SI Appendix 2

Calculation and characterization of the association between the connectivity of duplicated genes to their tendency to have redundant partners. Figure 2 in the main article shows that at every degree of connectivity, *k*, the singletons are more essential than the duplicates and that the difference in the essentiality of the two gene sets increases with *k*. In this section we analyze different regimes. Here, we provide a simple analytic treatment of the data to show that the difference in essentiality between the duplicates and the singletons results from a tendency of redundancy to be preferentially associated with the highly connected proteins.

Table of Variables

Variable	Description
k	The number of partners the genes' protein product has in the protein interaction network.
$P_{iv}^{S}(k)$	The (conditional) probability for an ***inviable*** deletion phenotype (*iv*) given that the gene is a singleton (S) with degree *k*. Could also be written $P(iv\|Singleton,k)$.
$P_{iv}^{D}(k)$	The (conditional) probability for an ***inviable*** deletion phenotype (*iv*) given that the gene is a duplicate (D) with degree *k*. Could also be written $P(iv\|Singleton,k)$.
$P_{ef}^{S}(k)$	The (conditional) probability that the gene performs an '*essential function*' (*ef*) given that the gene is a singtон with degree k. Could also be written, $P(ef\|k,S)$
$P_{ef}^{D}(k)$	The (conditional) probability that the gene performs an '*essential function*' (*ef*) given that the gene is a duplicate with degree k. Could also be written, $P(ef\|k,D)$
$P_{BU}(k)$	For duplicated genes, the probability that the duplicate of a gene with degree k can function as its backup, i.e. compensate against its loss (BU).

$P_{\overline{BU}}(k)$	For duplicated genes, the probability that the duplicate of a gene with degree k **<u>can not</u>** function as its backup. Obeys the relationship: $P_{BU}(k) = 1 - P_{\overline{BU}}(k)$
ϕ	The proportion of duplicated genes that **<u>do not</u>** have a redundant partner.
θ_t	The proportion of duplicated genes that **<u>do</u>** have a redundant partner. Obeys the relationship, $1 - \theta_t = \phi$
α	The slope of the centrality lethality in singletons
β	The intersect of the centrality lethality in singletons

Section I: Relating the essentiality of duplicate genes to that of singtons

Our reasoning equates the proportion of inviable singletons with the proportion of *essential-functions* in singletons. In other words, for singletons the phenotype directly reflects the contribution of the genes' function to the fitness. (Note that this is not the case for duplicates that have redundant partners.). To formally state this we write that the proportion of genes with an inviable deletion phenotype, $P_{iv}^{S}(k)$, equals the proportion of genes that perform essential functions, $P_{ef}^{S}(k)$.

Eq. **1**: $P_{iv}^{S}(k) = P_{ef}^{S}(k)$

We assume that the association stated above between the *function-essentiality* of genes to their degree, *k*, is not exclusive to singleton genes but exists in both singltons and duplicates. We thus state,

Eq. **2**: $P_{iv}^{S}(k) = P_{ef}^{S}(k) = P_{ef}^{D}(k)$

For a validation of the above relationship see figure 6 in the main article. For duplicate genes, the proportion of essential genes does not directly reflect the proportion of genes associated with *essential functions* as the phenotypes of some functionally essential duplicates are buffered by redundant partners. Thus, the probability that a duplicate is associated with an inviable phenotype is proportional to the probability that it performs an essential function and to the probability that it is *not* backed-up by a redundant partner.

Eq. **3**: $P_{iv}^{D}(k) = P_{ef}^{D}(k)\Big(1 - P_{BU}(k)\Big)$

$= P_{ef}^{D}(k) P_{\overline{BU}}(k)$

Now from Eq. 2 we have $P_{iv}^{S}(k) = P_{ef}^{D}(k)$. (Note that the left hand of the eq. is a measurable quantity that we have obtained.) So we have

Eq. **4**: $P_{iv}^{D}(k) = P_{iv}^{S}(k) P_{\overline{BU}}(k)$

Rearranging we have $P_{\overline{BU}}(k) = \dfrac{P_{iv}^{D}(k)}{P_{iv}^{S}(k)}$ and

Eq. **5**: $P_{BU}(k) = 1 - \dfrac{P_{iv}^{D}(k)}{P_{iv}^{S}(k)}$

Now from observation we know that $P_{iv}^{S}(k)$ is linear with *k*.

