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

Summary

Gene duplicates provide protection against loss-of-function mutations. This protective redundancy partly relies on the fact that paralogs carry their functions independently, i.e. the inactivation of one gene copy does not impair the function of the other copy. 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 fitness landscape of gene loss-of-function by CRISPR-Cas9 in more than 450 human cell lines. Our analysis revealed a robustness landscape of human cells showing regions of higher vulnerability to gene inactivation as a function of key paralog properties. We find that heteromerizing paralogs are indeed less protective than non-heteromeric ones, but this association is largely due to their higher abundance and their larger number of protein-protein interaction partners.

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). Although 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)—paralogous genes nevertheless affect how biological systems globally respond to loss-of-function 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 (Gu et al. 2003). Similarly, 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 (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 loss-of-function at the molecular level.

The ability of paralogs to compensate for each other's loss is however not universal (Ihmels et al. 2007) and opposite results have been reported. For instance, Chen et al. (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 redundancy would be associated with greater susceptibility to disease mutations, contrary to the robustness expected from gene redundancy. A better understanding of if and how paralogs can compensate 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 a major obstacle in the development of drug treatments (Lavi 2015).

The mechanisms by which paralogs compensate for each other's loss-of-function 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 with important consequences, the occurrence of this phenomenon is however very likely limited. Indeed, a systematic assessment of this mechanism at the protein level 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 are able to 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 (Diss et al. 2017) is that paralogs appear as having 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 its 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 (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. Finally, Rajoo et al (Rajoo et al. 2018) examined the composition of the yeast nuclear pore complex and as (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 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 loss-of-function of the second, 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 by 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 loss-of-function mutations would be dependent on whether their products form heteromeric complexes with each other. 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 loss-of-function 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 loss-of-function on cell proliferation, mRNA abundance from more than 350 cell lines and protein-protein interactions (Figure S1 for description of all data used for the meta-analysis) revealed patterns which strongly support our hypothesis that paralogs that assemble are less protective, but through other factors than heteromerization. Our analyses also show that mRNA abundance is a key determinant of relative consequences of loss-of-function within pairs of paralogs.

Results

Paralogous genes protect against the effect of gene loss-of-function across all cell lines

We used two datasets of paralogous genes, one of relatively young paralogs for which we have divergence time estimates (Lan and Pritchard 2016) and another set of relatively old paralogs derived from whole-genome duplication (total 3842 pairs of paralogs, see Methods, Table S1). We first examined whether paralogous genes protect against the deleterious effects of loss-of-function mutations in set of 455 human cell lines from three independent studies. These experiments yield a CRISPR-score (CS) per gene as the relative depletion of gRNAs during the experiment, which reflects the deleteriousness of loss-of-function on cell proliferation. These datasets are (1) CS1 from four cell lines (n=4) (Wang et al. 2015), (2) CS2 from 450 cell lines (Meyers et al. 2017), (3) CS2.1 from 450 cell lines (DepMap 2018) and (4) CS3 from 1 primary cell line (Shifrut et al. 2018) (Figure S1A). Dataset CS2 takes into account copy-number variation effects in each cell line and is derived from the same data as CS2.1, therefore CS2 and CS2.1 are not biologically independent but represent two different methods of analysis (Meyers et al. 2017). 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 CS1, CS2.1 and CS3 datasets by considering only the gRNAs that uniquely align to the genome (see Methods). CS estimates for CS2 dataset were used directly as computed by the authors. CS values among datasets are correlated, indicating reproducible measurements of fitness effects across platforms, methodologies, cell lines and cell types (Figure S2).

As expected, we find that paralogs buffer the effect of gene loss-of-function. Genes that have paralogs have higher and less negative CS values than genes that are singletons (see methods for classification of singletons) for the three independent datasets considered (Figure 1A). To confirm that these effects were systematic and were not driven by few cases with strong effects, we compared the mean CS for paralogs and singletons across cell lines (Figure 1B and Figure S3A for 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 and strict singletons or to genes that are not in the set of paralogs nor classified as singletons (Figure 1C and Figure S3B for CS2 dataset). These results are therefore highly reproducible and cell line independent.

Older paralogs tend to be less protective

A subset of paralogs have been compared in terms of sequence divergence at silent sites (dS) that can be used to classify their timing of duplication (Lan and Pritchard 2016). We thus examined if the age of the paralogs could affect the fitness effect of their deletion. We find that, across all datasets, older paralogs tend to have more deleterious effects on cell proliferation (measured as average CS value for a pair) than younger paralogs (see Figure 1D for analysis with CS2.1 dataset, Figure S4 for analysis with other datasets). Also, within pairs, the most deleterious paralog (measured as minimum

CS value for a pair) is generally more deleterious as pairs get older (see Figure 1D for analysis with dataset CS2.1, Figure S4 for analysis with other datasets). The overall trend 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).

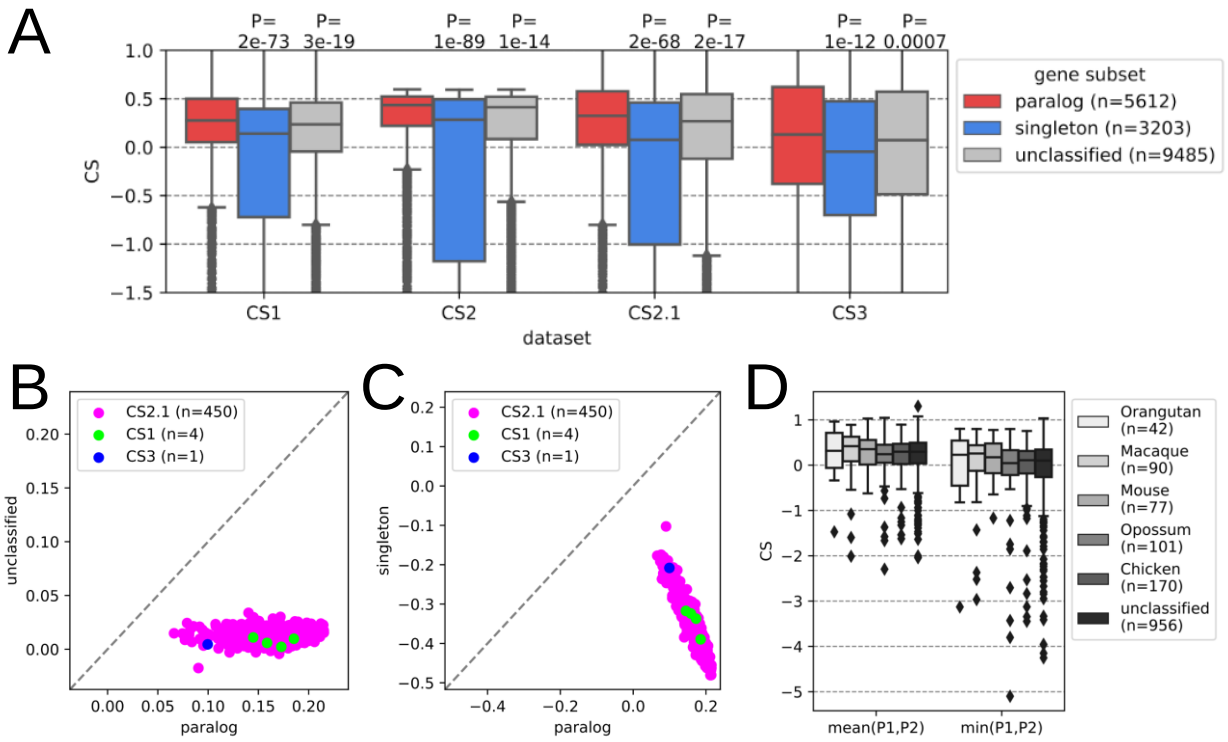


Figure 1: The loss-of-function of paralogs is less harmful than that of singletons in human cell lines.

A) Loss-of-function data derived from large-scale CRISPR-Cas9 screening experiments. The effect of loss-of-function is estimated by the depletion of gRNAs during the experiment, which reflects the deleteriousness of loss-of-function on cell proliferation. The extent of depletion is measured as a CRISPR-score (CS). Negative CS means that the genotype decreases in frequency through time. CS values across cell lines from three 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. For clarity, outlier values at the upper and lower tails of the distribution are not shown in the boxplots.

B) Comparisons of CS values between paralogs and singletons and **C)** between paralogs and unclassified genes (neither paralog nor 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 Figure S3.

D) Older paralogs tend to be less protective, i.e. more deleterious upon loss-of-function, than younger ones. The age groups are ordered in increasing distance of

phylogenetic node of duplication relative to human, starting from Orangutan to Chicken. The mean CS value of a pair (mean(P1,P2)) and minimum CS value of a pair (min(P1,P2)) is used as a measure of deleteriousness of the loss-of-function. CS values from dataset CS2.1 are shown in the plot. Similar plots for other datasets are shown in Figure S4.

Heteromeric paralogs derive 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 therefore first examined whether paralogous heteromers in humans likely derive from ancestral homomers and found that they indeed do, consistent with previous observations (Wagner 2003; Pereira-Leal et al. 2007). 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 derive from ancestral homomers if each paralog also form a homomer. Among the 3842 pairs of paralogs, 1141 physically interact (all PPIs considered) and at least 155 interact directly (direct PPI considered, see methods). Considering all PPIs, paralogs are 5.38 times more likely (67% versus 13%) to form heteromeric pairs (Fisher's exact test, p-value < 2.2e-16) if they also both form homomers than if none of them does. These proportions are 47% and less than 1% for direct heteromers (p-value < 2.2e-16) (see Table S4 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.

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

We looked at the effect of loss-of-function of paralogs that form heteromers and those that do not. Consistent with the dependency hypothesis, the loss-of-function of genes that heteromerize with their paralogs leads to larger effect on cell proliferation than the deletion of non-heteromeric ones in all datasets (Figure 2A, similar analysis with direct physical interactions is shown in Figure S5A). We also see that the effect is consistent across cell lines by looking at the mean CS of paralogous heteromers or non heteromers within each cell line (Figure 2B, similar analysis with dataset CS2 is shown in Figure S5B and analysis with direct physical interactions is shown in Figure S5C,D), with a majority of cell lines showing stronger effects for the deletion of paralogs forming heteromers. Similar analysis with paralogs that are both heteromers and homomers compared with paralogs which are only homomers show that interacting paralogous heteromers are relatively more deleterious (Figure 2C, similar analysis with dataset CS2 is shown in Figure S5E and with direct physical interactions is shown in Figure S5F,G). The effects are therefore not due to homomerization but due to heteromerization. These results support the hypothesis that interacting pairs of paralogs are less likely to buffer each other's loss-of-function.

One potential confounding factor with this analysis is the fact that the frequency of heteromers could covary with the age of the paralogs, which we showed above to affect

at least partially the fitness effect of gene loss-of-function. We therefore examined how the probability of being a heteromer ($P_{heteromer}$) covaries with the age of the paralogs. Using dS as a proxy for age, we find a relationship with age that tends to be positive, albeit non-monotonic (Figure 2D). We reasoned that this non-monotonicity could be due to the fact that young paralogs are largely indistinguishable in proteomics experiments and this method contributes a large fraction of protein interaction data (Livstone et al. 2011), which could lead to an overestimation of young paralogs among heteromeric ones.This effect is indeed absent when using direct physical interaction assays that do not rely on mass spectrometry protein identification (Figure 2D). These observations suggest that homomeric proteins that duplicate are more likely to be preserved than non-homomeric ones and thus more frequent among older pairs (see Figure S5H). The reason for this needs to be determined but this could be attributed to either this property directly or to other correlated ones such as protein abundance, which could correlate both with the probability of heteromeric assembly and with the probability of preservation (Jean-Francois Gout and Lynch 2015; Jean-François Gout et al. 2010).

We therefore looked at CS values of paralog loss-of-function corrected for age, by using the same bins and the age inferred by (Lan and Pritchard 2016). We observed that for all age groups, apart from the three youngest ones, CS values for heteromers are indeed lower than for non-heteromers, suggesting that this effect is independent from age (Figure 2E, similar plots with age (dS) and direct physical interactions is shown in Figure S6). The three youngest classes, which as mentioned above could be biased in terms of identification of heteromers, could also be biased because very young paralogs are more likely to show CRISPR-Cas9 cross-effects (i.e. double paralog knockout), leading to a potential overestimation of their fitness effects (Fortin et al. 2019). A similar effect is observed when considering age based on a phylogenetic classification (Lan and Pritchard 2016). Although the gRNA that multiply matched genomic locations were removed from CS1, CS2.1 and CS3, few mismatches could still be tolerated by Cas9 (Anderson et al. 2015), potentially leading to the co-deletion of paralogs for very young pairs.

