Description of Supplemental Spreadsheet

Contained in the supplemental spreadsheet are 9 worksheets with results regarding WGD paralogs. Specifically those worksheets are:

Kellis list – The original list of WGD paralogs published by Kellis et al (Kellis et al. 2004)

KAN and NAT – Names of deletion strains used for WGD and RSA experiments, as well as the 100 double-mutant controls

Paralog + ID – All paralogs with results of amino-acid sequence alignment using both Smith-Waterman and Needleman-Wunsch algorithms

Paralog + RSA – Results of RSA screening as assessed by the two independent scorers.

Paralog + GCA – Results of GCA indicating the potential interactors and all data used in multiplicative model.

Paralog + SS – Hybridization data for the 61 pairs determined to have relative expression below 75% (following normalization, see **Methods**) in a given condition.

Paralog + PPI – Protein interaction data from Krogan et al (Krogan et al. 2006), Gavin et al (Gavin et al. 2006), and bioGRID (Reguly et al. 2006) datasets for all paralog pairs.

Paralog + type – Paralog pairs with classification as either metabolic enzyme or ribosomal protein.

Paralog + GO – Results of BinGO (Maere et al. 2005) analysis for buffering, non-buffering and suspected SE paralogs.

Supplemental Methods

Assessment of synthetic lethality for RSA and GCA

RSA

Cells were grown for 2-3 days at 28°C, photographed, and classified by two independent observers as being either synthetic lethal, synthetic sick, having no interaction, being unclear, or being abnormal (abnormal indicates that a single deletion or control strain was inviable). Any strain determined by both observers to be either synthetic lethal or synthetic sick was classified as such. Strains classified by one observer as synthetic lethal or synthetic sick but by the other as unclear were further investigated by tetrad analysis (see Fig. S1). Lastly, all abnormal crosses (either one deletion strain or control strain inviable) were re-analyzed.

GCA

The resulting growth-curves deemed to either: have a lower point of saturation, have an obvious growth lag, or have a decreased slope in exponential growth phase when compared to plate-specific controls were labeled potential interactors. Next, growth rates of the constitutive single-mutant deletion strains of potential interactors were assayed in duplicate, and area under growth curve calculated after 20 hours growth. Genetic interaction between gene pairs was assessed as that beyond the predictions of the multiplicative model (analogous to (St Onge et al. 2007)):

Let W_x equal the fitness of mutant strain x (as compared to plate-specific control), W_y the fitness of the strain carrying the deletion for the corresponding sister, and

W_{xy} the fitness of the dual-deletion strain. A genetic interaction then is characterized as:

$$W_{xy} < (W_x * W_y) - (\sigma_x + \sigma_y)$$

Where σ_k represents the standard deviation of deletion strain k measured over replicates.

Determination of condition-specific synthetic lethality

For each of the five media conditions tested, intensity values resulting from hybridization (described above) were Lowess normalized using ~1000 barcoded strains which had been independently grown in YPD and additionally hybridized to each chip used (done to remove potential spatial bias). Expression data for WGD paralogs were then normalized both by row and column. Row values (dual-deletion strain expression) for the five experimental conditions were normalized using each strain's hybridization in control YPD. Columns (conditions) were normalized using the average expression value of hybridizing non-WGD double-mutant deletion strains grown in each condition. Normalization was performed independently for both experimental runs and for UPTAG and DNTAG expression. Data were then combined, and those strains with below 75% normalized expression in a given condition were treated as potential interactors.

A subset of the most consistently affected potential interactor strains was then selected to be further analyzed through GCA. To determine this subset, all strains in the initial pool of 499 (399 paralog double-deletion strains + 100 internal controls) with expression two-fold beyond background in control YPD (background determined as the

average expression value of non-existing strains) were given an incremental rank in each condition according to their expression magnitude. For each strain, the difference between rank within the control state and in any given media state was calculated (changes in rank for entire cell population were found to be roughly normally distributed around 0 for each condition). Those strains indicated as above to be potential interactors and with changing rank beyond one standard deviation in both experimental runs (in both UPTAG and DNTAG expression) were selected. Corresponding single-mutant and double-mutant strains were then grown and analyzed similar to described above for GCA, however in this instance, GCA was performed entirely in the given media condition. Those paralog pairs passing the multiplicative model (see above) were confirmed as being sensitive to the given condition.