Observation A: $P_{iv}^{S}(k) = \alpha k + \beta$

Note that the linear relationship stated in the above *Observation A* reflects an approximated fit and not the true functional form of $P_{iv}^{S}(k)$ as evident from the fact that $P_{iv}^{S}(k)$ is bound by {0,1}. Nevertheless, the data shows that this approximation is valid for a certain and significant interval of *k.* Combining *Observation A* with Eq. 4 obtain that:

Eq. 6: $P_{iv}^{D}(k) = (\alpha k + \beta)\, P_{\overline{BU}}(k)$

And

$P_{v}^{D}(k) = 1 - (\alpha k + \beta)\, P_{\overline{BU}}(k)$

Section II: The null model – assuming no preferential association of redundancy with protein connectivity

Our null model hypothesis is that there is no association between the number of physical interactions partners of a given protein to the probability, $P_{\overline{BU}}$, that it has a redundant

duplicate partner. In other words, $P_{\overline{BU}}$ is not a function of *k*. For simplicity we can write $P_{\overline{BU}} = \phi$ and then Eq. 6 becomes:

Null Hypothesis: $P_{iv}^{D}(k) = \phi\alpha k + \phi\beta$

This provides us with a strong tool to detect whether the difference in the slopes of duplicate essentiality and singleton essentiality is sufficiently large to suggest a preferential association of redundancy with *k*. Specifically, we calculate from the data the slopes and intercepts of both $P_{iv}^{D}(k)$ vs. *k* and for $P_{iv}^{S}(k)$ vs. *k*. By comparing the 'Null Hypothesis' to 'Observation A' we see that the ratio between the two intercepts should equal ϕ.

$$\frac{intercept \quad duplicates}{intercept \quad singletons} = \phi$$

And therefore the slope for $P_{iv}^{D}(k)$ vs. *k* should equal

$$\begin{matrix} predicted \quad slope \\ of \quad duplicates \end{matrix} = \left(\frac{intercept \quad duplicates}{intercept \quad singletons} \right) \times \left(slope \quad singletons \right)$$

A slope of $P_{iv}^{D}(k)$ vs. *k* that is smaller than the predicted slope would indicate preferential association of redundancy with degree of connectivity.

	Slope under null hypothesis	Measured slope
All duplicates	8.81×10^{-3}	5.7×10^{-3}
Duplicates with mean expression similarity >0.3	1.42×10^{-2}	2.4×10^{-3}
Duplicates excluding genes that are involved in protein complexes.	1.0×10^{-2}	3.0×10^{-3}

Table S1: We relied on the relationship above to calculate the slope for $P_{iv}^{D}(k)$ under the null assumption that redundancy is uniformly distributed among duplicates. These

predicted slopes are compared with the real measured slopes for i) all dupicates, ii) only duplicates that are dissimilarly expressed and iii) duplicates excluding duplicated genes that are involved in protein complexes.

Section III: Calculating the association between redundancy and protein connectivity

An association between redundancy and protein connectivity is captured by the function $P_{BU}(k)$, which is the probability that a duplicated gene is backed-up by a redundant partner as a function of its degree (*k*). From Eq. 5 we see that this function could directly be calculated from the dependencies of the proportions of essential duplicates and singletons on *k* (see Figure S1 below).

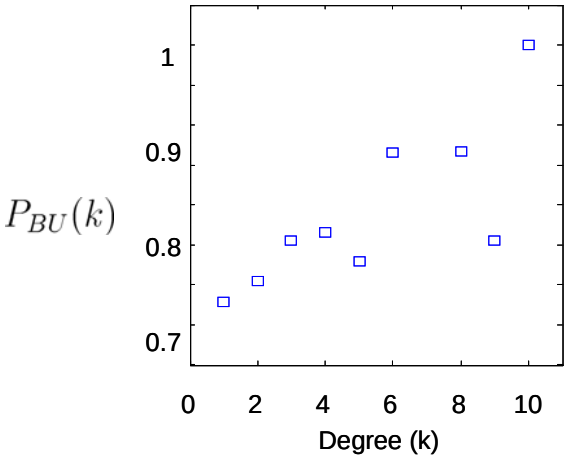

Figure S1: Probability of a gene to be backed up by its' duplicate as a function of its degree calculated from Eq. 5. We calculated $P_{BU}(k)$ only for duplicates with mean expression similarity below 0.3 as these were shown to contain the largest proportions of redundant pairs.

We can further ask more generally what difference between the slopes $P_{iv}^{D}(k)$ vs. *k* and $P_{iv}^{S}(k)$ vs. *k* would significantly suggest a preferential backup of connected proteins. The answer to this depends largely on the total proportion of duplicates that have redundant partners. To work this relationship out quantitatively we make a simplifying assumption that $P_{BU}(k)$ is linear with *k* and write,

Eq. **7**: $P_{BU}(k) = ak + b$

To keep $P_{BU}(k)$ within the bounds {0,1} we specify:

$$P_{BU}(k) = \begin{cases} ak + b & \text{for } 0 < (ak + b) < 1 \\ 1 & \text{for } (ak + b) > 1 \\ 0 & \text{for } (ak + b) < 0 \end{cases}$$