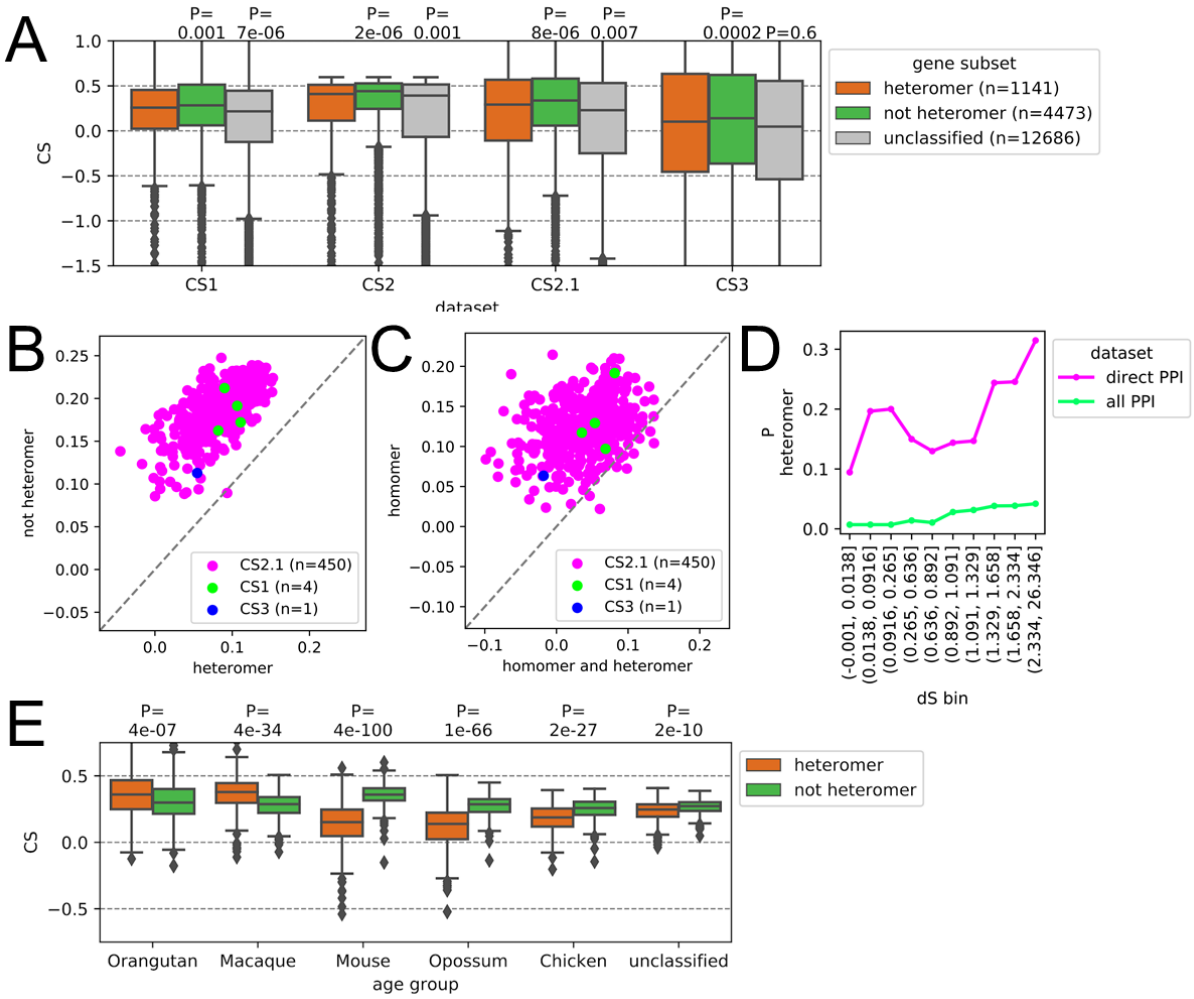


Figure 2. The loss-of-function of paralogs that form heteromers is more deleterious than the deletion of non-heteromers.

A) The effect of loss-of-function on cell proliferation (CS values) is on average more negative for paralogous heteromers than non-heteromers across all datasets. P-values of two-sided Mann-Whitney U tests are shown. Similar plot for heteromers detected with direct interactions only (see Methods) is shown in Figure S4A.

B) Mean CS values of paralogous heteromers and non-heteromers are shown across cell lines. Each point represents the mean CS value for a class in an individual cell line. All except one of the points are above the diagonal (dashed gray line), showing that the effect is systematic and largely cell-line independent. A similar plot for CS2 dataset is shown in Figure S5B and similar plot for heteromers detected with direct interactions only (see Methods) is shown in Figure S5C and D.

C) Similar to panel **B,** but comparing paralogs that form heteromers and homomers to those that form homomers only. This result shows that the difference between heteromers and non-heteromers is not caused by the fact that heteromers are also enriched for homomers. A similar plot for CS2 dataset is shown in Figure S5E and similar plot for heteromers and homomers detected with all interactions (see Methods) is shown in Figure S5F and G.

D) Paralogs that form heteromers tend to have been duplicated earlier in evolution. The probability of a paralog pair forming an heteromer is shown as a function of dS (proxy for age). Data is shown for interactions derived from all detection methods (1141 heteromers), and those that are more likely to detect direct interactions only (155 heteromers). Similar plot with paralogous homomers is shown in Figure S5H.

E) Paralogs that form heteromers tend to be more deleterious upon loss-of-function than other paralogs. Data from CS2.1 is shown. On the X axis, paralogs are ordered by their age. The CS values per class of paralogs (heteromer or not) and their age group are aggregated by taking average across cell lines. P-values are from two-sided Mann-Whitney U test. Similar plots are shown in terms of age of paralogs (dS) in Figure S6A and for heteromers detected with direct physical interactions only (see Methods) in Figure S6B and C.

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 found that some molecular functions are indeed enriched among heteromers of paralogs when considering all detection methods and methods that are specific to direct interactions (Table S5 and S6). For instance, proteins that have catalytic activity such as protein kinases (Table S5), which are known to directly regulate and thus interact with each other (Breitkreutz et al. 2010) are enriched. Other functions are depleted in paralogous heteromers, including G protein-coupled receptors (GPCRs) (Table S5). In the direct interaction set, paralogs that bind DNA and nucleic acids are enriched among heteromeric ones (Table S6), consistent with what is known about large families of dimeric transcription factors that evolve through duplication (Amoutzias et al. 2008) and that been shown to have co-dependent evolution (Baker, Hanson-Smith, and Johnson 2013). On the other hand, paralogs that are molecular transducers and transporters are depleted among heteromers (Table S6). The heteromerization of paralogs therefore correlates at least partly with some of their molecular functions.

We performed a systematic analysis to look for an association between heteromeric paralogs and molecular functions. We found that molecular functions that are enriched for heteromeric paralogs tend to have lower and more negative CS values (Spearman rho = -0.16, p-value = 3.082e-12, CS2.1), suggesting an association between molecular functions of paralogs, their probability of heteromerization and the impact of paralog deletion on cellular proliferation (Figure 3A, similar plots with other datasets are shown in Figure S7, S8 and S9, see Table S7 for ranking of GO terms shown in the plot). We illustrate this for four molecular functions at the extremes of this distribution. Paralogs that code for ion channels and SH3/SH2 adaptor activity have little deleterious effects on average when inactivated and rarely heterodimerize, whereas protein kinases and transcription factors often heterodimerize and their loss-of-function has strong effect on proliferation (Figure 3B).

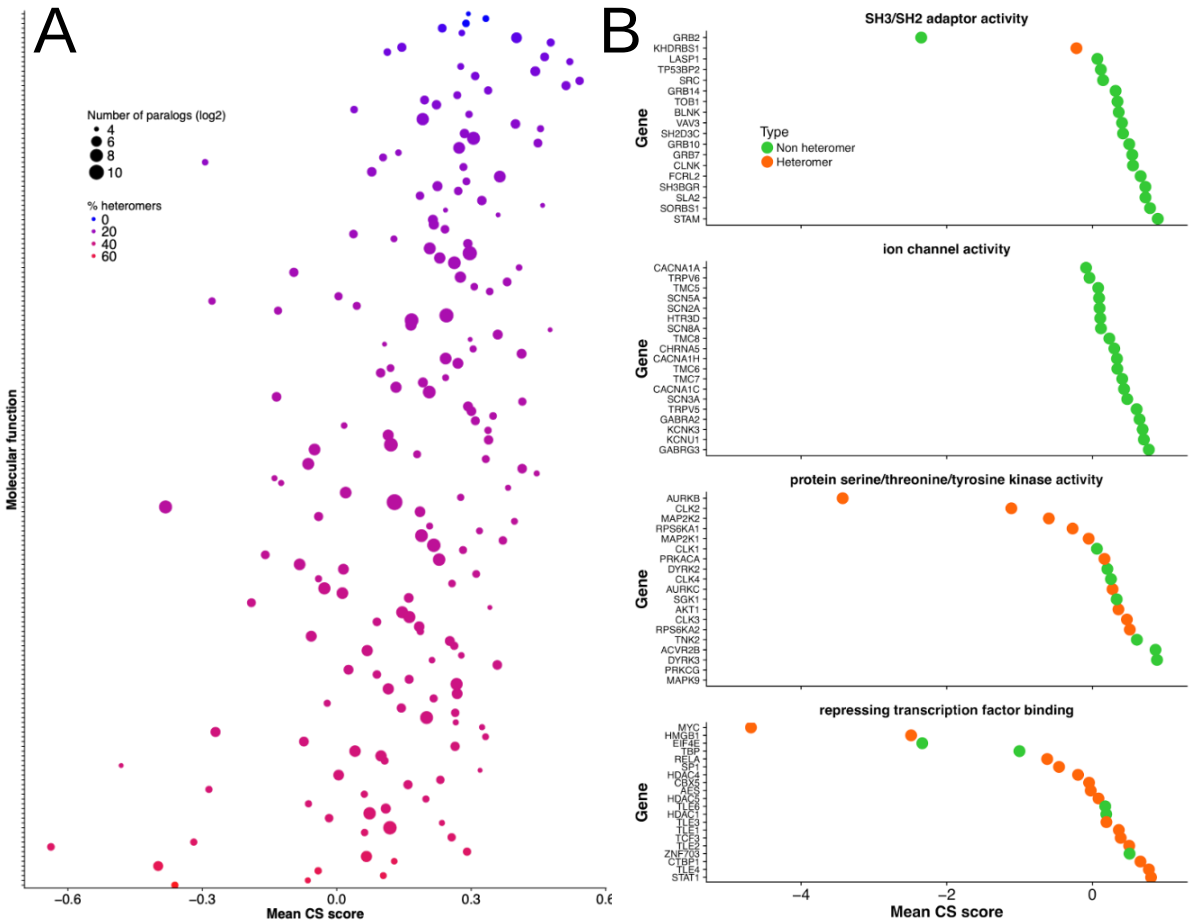


Figure 3. Association between the molecular functions of paralogs, their probability of heteromerization and the effect of gene loss-of-function on cell proliferation.

A) The average CS across cell lines (shown for CS2.1) per molecular function is negatively correlated with the percentage heteromeric paralogs per category (Spearman's rho = -0.16,p-value = 3.082e-12). Only GO molecular functions with more than 15 paralogs are shown. Molecular functions are sorted according to their proportion of paralogs that form heteromers. The size of the circles represent the number of pairs in a category and the color represent the percentage of heteromers in the category. Similar figures for the other datasets: CS1, Figure S7; CS2, Figure S8; and CS3, Figure S9. See Table S7 for ranking of GO terms shown on this figure. Note that not all molecular functions are independent because some genes are in several categories. Data derived form 5097 genes with GO molecular function annotation (Table S5).

B) Examples of paralogous genes for categories in which heteromers are rare (top) or frequent (bottom) to illustrate the general correlation between the fraction of heteromers in a molecular function and the effect of gene loss-of-function on cell proliferation.

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 more negative CS values. We first find that there is a strong negative correlation between the average CS values in the three independent datasets and the number of PPI partners of a protein (Table S7) and its mRNA abundance (Figure 4A, Data S2). As hypothesized, we also find that heteromers of paralogs are more highly expressed (Figure 4B, see Figure S10 for cell-line wise comparison) and have a larger number of PPI partners, both considering all interactions (Figure 4C) and direct interactions only (Figure S11). We therefore tested if heteromer formation has a significant effect on CS while taking expression levels and number of interaction partners into account. We found that for most expression levels and across the range of number of protein-protein interaction partners (Figures 4D and E), the larger effects of the loss-of-function of paralogs that form heteromers appears to be explained by their abundance and numbers of protein-interaction partners.