Supplemental Results

Presence of additional duplicates does not affect epistasis

To analyze the affect of multiple paralogy on epistasis we sub-divided WGD paralogs based on the existence of additional duplicates and compared frequencies of epistasis. Additional (i.e. non-WGD-resultant) cases of paralogy were determined for 449 of the initial set of 457 WGD paralog pairs (the 7 pairs initially described (Kellis et al. 2004) as being split into multiple open reading frames as well as one pair containing a categorized pseudogene were excluded), similar to the method previously described (Gu et al. 2002). Briefly, protein sequences corresponding to every known cDNA sequence in *S.cerevisiae* (excluding hypothetical or dubious open reading frames, 5880 total) were

downloaded from the SGD database (www.yeastgenome.org) and BLAST analysis was performed aligning each of the 898 individual WGD paralog genes against all *S.cerevisiae* protein sequences with an expectation (e-value) cutoff of 0.1. From there, resulting alignments in which the aligned region covered at least 50% of the sequence of the larger protein were retained (50% used as the cutoff for the aligned region as opposed to the more common value of 80% in order to identify a greater number of paralogs (as previously described (Gu et al. 2003)). As a final criterion, alignments were required to meet a threshold of percent sequence identity in order to be retained (Gu et al. 2003). For alignments where the aligned region was greater than 150aa in length, a minimum sequence identity of 30% was required. Based on previous empirical evidence indicating that a more stringent sequence identity cutoff is needed for smaller proteins (Rost 1999), for all alignments shorter than 150aa, sequence identity was required to surpass the value determined by a pre-determined formula (Rost 1999):

$$n + 480 * L^{-0.32 \times (1 + e-L/1000)}$$

Where n = 6 (as previously established (Gu et al. 2003)) and L represents the length of the aligned region. Based on these identifications, we sub-divided WGD paralog pairs into three groups: those pairs where both members had additional paralogs, those pairs where only one member had an additional paralog, and those pars where neither had additional paralogs (21%, 4% and 75% of all WGD paralog pairs respectively). Upon comparison, epistatic and non-epistatic paralogs showed no appreciable differences in the representation of the three groups of WGD duplicates (see Fig. S2). Therefore we

conclude that presence of an additional duplicate does not influence the propensity of a WGD paralog pair to be epistatic.

Expression properties correlate with sequence similarity for all paralogs

As mentioned in the manuscript, buffering paralogs (IG and UG) exhibited increased expression magnitude (protein & mRNA) and co-ordination when contrasted against respective non-buffering paralogs. However, correlations between sequence identity and gene expression (r=0.486), protein expression (r=0.529), expression correlation (r=0.336), and CAI (r=0.748) were all highly significant ($p < 1x10^{-10}$ for each). As buffering paralogs were more highly conserved than non-buffering, these correlations suggested that it would be impossible to determine whether increased sequence similarity caused increased expression, or vice versa.

Supplemental Figures & Tables

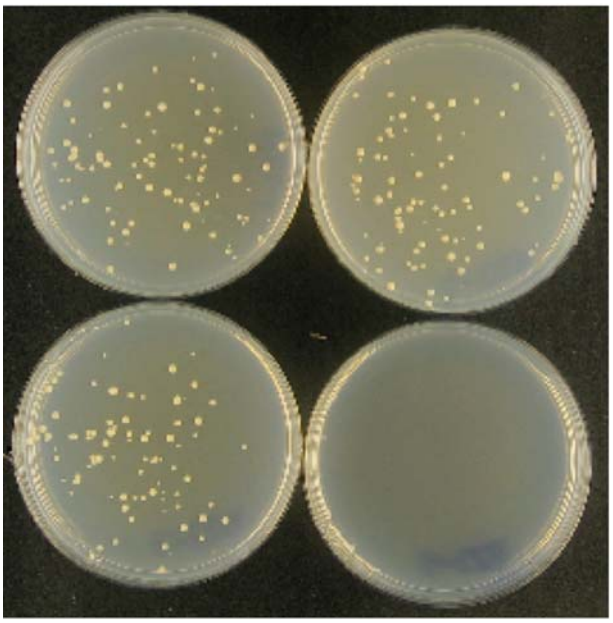
		Scorer 1				
		SyS	SL	Unclear	No Interaction	Dead Cells
Scorer 2	SyS	37	5	10	0	0
	SL	1	8	8	0	0
	Unclear	0	1	12	10	0
	No Interaction	0	0	1	278	0
	Dead Cells	0	0	3	13	12

Table S1 – Results of Random Spore Analysis (RSA). Resulting colonies from all assayed WGD paralog pairs were photographed and assessed by 2 independent researchers (indicated as Scorer #1 and Scorer #2). Pairs evaluated by both scorers as being either synthetic sick (SyS) or synthetic lethal (SL) were treated as genetic interactors (indicated in red). Any pairs evaluated by either researcher as being 'Unclear' (indicated in blue) were analyzed through tetrad dissection.