Relying on Eq. 7, Eq. 4 and the data for $P_{iv}^{S}(k)$ we can now calculate predicted slopes for $P_{iv}^{D}(k)$ vs *k* by assuming different values for the overall proportion of redundant duplicates, θ_t, and the extent to which duplicate redundancy is biased towards connected proteins (see figure S2). To do that we must first express $P_{BU}(k)$ as a function of *a* and θ_t. We thus write the constrain,

$$\sum_k (ak + b)\, P(k) = \theta_t$$

Opening the sum we get,

$$\sum_k (ak + b)\, P(k) = \sum_k akP(k) + \sum_k bP(k)$$

$$= a \sum_k kP(k) + b \sum_k P(k)$$

$$= a\langle k \rangle + b = \theta_t$$

Thus,

$$b = \theta_t - a\langle k \rangle$$

Substituting this into Eq. 7 we get:

$$P_{BU}(k) = ak + \theta_t - a\langle k \rangle$$

And rearranging,

Eq. **8**: $P_{BU}(k) = a\,(k - \langle k \rangle) + \theta_t$

And

Eq. **9**: $P_{\overline{BU}}(k) = 1 - (a\,(k - \langle k \rangle) + \theta_t)$

Combining with Eq. **6**

Eq. **10A**: $P_{iv}^{D}(k) = \Big(\alpha k + \beta\Big)\Big(1 - a\,(k - \langle k \rangle) - \theta_t\Big)$

And so

Eq. **10B**: $P_{v}^{D}(k) = 1 - \Big(\alpha k + \beta\Big)\Big(\phi_t - a\,(k - \langle k \rangle)\Big)$

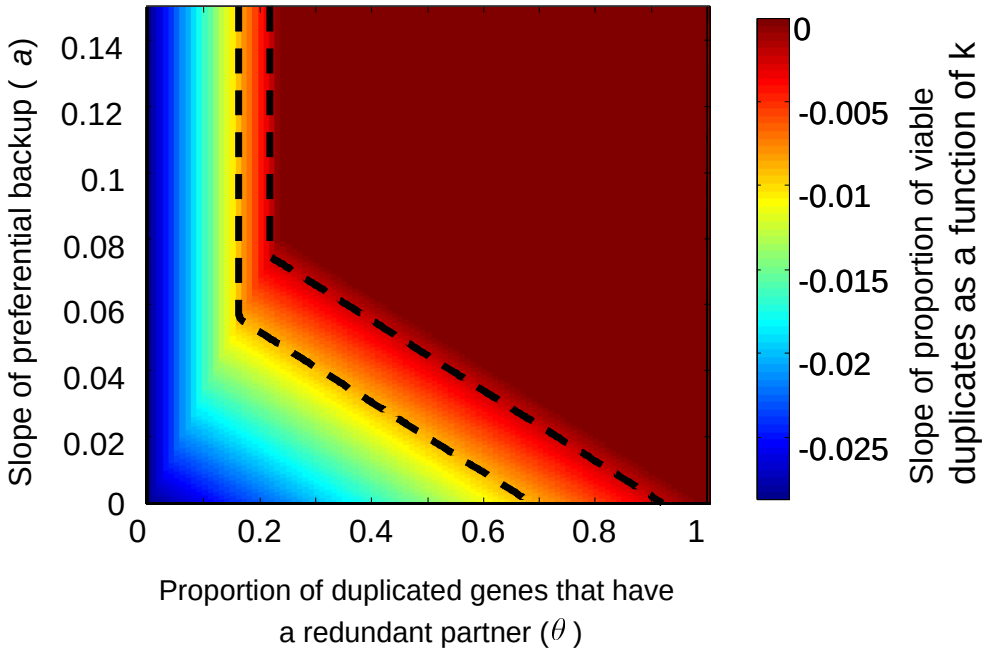

Figure S2: Slopes of the proportion of viable duplicates as a function of the overall proportion of redundant duplicates, θ_t, and the tendency of redundancy to be associated with connected proteins (quantified by the slope a, defined in eq. 7). Values of $P_{v}^{D}(k)$ were calculated from $P_{iv}^{D}(k) = P_{iv}^{S}(k)P_{\overline{BU}}(k)$ by assuming a linear form of $P_{\overline{BU}}(k)$ (Eq. 7). The explicit formula for this is given in Eq. 10B. Slopes were calculated by performing linear fits to the calculated $P_{iv}^{D}(k)$ data. The dotted black line contains the region in which with slope values corresponding to those that we have observed for duplicates with appropriate error interval. The plot, thus, shows that or results could have been obtained without preferential backup if over 90% of all duplicates were redundant. For reference, the corresponding slope of singletons, $P_{iv}^{S}(k)$, is -0.028.