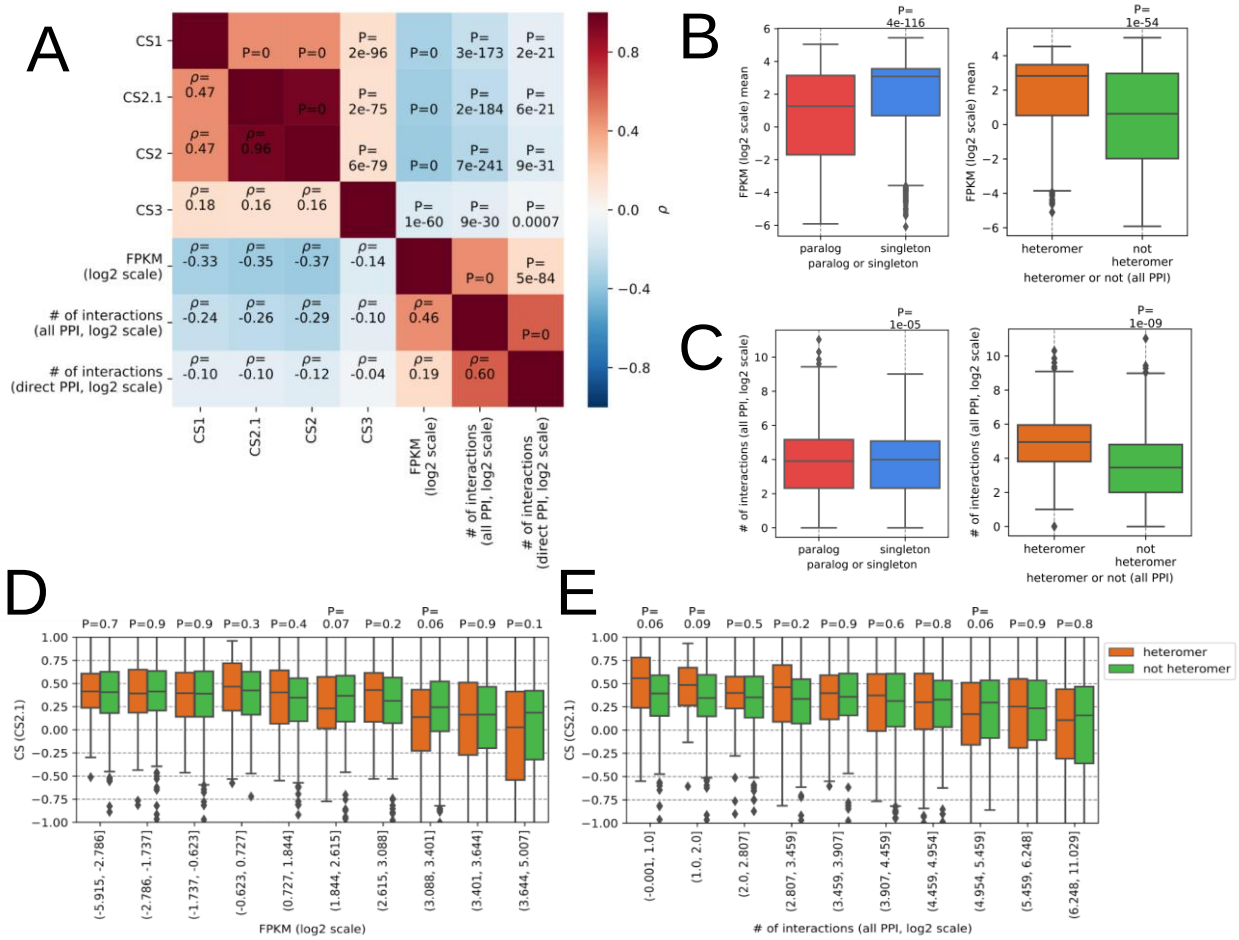


Figure 4: Relationship between the effect of loss-of-function of a gene on cell proliferation, mRNA abundance and number of protein-protein interaction partners.

A) The effect of gene loss-of-function on cell proliferation as measured by CS from different datasets is correlated with mRNA abundance and number of protein-protein interaction partners. Spearman correlation coefficients (ρ) between each pair of factors are shown on the heatmap. Correlation coefficient (shown below the diagonal) and the associated P-values (shown above the diagonal) are also shown. Cell-line wise comparison of the RNA expression across different classes is shown in Figure S10.

B) Paralogs have lower mRNA levels than singletons (left). Paralogs that form heteromers show higher expression than non-heteromers (right). mRNA abundance values are represented in terms of log10 scale FPKM scores. P-values for two-sided Mann-Whitney U tests are shown. Similar plot with heteromers of paralogs detected with only direct physical interactions is shown in Figure S11.

C) Similar to panel B but showing that the number of protein interaction partners is larger for paralogs than singletons. Paralogs that form heteromers have more interacting partners compared to non-heteromers (right). Similar plot with paralogous heteromers detected with only direct physical interactions is shown in Figure S11.

D) and **E)** The dependence of CS values of paralog classes on the mRNA abundance (panel **D**) and number of PPI interactions (panel **E**) is shown by binning mRNA levels and the number of protein interaction partners. Average CS for dataset CS2.1 and P-values from two-sided Mann-Whitney U tests are shown.

Heteromerizing paralogs occupy a space of the fitness landscape where gene loss-of-function is more deleterious

Overall, these results can be summarized in a fitness landscape of human cell lines in which the robustness against gene inactivation is shown as a function of the number of physical protein interaction partners and mRNA abundance. This clearly shows that strong negative effects are localized in the upper right corner, where expression is high and number of interaction partners is large (Figure 5A). The overlay on this landscape of the density of singletons and paralogs shows that they occupy a similar space, alhought singletons are in a space in which genes are slightly more expressed and have a slightly larger number of interactions (Figure 5B). Paralogous genes that form heteromers clearly occupy a space that is different from non-heteromeric ones, which brings them closer to a more deleterious parameter space in which genes are more expressed and have more protein interaction partners (Figure 5C). Similar plots with direct physical interactions are shown in Figure S12. We also show on the map representative pairs of heteromeric and non-heteromeric pairs of paralogs (Figure 5D). For instance, Ubiqulin 1 and 4 (UBQLN 1 and 4) form a heteromer, are highly expressed, have a large number of protein interaction partners and their loss of function is highly deleterious.

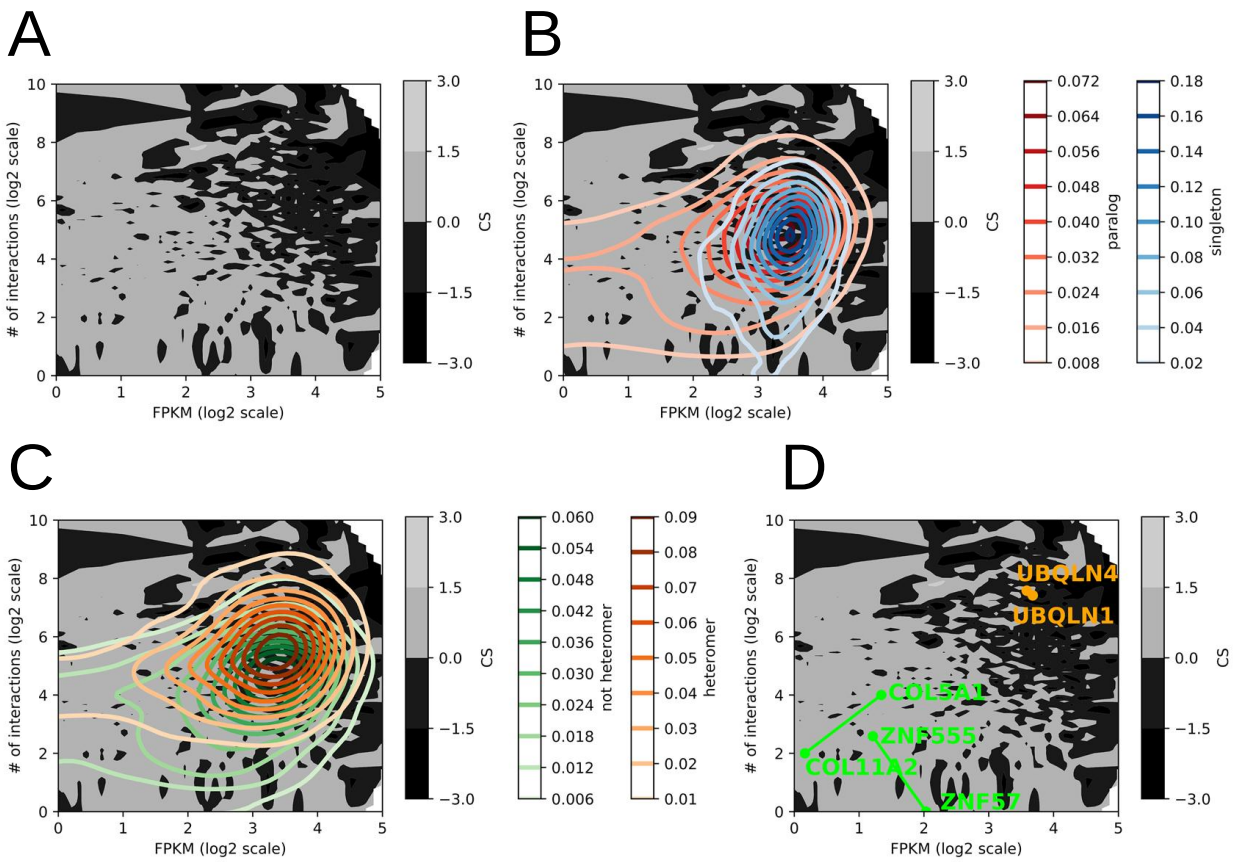


Figure 5: Robustness landscape of human cell lines shows regions of high vulnerability to paralog loss-of-function. mRNA abundance (log2(FPKM)) and number of protein-protein interaction partners (log2 scale) are strong determinants of the deleteriousness of paralog loss-of-function.

A) The landscape shows the effect of loss-of-function of paralogs 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 negative 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 towards lower expression levels and small numbers of protein interaction partners, compared to singletons.

C) Similar to B, paralogous heteromers and paralogous non-heteromers are overlaid on the landscape. The density of heteromers is biased towards higher expression levels and larger 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.

Similar plots with direct physical interactions only are shown in Figure S12.

The most expressed gene of a pair of paralogs has the strongest effect when deleted

Given that gene expression appears to be a major determinant of the fitness effect of loss-of-function, it is possible that the ability of a paralog to buffer for the loss of its sister copy depends on their relative abundance. For instance, when highly asymmetrically expressed, the loss-of-function of the most expressed gene of a pair is unlikely to be buffered by the least expressed one. However, when both are expressed at a comparable level, both would be expected to buffer each other in a similar manner. We examined whether asymmetry in mRNA abundance influences which gene in a pair is more deleterious upon loss-of-function (Figure 6A).

We found that across many cell lines, the paralog of a pair with the highest level of expression shows significantly more negative CS values (Figure 6B). We also found that, globally, the asymmetry of expression affects the asymmetry of fitness effect in a graded manner (Figure 6C,D) across the data. The difference of average CS value between paralogs correlates with their asymmetry of expression levels (Figure 6C). The fraction of pairs across cell lines for which the loss-of-function effect of a paralog (P1) is greater than the other (P2) is strongly associated with their asymmetry of expression (Figure 6D). This means that the loss of P2 is more deleterious when it is the most expressed and vice-versa. Interestingly, this effect appears to be stronger for heteromeric paralogs, which means that for the same extent of expression asymmetry, the difference in CS values is larger for heteromers (Figure 6C). The stronger association between extent of asymmetry of expression and asymmetry of CS values for heteromers may be counterbalanced overall by the fact that heteromeric paralogs are less likely to have asymmetrical expression (Figure S13), most likely due to the need for their stoichiometric balance imposed by their physical assembly. Some genes are exception to this general trend however (Figure 6E), showing that some pairs may have buffering relationships that are not driven by differences in abundances.

