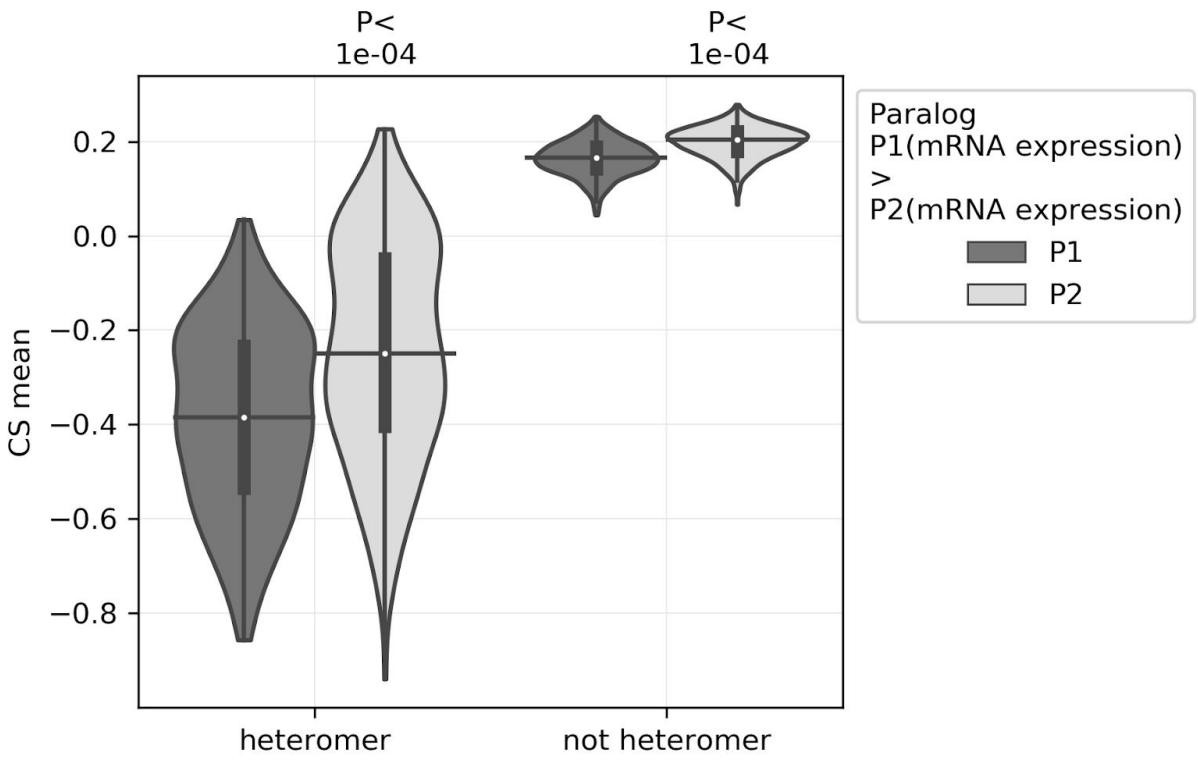
Similar analysis as that of Fig 4E. Analyses with CS datasets CS1, CS2, CS2.1 and CS3 are shown in panel A to D, respectively.



Appendix Fig S10: Dependence of the robustness of paralogs (shown on y-axis in terms of CS values) on mRNA expression (y-axis), across 4 CS datasets. Related to Fig EV3E.

mRNA expression of the genes was binned into 5 equal sized bins.

P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.