Conclusion

Thus using a simple formal analysis we were able to show that the difference between the proportion of non-viable singletons to the proportion of non-viable duplicates can only be explained by a preferential association of redundancy with the highly connected proteins. Thus, not only are the duplicates less essential than singletons at any value of *k* but they are also more dispensable than would have been expected if redundancy was uniformly distributed among the paralogous pairs irrespective of their connectivity, *k*.

Duplicate gene dataset and protein-protein physical interaction data

A total of 2,216 duplicate genes were collected based on PBLAST as previously described (24). The list of paralog pairs used in this study, along with the paralogs' corresponding values of mean expression similarity and degree connectivity, are provided in SI Table 2. The degree of connectivity of each of the genes in the protein interaction network was retrieved from the GRID database (40) (http://biodata.mshri.on.ca/yeast_grid/servlet/SearchPage), which combines literature-derived and high-throughput physical protein-protein interactions. (See further details in *SI Appendix 4*). Analyses were initially performed using the June 2005 version of the GRID database. Based on this version, hubs were defined as proteins with >10 interactions, and were subsequently chosen for experimental analysis. We then repeated all computational analyses with the September 2006 version of GRID. Thus, all computational analyses described herein are based on this latter version; yet qualitatively, they are indistinguishable from our analyses of the June 2005 version (data not shown). Because GRID contains multiple sources for protein-protein interactions, we considered an interaction valid only if it was derived from at least one of the following methodologies (classification defined in SGD, http://www.yeastgenome.org/help/glossary.html): (*i*) Affinity Capture-MS; (*ii*) Affinity Capture-Western; (*iii*) Cocrystal structure; (*iv*) Cofractionation; (*v*) Copurification; (*vi*) Far Western; (*vii*) Reconstituted Complex; or (*viii*) Two-hybrid.

---

[1] ?/Au: Please confirm the SI callouts in this document.

We also performed all analyses on the "High Confidence" (HC) dataset (41), which is based on literature-derived, multi-validated dataset of protein interactions. The trends reported here based on the GRID interactions are qualitatively similar to those obtained using the HC dataset (see SI File 5).

Synthetic sick and synthetic lethal experiments: strains, media, growth conditions and tetrad analysis

The following criteria were used when choosing genes for the double knockout experiments: For highly connected proteins, we examined all non-essential dispensable hubs (with >10 physically interacting partners) that had a non-similarly expressed paralog (0 < mean expression similarity < 0.3). Based on the June 2005 version of the GRID database. For sparsely connected proteins, we examined all dispensable non-hubs (0-1 physically interacting partners for both paralogs) that had only one duplicate. Based on the June 2005 version of the GRID database.
All *S.cerevisiae* disruption strains used in the present work are based on the following genetic backgrounds:

BY4741: *MATa*; *his3Δ1*; *leu2 0*; *met15Δ0*; *ura3Δ0*
BY4742: *MATα* ; *his3Δ1*; *leu2Δ0*; *lys2Δ0*; *ura3Δ0*
All disruptions were marked by *kanMX4* (43)

Yeast cells were grown in YEPD (1% yeast extract, 2% Bacto peptone, 2% dextrose). Sporulation was carried out in SPO medium (1% potassium acetate, 0.1% yeast extract, 0.05% dextrose) by incubating cells for 72h at 25^0C.

Diploid selection and tetrad analysis were carried out using the Singer MSM Manual Micromanipulator, according to the manufacturer's instructions. Genetic interactions were scored by conventional tetrad analysis. Briefly, since both deletions are marked by the same marker (*kanMX4*) we use both spores viability and their ability to grow on G418 containing media to deduce synthetic lethality. The following growth

pattern are expected in case of synthetic lethality: in tetra type tetrads (TT) – three viable spores from which two are G418 resistant (compared with four viable spores of which three are G418 resistant in case of no genetic interaction); in nonparental ditype tetrads (NPD)– two viable spores both of which are sensitive to G418 (compared with four viable spores, of which two are G418 resistant in case of no genetic interaction); in parental ditype tetrads (PD) - four viable spores, all G418 resistant, this pattern is similar in the existence and lack of genetic interaction. For each cross we analysed more than 40 tetrads and use the expected growth patterns describe above to deduce genetic interactions. Spores were plated on YEPD plates, incubated for 72h at 25^0C, following by replica plating onto YEPD and selective plates. The following selective plates were used: SD (0.67% yeast nitrogen base, 2% dextrose, 1.8% agar and the appropriate nutrients) lacking methionine, SD lacking lysine, SD lacking serine (in crosses involving disruption of the gene *YER081W*); YEPD containing G418 (200 mg/liter), and YEPGal plates (1% yeast extract, 2% Bacto peptone, 2% galactose-in crosses involving disruption in *YMR105C*).