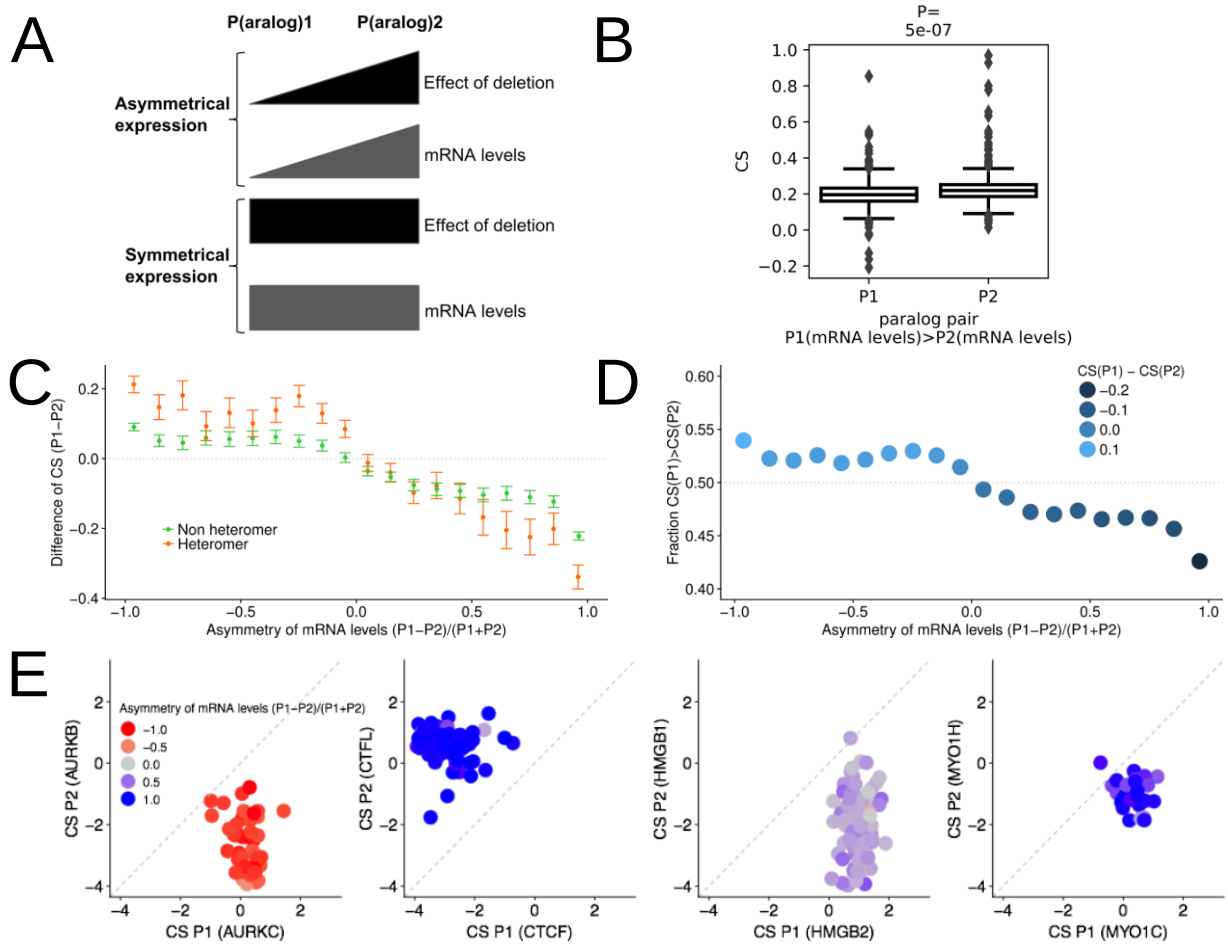


Figure 6: Asymmetry of expression levels between paralogs creates an asymmetry in their functional importance.

A) If expression level is a determinant of the functional consequence of gene deletion within a paralog pair, the most expressed one is expected to have a stronger deleterious effect upon loss-of-function. The least expressed paralog of a pair cannot compensate the loss of the most expressed one.

B) The most expressed paralog (P1) of a pair is more likely to be deleterious than the least expressed (P2) across 374 cell lines. Cell-line wise RNA expression data (represented in terms of log2 scale FPKM scores) for 374 cell lines and CS values for the same cell lines were used for the analysis. Each point represents CS value of an individual cell line. P-value is from two-sided Mann-Whitney U test.

C) Average difference of CS value between P1 and P2 as a function of asymmetry of expression. The 95% confidence interval are shown and the data is plotted for heteromers and non-heteromers. The CS2.1 dataset was used for this analysis. Expression level asymmetry was binned into 20 equal size bins.

D) The probability that a paralog P1 has higher CS than P2 as a function of its normalized relative expression to P2. Probabilities above 50% 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 P2 has higher expression. On the right, P1 is more likely to have a lower and more negative CS value (more deleterious) and

it is more expressed than P2. Expression level asymmetry was binned into 20 equal size bins. The color of the points represent the average difference of CS value in the bin.

E) Examples of pairs of paralogous genes with asymmetrical expression level (shown in the color of points) and with asymmetrical CS values. The two pairs on the left show result consistent with the overall trend and the two on the right show exceptions.

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 (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 thanks to large-scale CRISPR-cas9 screening (Wang et al. 2015). Here, using 3 large-scale datasets of gene loss-of-function that represent diverse cell lines and different experimental approaches (Figure S1A), we find that paralogs systematically contribute to cellular robustness across all cell lines (Figure 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. One of them 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 loss-of-function 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 loss-of-function 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 loss-of-function (Fortin et al. 2019). For this reason, we re-analyzed all data and considered only gene-specific gRNAs. 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 loss-of-function remains a difficult task.

We examined whether specific features of paralogous genes could affect their ability to buffer each other's loss-of-function 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 loss-of-function (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 loss-of-function mutations (Figure 2A). However, this is largely explained by higher expression levels and larger number of protein interaction partners for this class of paralogs (Figure 4 B

and C). On the robustness landscape outlined by the two factors (Figure 5A), expression levels and number of protein interaction partners clearly separate genes based on their deleteriousness. It also helps in understanding the major determinants of buffering effect of paralogs in general (Figure 5B) and the greater deleteriousness of heteromers (Figure 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. Dependency between paralogs could also evolve by other means than physical interactions. In addition, 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.

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 (Lan and Pritchard 2016) 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 loss-of-function variants are rarer in humans for proteins with large number of protein interaction partners (Karczewski et al. 2019). Here we observe the 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 (Figure 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 loss-of-function 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 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 of the 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 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 loss-of-function, 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) 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 because only their cumulative abundance is gauged by natural selection. Accordingly, functional divergence at the protein level would 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 (Figure 6B,C,D). If gene expression levels evolve at 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 loss of function mutation. Interestingly, we found that in general, heteromerizing pairs of paralogs have more symmetrical expression levels than non-heteromerizing ones. Heteromerization could slow down gene expression drifting and contribute to paralog maintenance, which could explain the tendency for the higher frequency of heteromers of paralogs among older pairs. Interestingly, for the same level of expression asymmetry, the difference of deleteriousness between paralogs is greater for heteromerizing ones.

Overall, our analyses show that not all paralogs are equally likely to buffer each other's loss-of-function in human cells. The underlying mechanisms for this ability remain to be fully understood beyond mRNA abundance and protein-protein interactions, and may depend on the specific function of the paralogs. 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 loss-of-function 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) (file BIOGRID-ORGANISM-3.5.167.tab2). While defining 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'.

Gene sets: paralogs and singletons

The set of human paralogous was obtained from (Lan and Pritchard 2016) and and is enriched for small-scale duplication events (1436 pairs associated with gene symbols). dS values and timing of duplication for these pairs were also obtained from this study. One gene pair had a dS value greater than 6 so it was set to 6 because larger values are likely saturated and non reliable. This set of paralogs was completed with a set of paralogs from whole-genome duplication events (obtained from http://ohnologs.curie.fr/ (Singh, Arora, and Isambert 2015)) using the strictest set (HUMAN.Pairs.Strict.2R.txt). As a complete set, 3842 non-redundant pairs of paralogs were used in the study (Table S1).

Pure singletons were identified using BLASTP (Altschul et al. 1990) searches of the human proteome (Homo_sapiens.GRCh38.unique.fa) against itself. Any protein that had no hits with E-value smaller than 0.001 over a segment longer that 0.6 times the smaller protein was considered as singleton. Gene symbols were used to merge the data from paralogs and protein-interaction data.

Gene sets: heteromers and homomers

From the protein-protein interactions obtained from BioGRID (Chatr-Aryamontri et al. 2015), heteromer of paralogs were identified as the pairs of paralogs that 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'. With 'all PPI' detection methods, 1141 heteromers and 394 homomers were identified. With 'direct PPI' detection methods, 155 heteromers and 326 homomers were identified.

CRISPR score dataset CS1 (Wang et al. 2015)

The CS values of set CS1 were derived from data from (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 quantile normalised as done for the original paper.The $\log_2$ fold change between final read counts and initial read counts was considered as CS value for each guide. CS values per gene were obtained by taking median of CS value for all guides that align to

the gene. Using the median allowed to avoid biases from extreme effects that could could from a single guide. The average of CS values per gene over cell line was then computed. CS1 dataset is included in Data S1.

CRISPR score dataset CS2 (DepMap 2018)

We also considered the recently published data (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 S2). CS values were obtained from the DepMap repository (DepMap 2018) (file name: gene_dependency.csv). As an aggregated CS value per gene for this dataset, the average CS value per gene over cell lines was computed. The CS2 dataset is included in Data S1.

CRISPR score dataset CS2.1 (DepMap 2018)

The CS2.1 dataset was generated by analysing data for 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 median value of fold change across guides per gene per cell line was used as gene score for a given cell line. As an aggregated CS value per gene for the dataset, the average of CS value per gene over cell lines was computed. CS2.1 dataset is included in Data S1.

CRISPR score dataset CS3

CS values for the CS3 dataset were obtained from a pooled screening in primary T-cells (Shifrut et al. 2018). gRNA wise $\log_2$ fold changes scores obtained from the study were first filtered to remove all the multi-hit gRNAs. CS values per gene were obtained by taking the median of CS value for all the guides for a particular gene. This dataset serves as an independent reference to the cancer or immortalized cell lines used in the other datasets. The CS3 dataset is included in Data S1.

Merging of CRISPR score datasets

For the comparative analysis of the 4 datasets, CS values in each dataset were first z-score normalized and individual datasets were merged by gene symbols. CS values from all the CS datasets are strongly correlated (Figure S2). Cell-line wise merged CS datasets are included in Data S1 and aggregated CS values are included in Table S3. Data S1 is available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-BSST233.

Gene expression

In order to obtained 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 CRISPR scores in the CS2 and CS2.1 datasets (see Table S2 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 overestimated 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, release:75) and $bamp.unique.bam is path to the BAM file containing only unique reads (as made in the preceding step). Cell-line wise RNA abundance dataset is included as Data S2. Data S2 is available in the BioStudies database (http://www.ebi.ac.uk/biostudies) under accession number S-BSST233.

GO enrichment analysis

The GO Molecular Function enrichment analysis was performed using PANTHER (Mi et al. 2017) online (http://pantherdb.org/geneListAnalysis.do) using the PANTHER Overrepresentation Test (Released 20181113) of PANTHER version 14.0 Released 2018-12-03. The reference set used was the list of all paralogs and the test set either the list of heteromeric paralogs from the “all PPI” data or from the “dir PPI”. Raw p-values from Fisher’s test and FDR corrected values are available in Table S5 and S6. The analysis of CS values per molecular function and fraction of heteromeric paralogs was performed using the GO molecular function terms obtained from PANTHER.

Acknowledgements

We thank the members of the Landrylab for discussions. We thank Carla Bautista, Johan Hallin, Angel Cisneros, Diana Ascencio, Ugo Dionne and Axelle Marchant for comments on the manuscript. 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

CRL and RD designed and performed research and wrote the paper. Equivalent contribution, authors shown in alphabetical order

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.

Anderson, Emily M., Amanda Haupt, John A. Schiel, Eldon Chou, Hidevaldo B. Machado, Žaklina Strezoska, Steve Lenger, et al. 2015. "Systematic Analysis of CRISPR-Cas9 Mismatch Tolerance Reveals Low Levels of off-Target Activity." *Journal of Biotechnology* 211 (October): 56–65.

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, 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.

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.

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

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.

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.

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.

Mi, Huaiyu, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas. 2017. "PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements." *Nucleic Acids Research* 45 (D1): D183–89.

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.

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.

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.

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.

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.
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.
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.
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.

Supporting information

Supplementary tables

Table S1: List of paralogs used in the study.
Table S2: Cell lines used in the study.
Table S3: CS datasets: aggregated across cell-lines.
Table S4: Number of homomeric (P1P1 or P2P2) and heteromeric (P1P2) paralogs.
Table S5: GO term enrichment analysis for heteromers detected with all PPI.
Table S6: GO term enrichment analysis for heteromers detected with only direct PPI.
Table S7: Number of physical interactions per gene.

Datasets

Data S1: Cell-line wise CS (CRISPR Scores) datasets (Biostudies Accession number: S-BSST233).
Data S2: Cell-line wise mRNA expression dataset (Biostudies Accession number: S-BSST233).

Supplementary figures

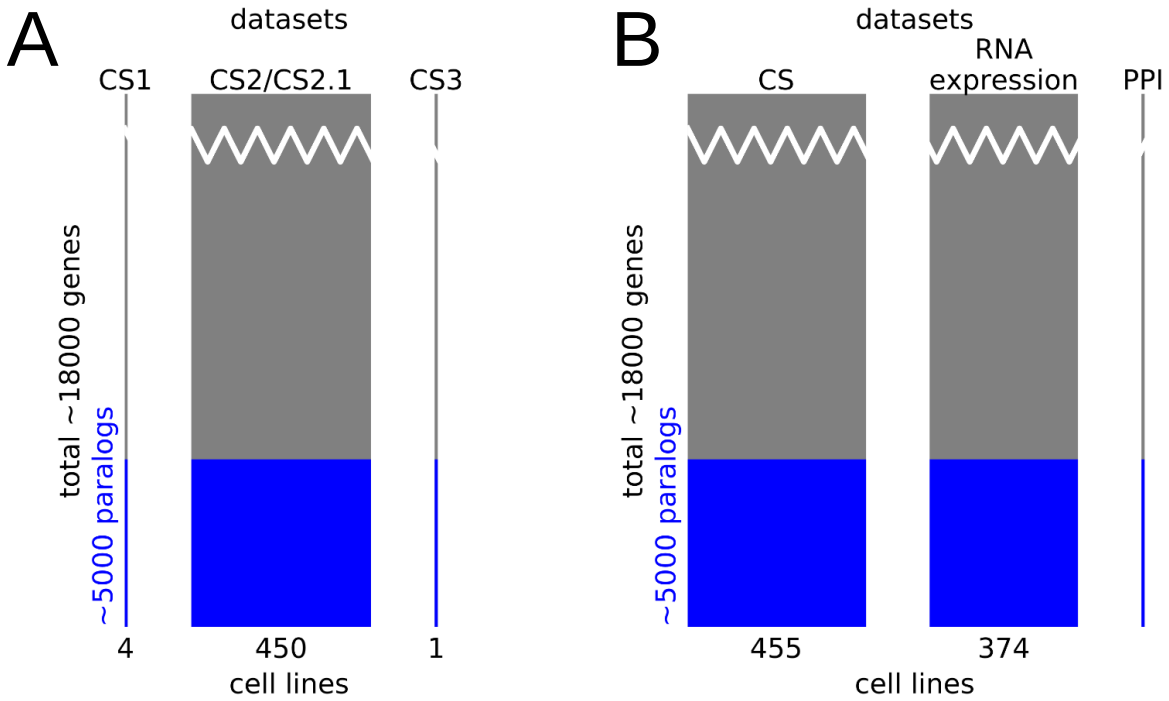


Figure S1: Scale of the data used in the study.