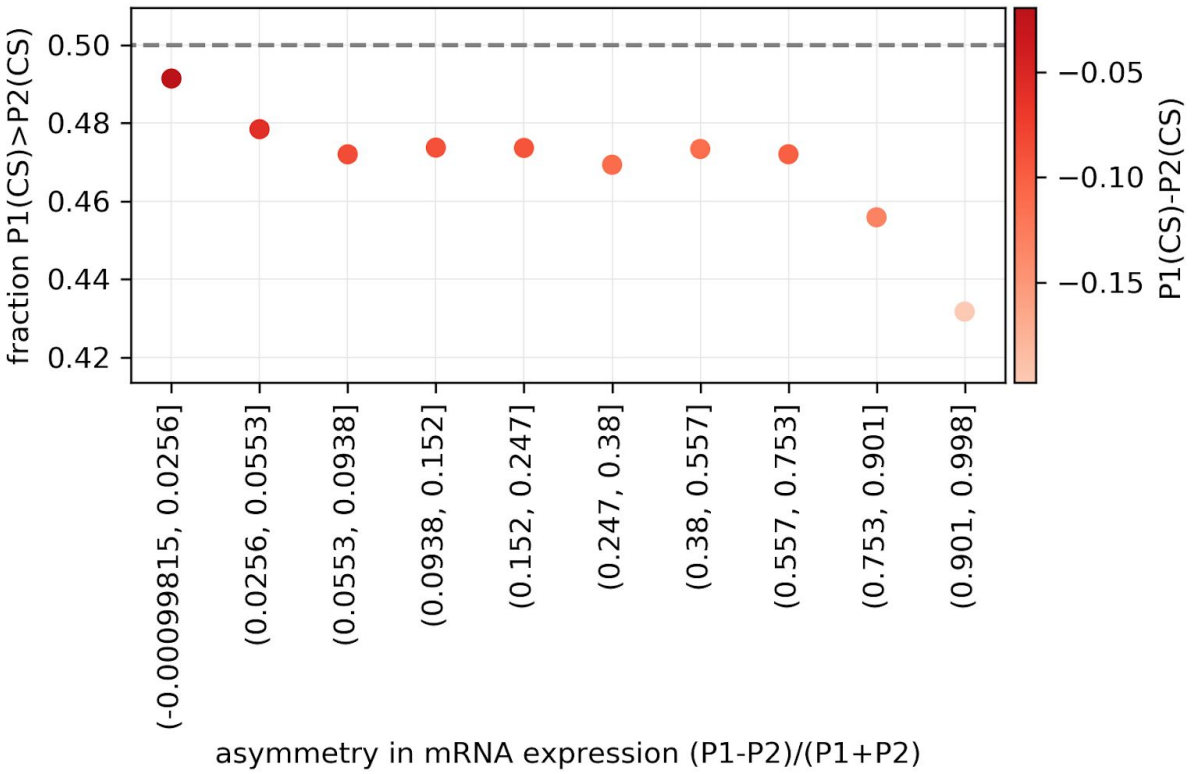


Appendix Fig S11: The most expressed paralog (P1) of a pair is more likely to be deleterious than the least expressed (P2), across 374 cell lines.

Similar analysis as that of Fig 6, but with 'direct PPI'.

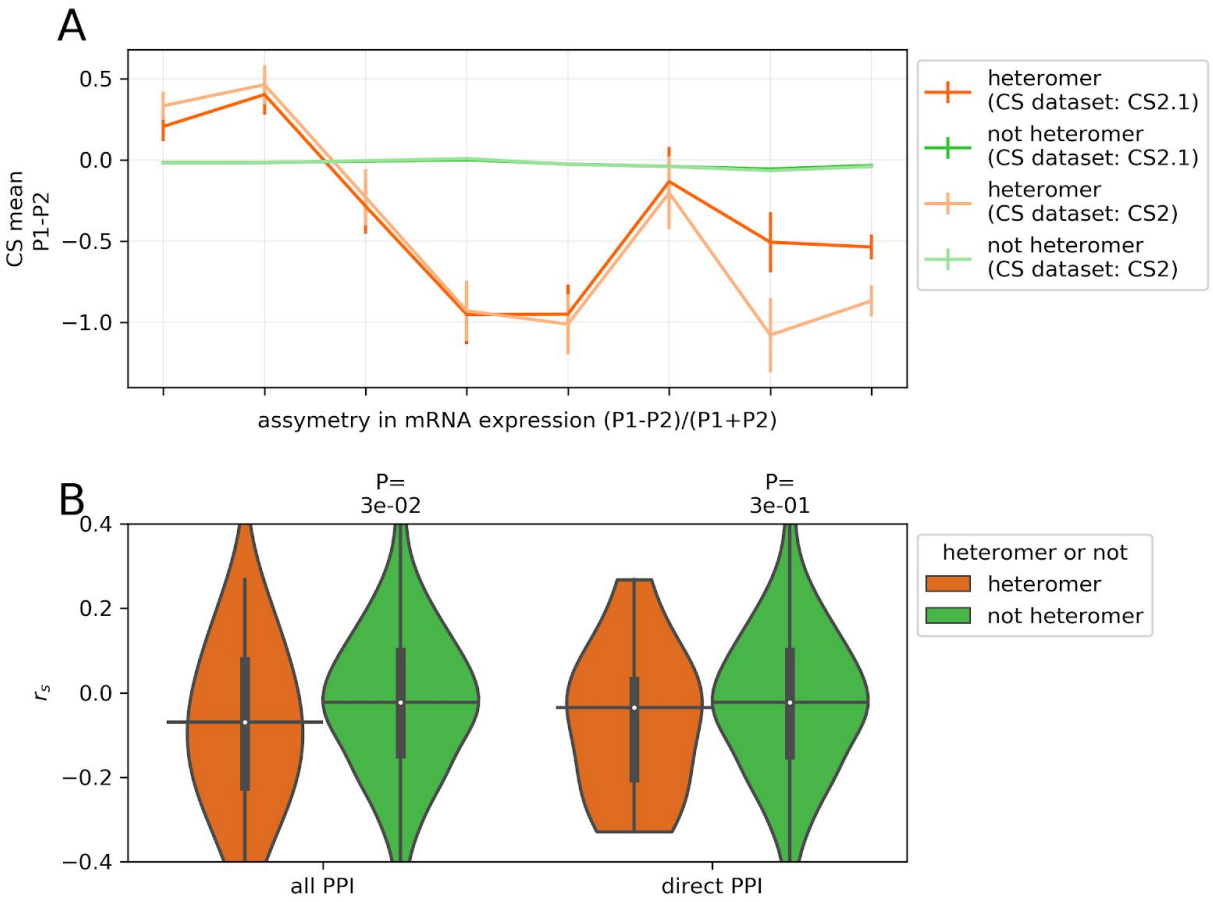
Each point represents CS value of an individual cell line.