A CRISPR screen data from 3 independent studies — CS1 (Wang et al. 2015), CS2/CS2.1 (Meyers et al. 2017; DepMap 2018) and CS3 (Shifrut et al. 2018) — were used to generate the CS datasets.

B CRISPR screen data for l 455 cell lines, RNA expression data for 374 cell lines (Barretina et al. 2012) and protein-protein interaction data obtained from Biogrid (Livstone et al. 2011) were used for the meta-analysis in the study. CS = CRISPR score.

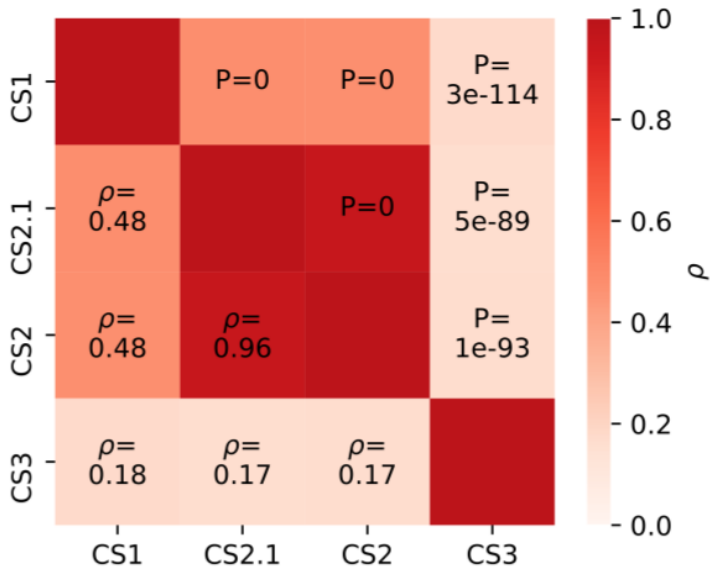


Figure S2: Correlation between CS values across datasets.

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

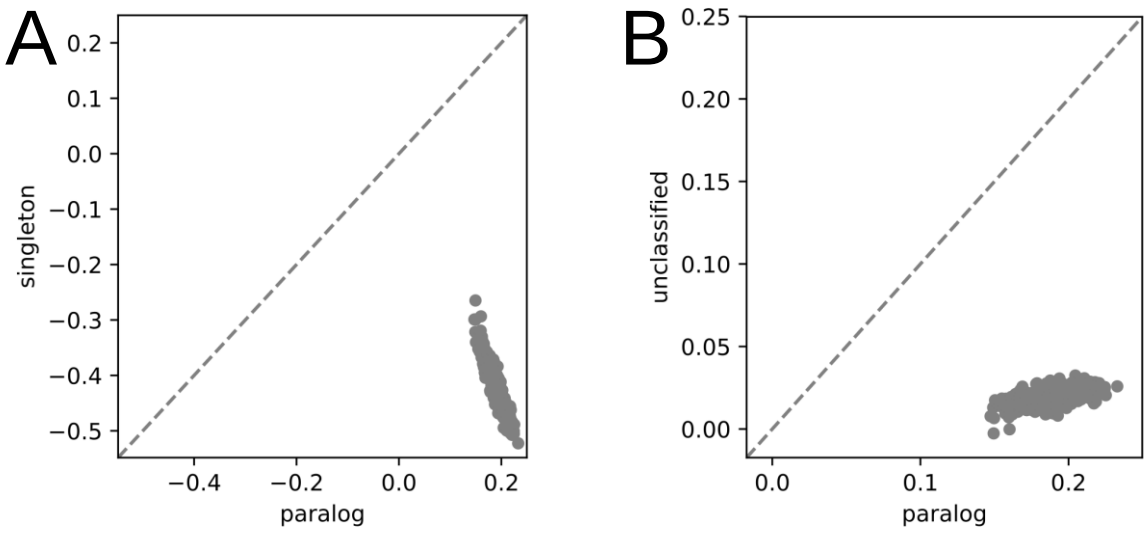


Figure S3: Effect of gene loss-of-function for singletons and paralogous genes across cell lines. Unclassified genes are genes that are not in the paralog datasets but that were not identified as singletons in the stringent identification of singletons. Each point represents the mean CS for a class in an individual cell line (450 cell lines from CS2 dataset).

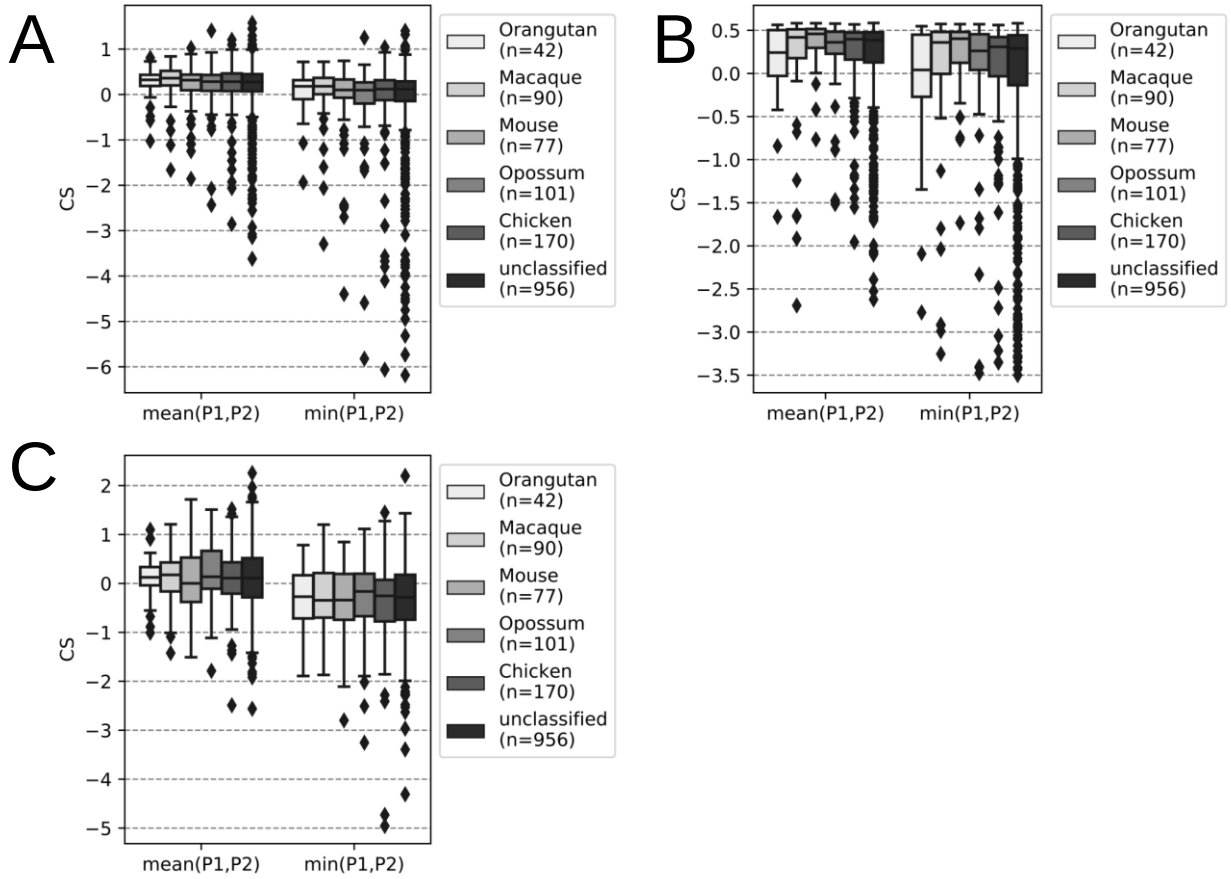


Figure S4: Older paralogs tend to be less protective i.e. more deleterious than younger ones. The age groups in the plot are ordered in increasing order from Orangutan to Chicken. The average CS value of the paralogs (mean(P1,P2)) and minimum CS value among the paralogs (min(P1,P2)) is used a measure of deleteriousness of the paralogs. The CS data from dataset CS1 (A), CS2 (B) and CS3 (C) is shown.

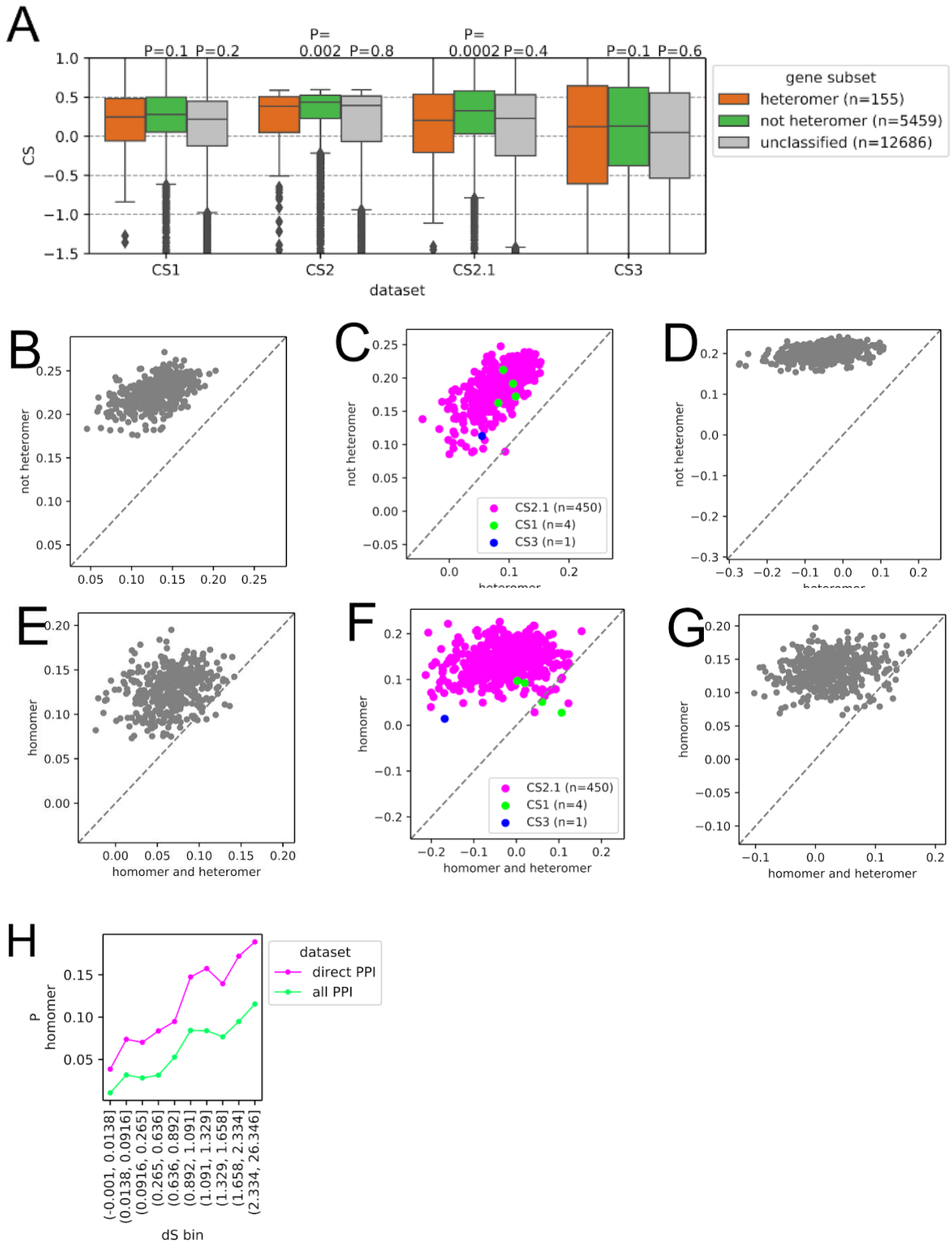


Figure S5. The deletion of paralogous heteromers (considering only direct physical interactions) is more deleterious than the deletion of non-heteromers and homomers.

A) Effects on cell proliferation (CS values) are on average more negative for paralogs that form heteromers than for paralogous that do not form heteromers across all datasets. P-values for two-sided Mann-Whitney U tests are shown.

B) Similar analysis as in Figure 2B, except that dataset CS2 is used. Mean CS of paralogs that form heteromers (defined from direct physical interactions) and paralogous that do not form heteromers are shown across cell lines. Each point represents the mean CS value for a class in an individual cell line.

C) Similar plot as Figure 2B except that classification of paralogs is based on direct interactions only.direct interactions.

D) Similar plot as B, except that classification is based on direct interactions only.

E) Similar analysis as in Figure 2C, except that dataset CS2 is used. Mean CS of paralogs that form heteromers (defined from direct physical interactions) and paralogous that do not form heteromers are shown across cell lines. Each point represents the mean CS value for a class in an individual cell line.

F) Similar plot as Figure 2C, except that classification is based on direct interactions only.

G) Similar plot as Figure E, except that classification is based on direct interactions only.

H) The probability of a paralog pair forming an homomer is shown as a function of dS (proxy for age). Data is shown for interactions derived from all detection methods (394 homomers), and those that are more likely to detect direct interactions only (326 homomers).

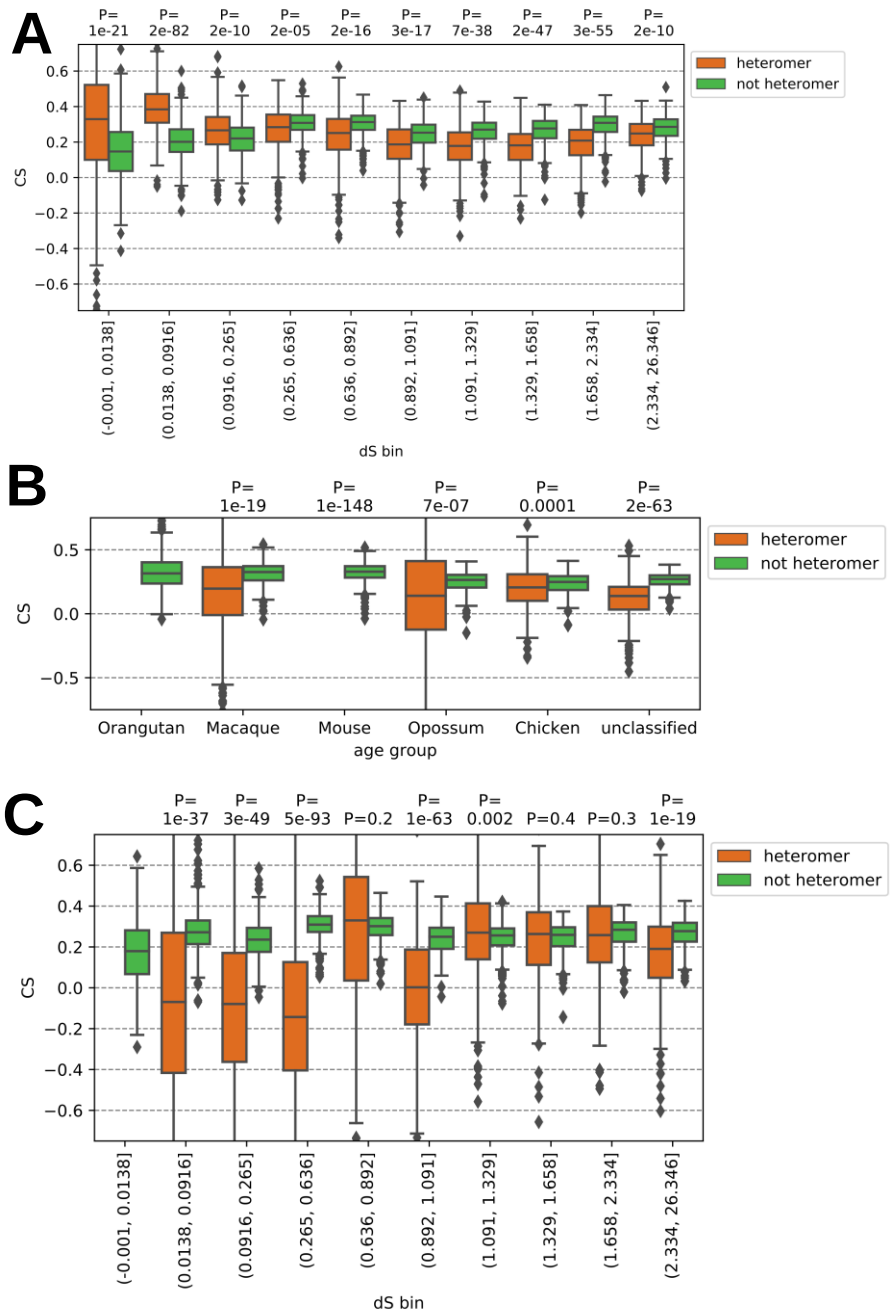
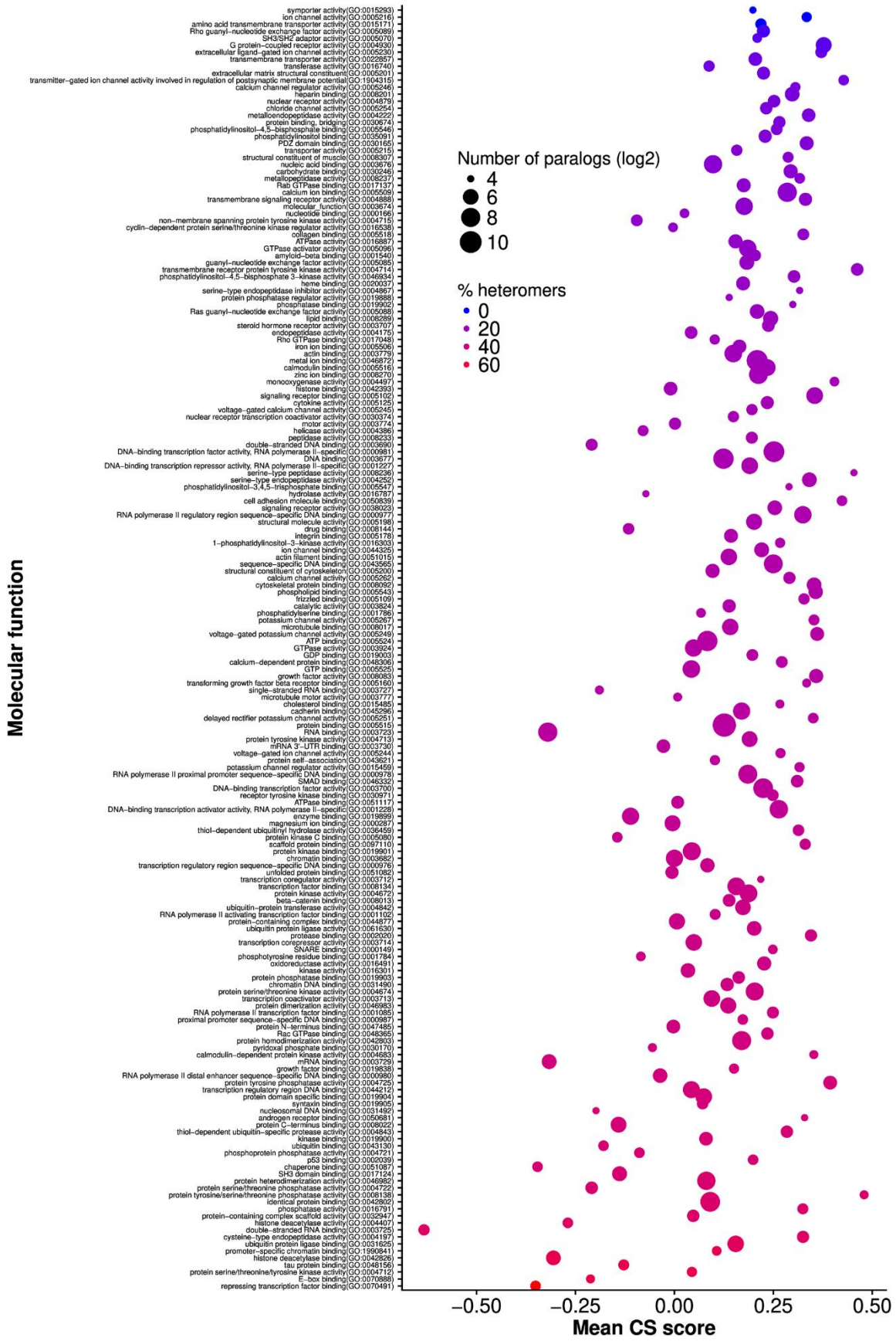


Figure S6. Paralogs that form heteromers tend to be more deleterious upon loss-of-function than other paralogs, systematically across cell lines. The effect of loss-of-function of heteromeric paralogs and non heteromeric paralogs on cell proliferation (CS) from CS2.1 dataset is shown. On the X axis, 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 (ds bins (**A and B**) or age group (**C**)) is aggregated by taking average across cell lines. P-values are from two-sided Mann-Whitney U tests. Classification of paralogs in Panel A is based on all interactions while that in panel B and C is based on only direct interactions.