P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.



Appendix Fig S12: Relationship between the asymmetry of expression and relative deleteriousness of paralog for the CS2 dataset.

The probability that a highly expressed paralog P1 has higher CS than comparatively weakly expressed paralog P2, as a function of its normalized relative mRNA expression to P2. Similar analysis as that of Fig EV5A, but with CS2 dataset.

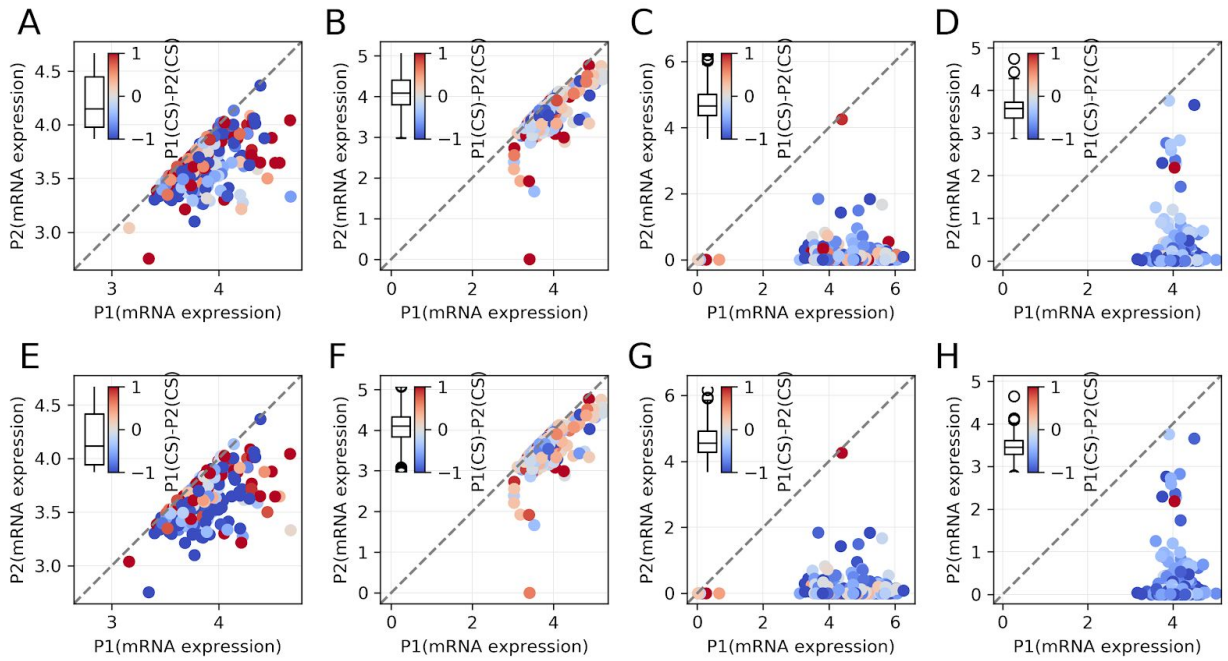


Appendix Fig S13: Relationships between asymmetry of the mRNA expression and difference in CS values. Related to Fig 6.