Molecular function
symporter activity(GO:0015293)
ion channel activity(GO:0005216)
amino acid transmembrane transporter activity(GO:0015171)
Rho guanyl−nucleotide exchange factor activity(GO:0005089)
SH3/SH2 adaptor activity(GO:0005070)
G protein−coupled receptor activity(GO:0004930)
extracellular ligand−gated ion channel activity(GO:0005230)
transmembrane transporter activity(GO:0022857)
transferase activity(GO:0016740)
extracellular matrix structural constituent(GO:0005201)
transmitter−gated ion channel activity involved in regulation of postsynaptic membrane potential(GO:1904315)
calcium channel regulator activity(GO:0005246)
heparin binding(GO:0008201)
nuclear receptor activity(GO:0004879)
chloride channel activity(GO:0005254)
metalloendopeptidase activity(GO:0004222)
protein binding, bridging(GO:0030674)
phosphatidylinositol−4,5−bisphosphate binding(GO:0005546)
phosphatidylinositol binding(GO:0035091)
PDZ domain binding(GO:0030165)
transporter activity(GO:0005215)
structural constituent of muscle(GO:0008307)
nucleic acid binding(GO:0003676)
carbohydrate binding(GO:0030246)
metallopeptidase activity(GO:0008237)
Rab GTPase binding(GO:0017137)
calcium ion binding(GO:0005509)
transmembrane signaling receptor activity(GO:0004888)
molecular_function(GO:0003674)
nucleotide binding(GO:0000166)
non−membrane spanning protein tyrosine kinase activity(GO:0004715)
cyclin−dependent protein serine/threonine kinase regulator activity(GO:0016538)
collagen binding(GO:0005518)
ATPase activity(GO:0016887)
GTPase activator activity(GO:0005096)
amyloid−beta binding(GO:0001540)
guanyl−nucleotide exchange factor activity(GO:0005085)
transmembrane receptor protein tyrosine kinase activity(GO:0004714)
phosphatidylinositol−4,5−bisphosphate 3−kinase activity(GO:0046934)
heme binding(GO:0020037)
serine−type endopeptidase inhibitor activity(GO:0004867)
protein phosphatase regulator activity(GO:0019888)
phosphatase binding(GO:0019902)
Ras guanyl−nucleotide exchange factor activity(GO:0005088)
lipid binding(GO:0008289)
steroid hormone receptor activity(GO:0003707)
endopeptidase activity(GO:0004175)
Rho GTPase binding(GO:0017048)
iron ion binding(GO:0005506)
actin binding(GO:0003779)
metal ion binding(GO:0046872)
calmodulin binding(GO:0005516)
zinc ion binding(GO:0008270)
monooxygenase activity(GO:0004497)
histone binding(GO:0042393)
signaling receptor binding(GO:0005102)
cytokine activity(GO:0005125)
voltage−gated calcium channel activity(GO:0005245)
nuclear receptor transcription coactivator activity(GO:0030374)
motor activity(GO:0003774)
helicase activity(GO:0004386)
peptidase activity(GO:0008233)
double−stranded DNA binding(GO:0003690)
DNA−binding transcription factor activity, RNA polymerase II−specific(GO:0000981)
DNA binding(GO:0003677)
DNA−binding transcription repressor activity, RNA polymerase II−specific(GO:0001227)
serine−type peptidase activity(GO:0008236)
serine−type endopeptidase activity(GO:0004252)
phosphatidylinositol−3,4,5−trisphosphate binding(GO:0005547)
hydrolase activity(GO:0016787)
cell adhesion molecule binding(GO:0050839)
signaling receptor activity(GO:0038023)
RNA polymerase II regulatory region sequence−specific DNA binding(GO:0000977)
structural molecule activity(GO:0005198)
drug binding(GO:0008144)
integrin binding(GO:0005178)
1−phosphatidylinositol−3−kinase activity(GO:0016303)
ion channel binding(GO:0044325)
actin filament binding(GO:0051015)
sequence−specific DNA binding(GO:0043565)
structural constituent of cytoskeleton(GO:0005200)
calcium channel activity(GO:0005262)
cytoskeletal protein binding(GO:0008092)
phospholipid binding(GO:0005543)
frizzled binding(GO:0005109)
catalytic activity(GO:0003824)
phosphatidylserine binding(GO:0001786)
potassium channel activity(GO:0005267)
microtubule binding(GO:0008017)
voltage−gated potassium channel activity(GO:0005249)
ATP binding(GO:0005524)
GTPase activity(GO:0003924)
GDP binding(GO:0019003)
calcium−dependent protein binding(GO:0048306)
GTP binding(GO:0005525)
growth factor activity(GO:0008083)
transforming growth factor beta receptor binding(GO:0005160)
single−stranded RNA binding(GO:0003727)
microtubule motor activity(GO:0003777)
cholesterol binding(GO:0015485)
cadherin binding(GO:0045296)
delayed rectifier potassium channel activity(GO:0005251)
protein binding(GO:0005515)
RNA binding(GO:0003723)
protein tyrosine kinase activity(GO:0004713)
mRNA 3'−UTR binding(GO:0003730)
voltage−gated ion channel activity(GO:0005244)
protein self−association(GO:0043621)
potassium channel regulator activity(GO:0015459)
RNA polymerase II proximal promoter sequence−specific DNA binding(GO:0000978)
SMAD binding(GO:0046332)
DNA−binding transcription factor activity(GO:0003700)
receptor tyrosine kinase binding(GO:0030971)
ATPase binding(GO:0051117)
DNA−binding transcription activator activity, RNA polymerase II−specific(GO:0001228)
enzyme binding(GO:0019899)
magnesium ion binding(GO:0000287)
thiol−dependent ubiquitinyl hydrolase activity(GO:0036459)
protein kinase C binding(GO:0005080)
scaffold protein binding(GO:0097110)
protein kinase binding(GO:0019901)
chromatin binding(GO:0003682)
transcription regulatory region sequence−specific DNA binding(GO:0000976)
unfolded protein binding(GO:0051082)
transcription coregulator activity(GO:0003712)
transcription factor binding(GO:0008134)
protein kinase activity(GO:0004672)
beta−catenin binding(GO:0008013)
ubiquitin−protein transferase activity(GO:0004842)
RNA polymerase II activating transcription factor binding(GO:0001102)
protein−containing complex binding(GO:0044877)
ubiquitin protein ligase activity(GO:0061630)
protease binding(GO:0002020)
transcription corepressor activity(GO:0003714)
SNARE binding(GO:0000149)
phosphotyrosine residue binding(GO:0001784)
oxidoreductase activity(GO:0016491)
kinase activity(GO:0016301)
protein phosphatase binding(GO:0019903)
chromatin DNA binding(GO:0031490)
protein serine/threonine kinase activity(GO:0004674)
transcription coactivator activity(GO:0003713)
protein dimerization activity(GO:0046983)
RNA polymerase II transcription factor binding(GO:0001085)
proximal promoter sequence−specific DNA binding(GO:0000987)
protein N−terminus binding(GO:0047485)
Rac GTPase binding(GO:0048365)
protein homodimerization activity(GO:0042803)
pyridoxal phosphate binding(GO:0030170)
calmodulin−dependent protein kinase activity(GO:0004683)
mRNA binding(GO:0003729)
growth factor binding(GO:0019838)
RNA polymerase II distal enhancer sequence−specific DNA binding(GO:0000980)
protein tyrosine phosphatase activity(GO:0004725)
transcription regulatory region DNA binding(GO:0044212)
protein domain specific binding(GO:0019904)
syntaxin binding(GO:0019905)
nucleosomal DNA binding(GO:0031492)
androgen receptor binding(GO:0050681)
protein C−terminus binding(GO:0008022)
thiol−dependent ubiquitin−specific protease activity(GO:0004843)
kinase binding(GO:0019900)
ubiquitin binding(GO:0043130)
phosphoprotein phosphatase activity(GO:0004721)
p53 binding(GO:0002039)
chaperone binding(GO:0051087)
SH3 domain binding(GO:0017124)
protein heterodimerization activity(GO:0046982)
protein serine/threonine phosphatase activity(GO:0004722)
protein tyrosine/serine/threonine phosphatase activity(GO:0008138)
identical protein binding(GO:0042802)
phosphatase activity(GO:0016791)
protein−containing complex scaffold activity(GO:0032947)
histone deacetylase activity(GO:0004407)
double−stranded RNA binding(GO:0003725)
cysteine−type endopeptidase activity(GO:0004197)
ubiquitin protein ligase binding(GO:0031625)
promoter−specific chromatin binding(GO:1990841)
histone deacetylase binding(GO:0042826)
tau protein binding(GO:0048156)
protein serine/threonine/tyrosine kinase activity(GO:0004712)
E−box binding(GO:0070888)
repressing transcription factor binding(GO:0070491)
Number of paralogs (log2)
4
6
8
10
% heteromers
0
20
40
60
−0.50
−0.25
0.00
0.25
0.50
Mean CS score

1

Figure S7. GO analysis with CS1

The average CS across cell lines per molecular function is negatively correlated with the heteromeric paralogs per category. Only GO molecular functions with more than 15 paralogs are shown. Molecular functions are sorted according to their proportions of paralogs that form heteromers. The size of the circles represent the number of paralogous pairs in a category and the color represents the percentage of heteromers in the category. See Table S7 for ranking of GO terms shown.