A) Relationship between the difference in CS of the paralog pair (P1-P2) and the asymmetry of mRNA expression levels i.e. (P1-P2)/(P1+P2), where mRNA expression of P1 is higher than P2. Values near 0 are cases in which the mRNA expression is symmetrical and asymmetrical for values near 1. The heteromers are defined by direct PPI'.

B) Average difference of CS value between P1 and P2 (P1(CS) -P2(CS)) is correlated with the asymmetry of mRNA expression (i.e. (P1-P2)/(P1+P2), where mRNA expression of P1 is greater than that of the P2), across cell lines. Similar analysis as Fig 4B, but with CS2 dataset.

P-values from two-sided Mann-Whitney U tests are shown. On the violin plots, the medians of the distributions are shown by a horizontal black line and quartiles by a vertical thick black line. For clarity, the upper and lower tails of the distributions are not shown.



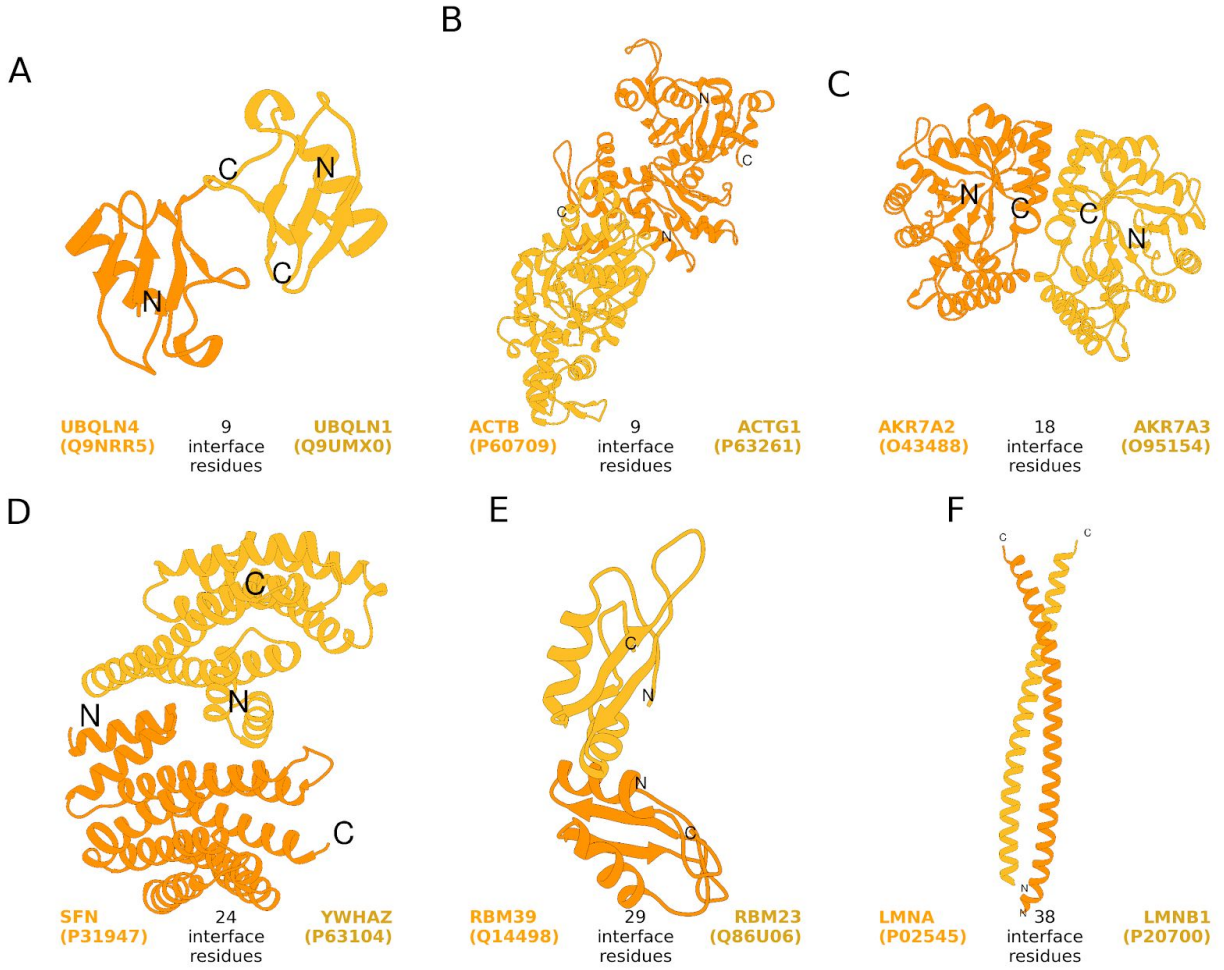
Appendix Fig S14: Relationships between the asymmetry of the mRNA expression and the difference in CS values is shown for representative heteromeric and non-heteromeric pairs of paralogs.

Each datapoint represents one cell line and is colored according to the difference of CS value for a pair. In the examples shown, the heteromers have more symmetrical expression levels than non heteromers. The results show that points away from the diagonal tend to be blue, showing that the asymmetry of expression varies per cell line and this generally correlate with the asymmetry of deleteriousness as well.

Representative heteromers: UBQLN1-UBQLN4 is shown in the 1st column of panels (panel A and E), while LMNA-LMNAB1 is shown in the 2nd column of panels (panel B and F).

Representative non-heteromers: SNX17-SNX31 is shown in the 3rd column of panels, while SPTA1-SPTAN1 is shown in the 4th column of panels.

The plots on the 1st row derive from the CS2.1 dataset, while those in the 2nd row derive from the CS2 dataset.



Appendix Fig S15: Structures of the representative heteromeric paralogs showing the number of interface residues.
Structures of the paralogs were obtained from Interactome3D (Mosca, Céol, and Aloy 2013) while the number of residues at the interaction interface were retrieved from Interactome INSIDER (Meyer et al. 2018). Uniprot ids of the paralogs are shown in brackets.

Appendix references

DepMap, Broad. 2018. “DepMap Achilles 18Q3 Public.” Figshare. https://doi.org/10.6084/M9.FIGSHARE.6931364.V1 [DATASET].

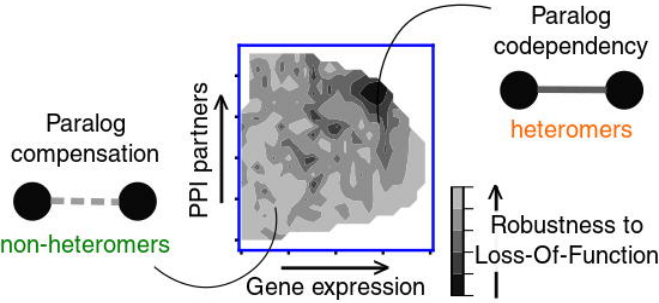
Meyer, Michael J., Juan Felipe Beltrán, Siqi Liang, Robert Fragoza, Aaron Rumack, Jin Liang, Xiaomu Wei, and Haiyuan Yu. 2018. “Interactome INSIDER: A Structural Interactome Browser for Genomic Studies.” *Nature Methods* 15 (2): 107–14.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. “Computational Correction of Copy Number Effect Improves Specificity of CRISPR–Cas9 Essentiality Screens in Cancer Cells.” *Nature Genetics* 49 (October): 1779.

Mosca, Roberto, Arnaud Céol, and Patrick Aloy. 2013. “Interactome3D: Adding Structural Details to Protein Networks.” *Nature Methods* 10 (1): 47–53.

Shifrut, Eric, Julia Carnevale, Victoria Tobin, Theodore L. Roth, Jonathan M. Woo, Christina T. Bui, P. Jonathan Li, Morgan E. Diolaiti, Alan Ashworth, and Alexander Marson. 2018. “Genome-Wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function.” *Cell* 175 (7): 1958–71.e15.

Wang, Tim, Kıvanç Birsoy, Nicholas W. Hughes, Kevin M. Krupczak, Yorick Post, Jenny J. Wei, Eric S. Lander, and David M. Sabatini. 2015. “Identification and Characterization of Essential Genes in the Human Genome.” *Science* 350 (6264): 1096–1101.


Paralog
compensation
non-heteromers
PPI partners
Gene expression
Paralog
codependency
heteromers
Robustness to
Loss-Of-Function
Robustness landscape