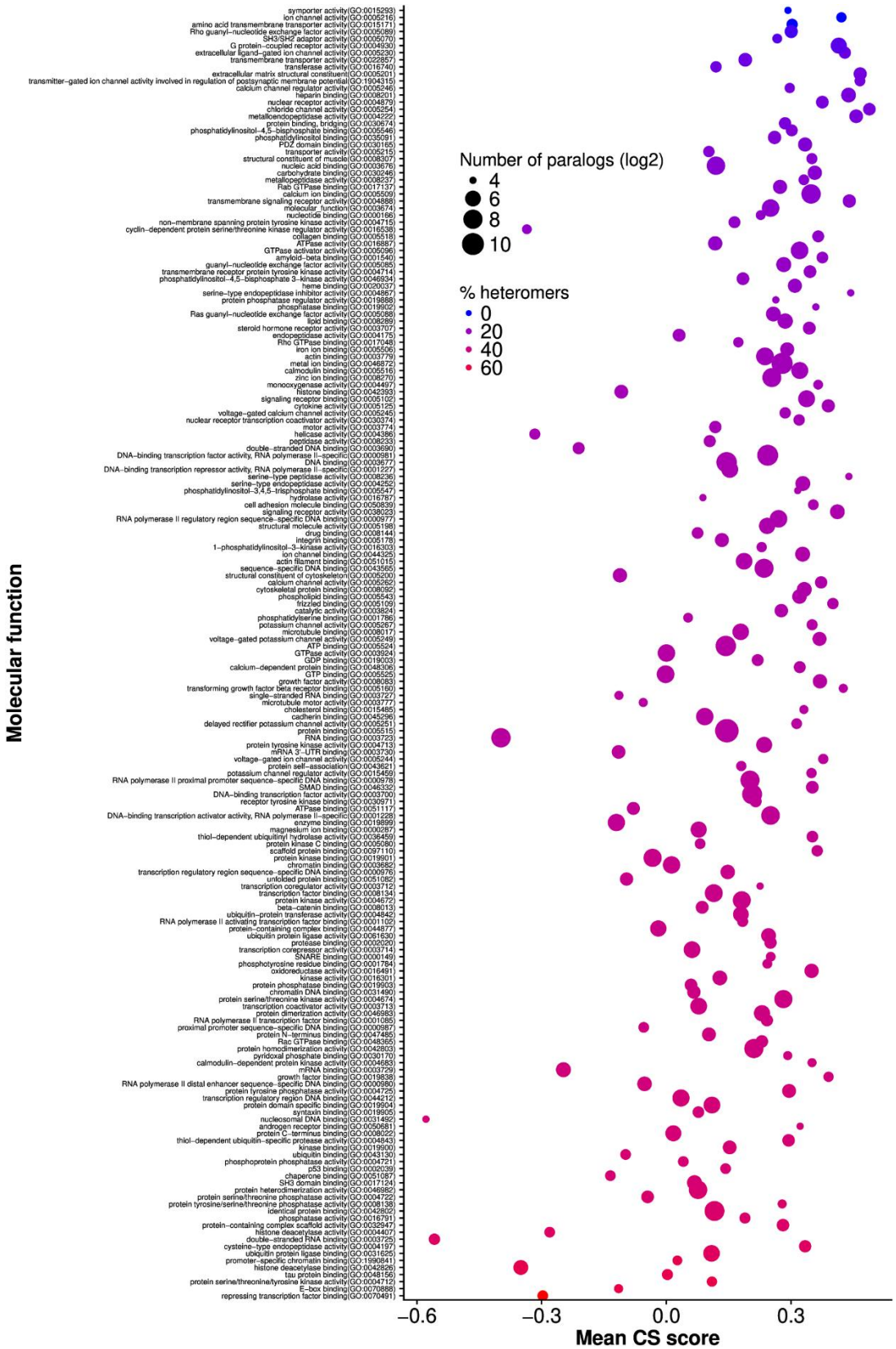


Figure S8. GO analysis with CS2

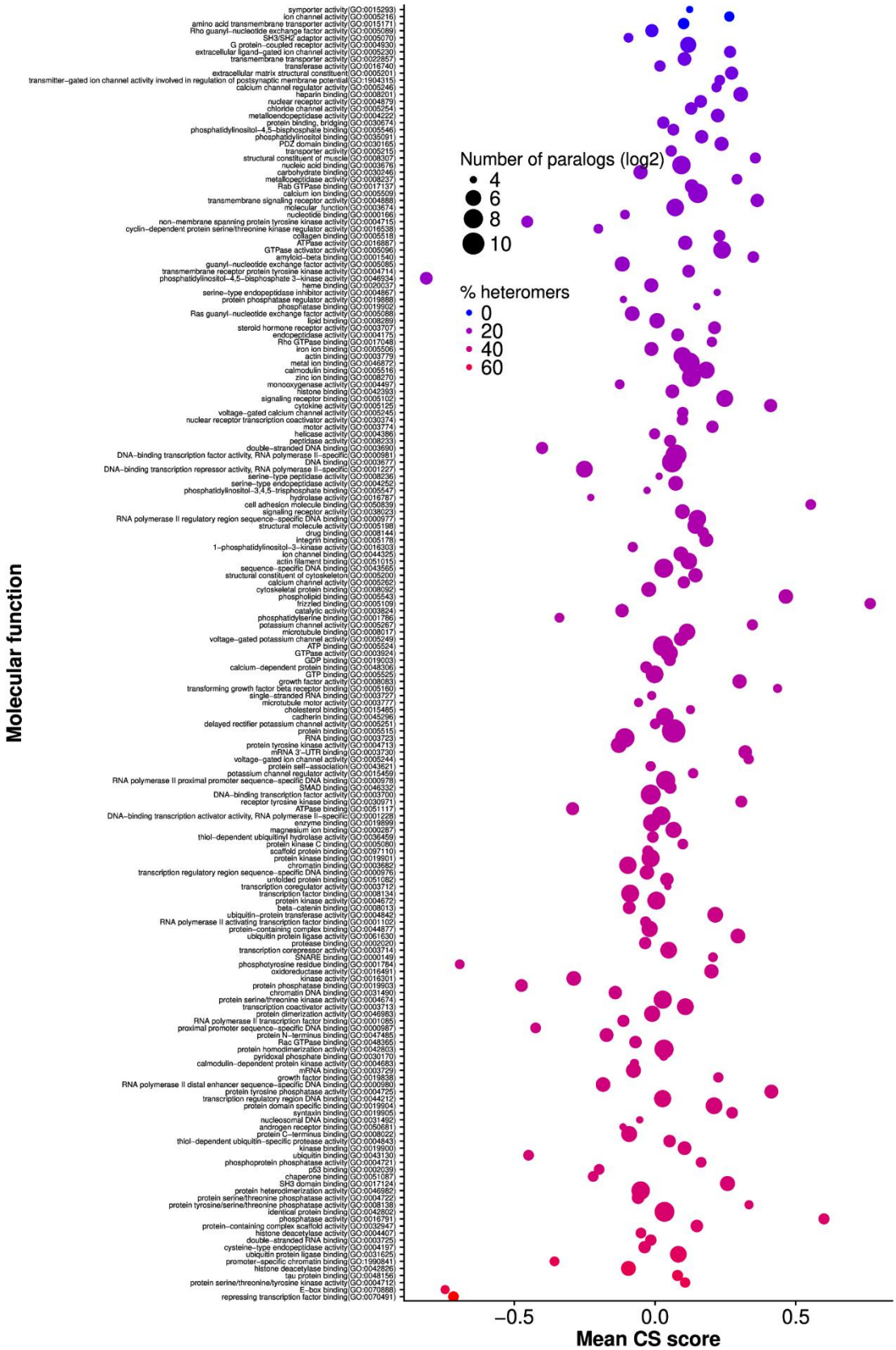


Figure S9. GO analysis with CS3

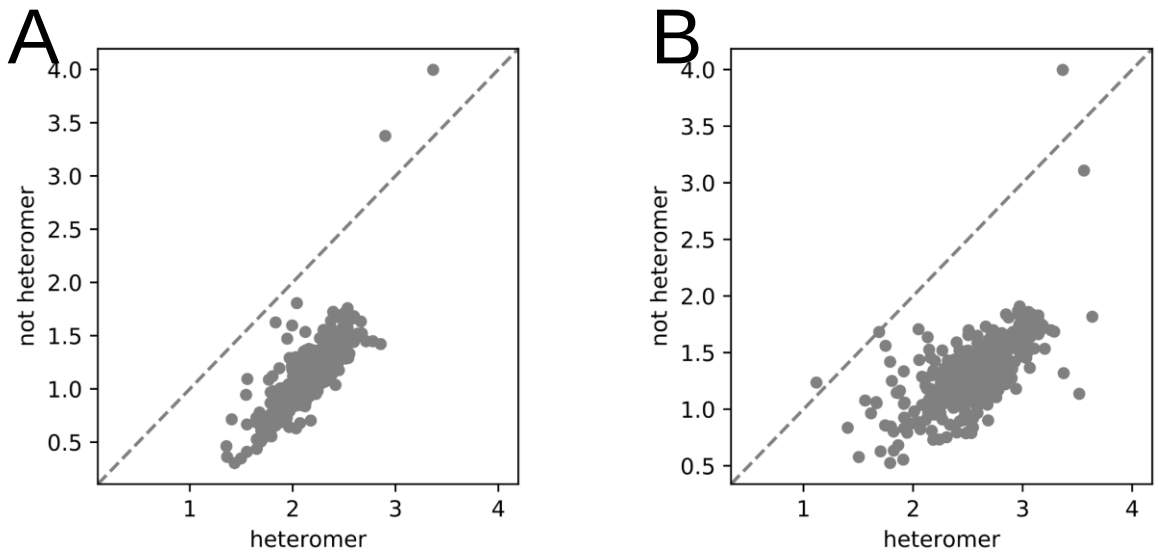


Figure S10. Cell-line wise comparison of RNA expression between paralogous heteromers and paralogous non-heteromers. Each point represents the mean FPKM (log2 scale) score for a class (singleton or paralog, heteromer or not heteromer) in an individual cell line (n=374). Data for heteromers of paralogs detected from all physical interactions is shown in **A**, while those detected from only direct physical interactions is shown in **B**. In each case the ~99% of the points are separated by the diagonal (dashed gray line) indicating cell-line independent systematic effects.

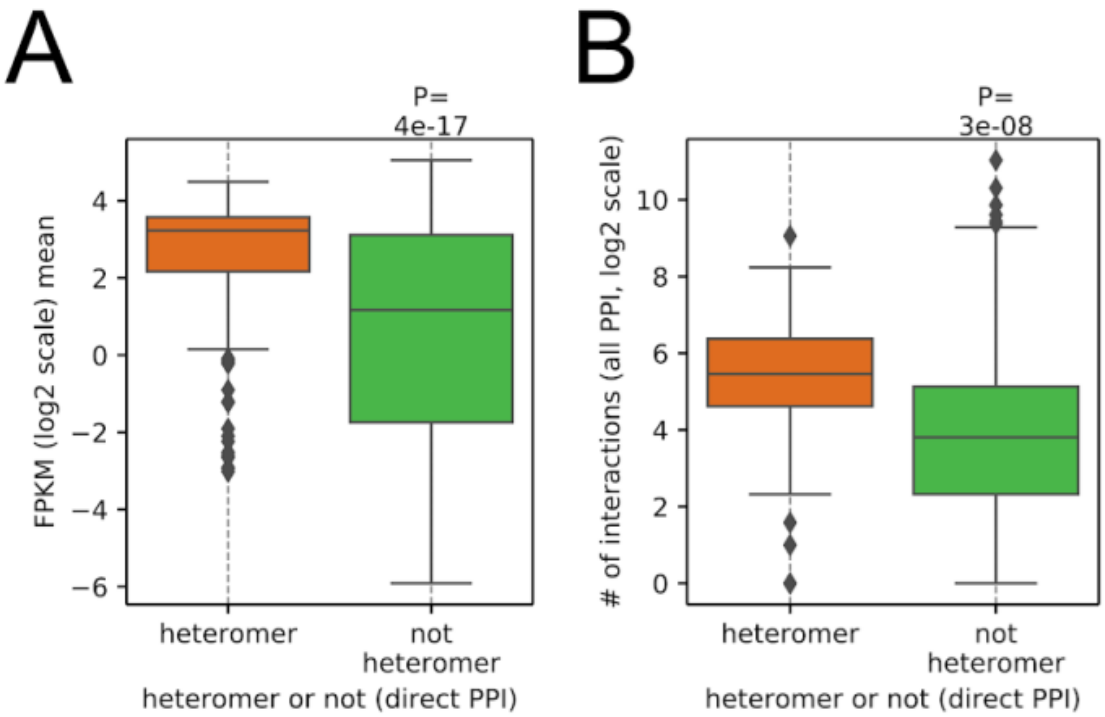


Figure S11. Number of protein-protein interaction partners and mRNA abundance of heteromeric paralogs. (A) Shows the number of physical interaction partners for paralogs that form heteromers and other paralogs for interactions detected using direct physical interaction methods. (B) Shows the mRNA abundance for paralogs that form heteromers and other paralogs for interactions detected using direct physical interaction methods. P-values are from two-sided Mann-Whitney U test.

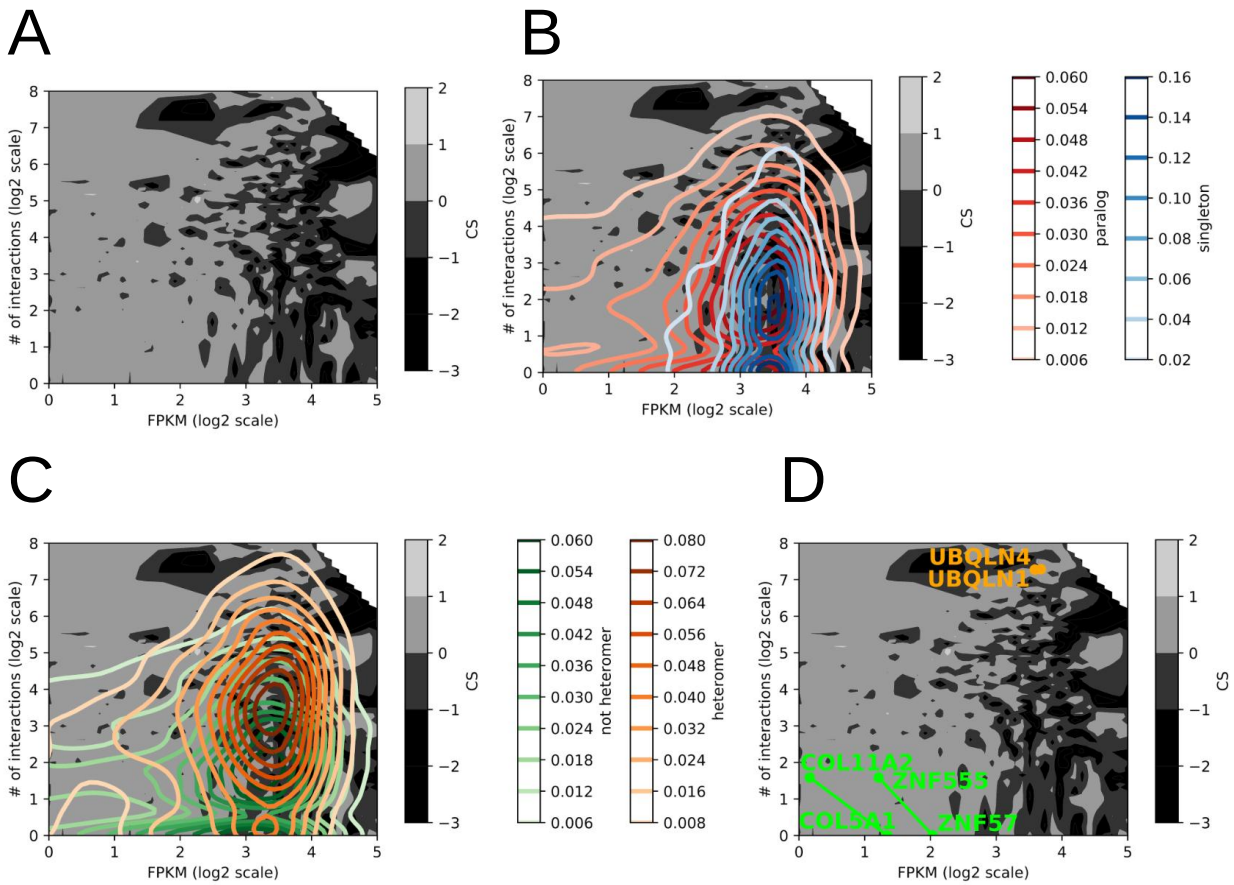


Fig. S12 Landscape of the robustness of human cell lines to paralog deletion in lines utilizing direct physical interactions only. RNA expression level (log_2 scale FPKM scores) and number of direct protein-protein interaction partners (log_2 scale) are strong determinant of the deleteriousness of gene loss-of-function.

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

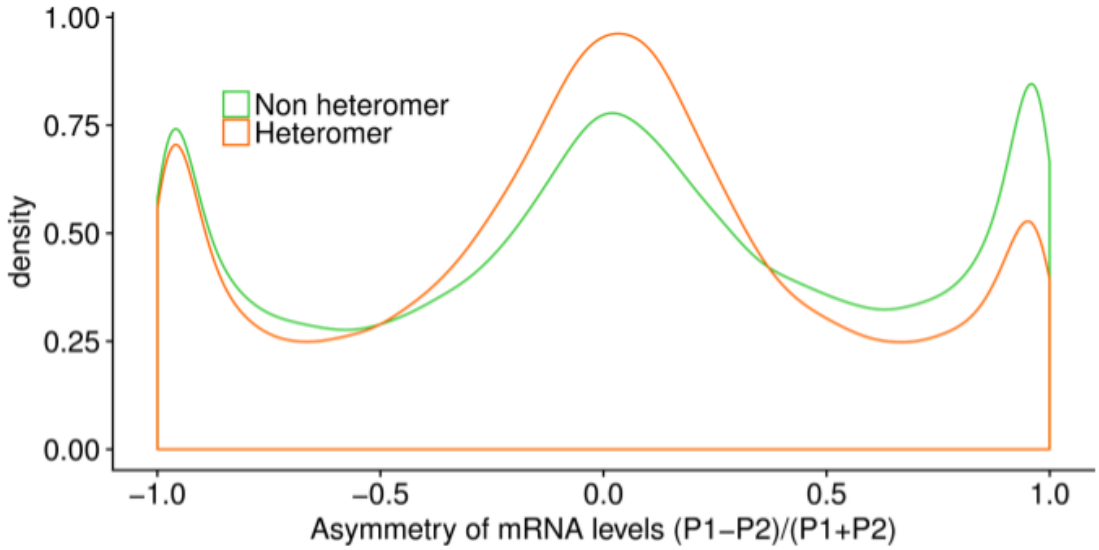


Fig. S14. Asymmetric mRNA abundance of the paralogs across cell lines. For each pair of paralogs, the extent of asymmetry across all cell lines was measured as the difference of expression (FPKM) over the sum of expression, setting arbitrarily a paralog of a pair as P1 and the other as P2. Accordingly, asymmetry follows a scale from 0 to 1. Values near -1 are cases in which the mRNA abundance for a pair of paralog is dominated by P2 and the opposite for values near 1. Pairs of paralogs that form heteromers tend to have more symmetrical expression levels (values around 0).

Supplementary tables

Table S4: Number of homomeric (P1P1 or P2P2) and heteromeric (P1P2) paralogs.

PPI type	gene1 homomer	gene2 homomer	heteromer	# of paralogs
all PPI	-	-	-	2569
all PPI	-	-	P1P2	368
all PPI	-	P2P2	-	210
all PPI	-	P2P2	P1P2	120
all PPI	P1P1	-	-	229
all PPI	P1P1	-	P1P2	115
all PPI	P1P1	P2P2	-	75
all PPI	P1P1	P2P2	P1P2	156
direct PPI	-	-	-	3336
direct PPI	-	-	P1P2	23
direct PPI	-	P2P2	-	180
direct PPI	-	P2P2	P1P2	18
direct PPI	P1P1	-	-	164
direct PPI	P1P1	-	P1P2	22
direct PPI	P1P1	P2P2	-	52
direct PPI	P1P1	P2P2	P1P2	47