ORF 1	Gene 1	ORF 2	Gene 2	RSA Result
YBL087C	RPL23A	YER117W	RPL23B	SL
YBR010W	HHT1	YNL031C	HHT2	SL
YBR118W	TEF2	YPR080W	TEF1	SL
YDL138W	RGT2	YDL194W	SNF3	SL
YDR098C	GRX3	YER174C	GRX4	SL
YER081W	SER3	YIL074C	SER33	SL
YGL076C	RPL7A	YPL198W	RPL7B	SL
YGR124W	ASN2	YPR145W	ASN1	SL
YGR254W	ENO1	YHR174W	ENO2	SL
YIL105C	SLM1	YNL047C	SLM2	SL
YKL129C	MYO3	YMR109W	MYO5	SL
YMR186W	HSC82	YPL240C	HSP82	SL
YOR226C	ISU2	YPL135W	ISU1	SL
YBR210W	ERV15	YGL054C	ERV14	SyS
YDL022W	GPD1	YOL059W	GPD2	SyS
YDR436W	PPZ2	YML016C	PPZ1	SyS
YGR192C	TDH3	YJR009C	TDH2	SyS
YJL098W	SAP185	YKR028W	SAP190	SyS
YJL133W	MRS3	YKR052C	MRS4	SyS
YKL032C	IXR1	YMR072W	ABF2	SyS
YKL043W	PHD1	YMR016C	SOK2	SyS

Table S2 – Strains determined to be inviable using RSA but not GCA. While most genetic interactions detected through RSA were also found through GCA, 21 RSA interactors had normal growth curves. As indicated above, 8 of the 21 were synthetic sick (small colony sizes) and thus may still have had normal rates of colony growth. The remaining 9 likely had a buffering relationship only presenting in the solid minimal media of the RSA experiments.

S1-a



S1-b

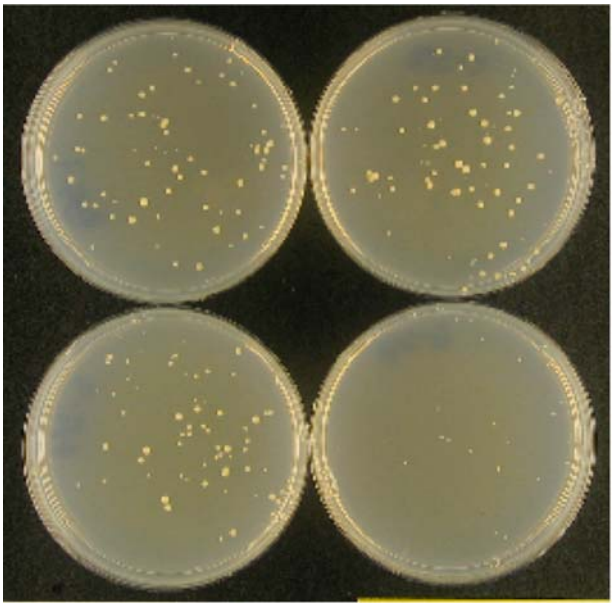
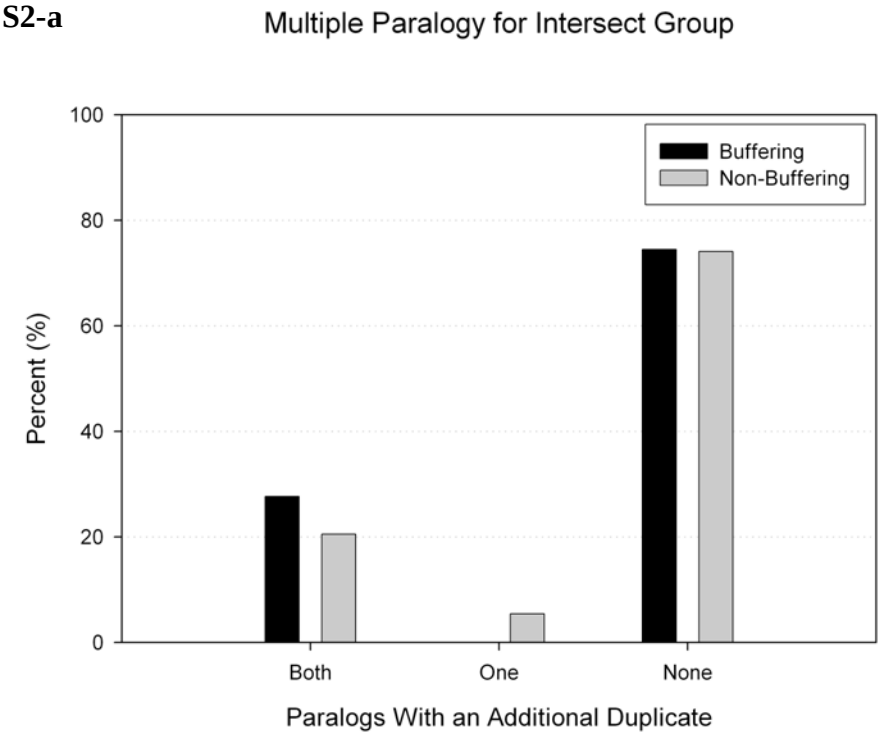


Figure S1 – Synthetic sickness and synthetic lethality as assessed through RSA. Double mutant colonies with obvious lethal (a) or sick (b) phenotypes scored as genetic interactions. In both (a) and (b) top left colony represents cells with no mutation, top right represents mutation of one paralog, bottom left deletion of the corresponding sister paralog, and bottom right the strain carrying the dual-deletion. All pictures are available upon request.

S2-a



S2-b

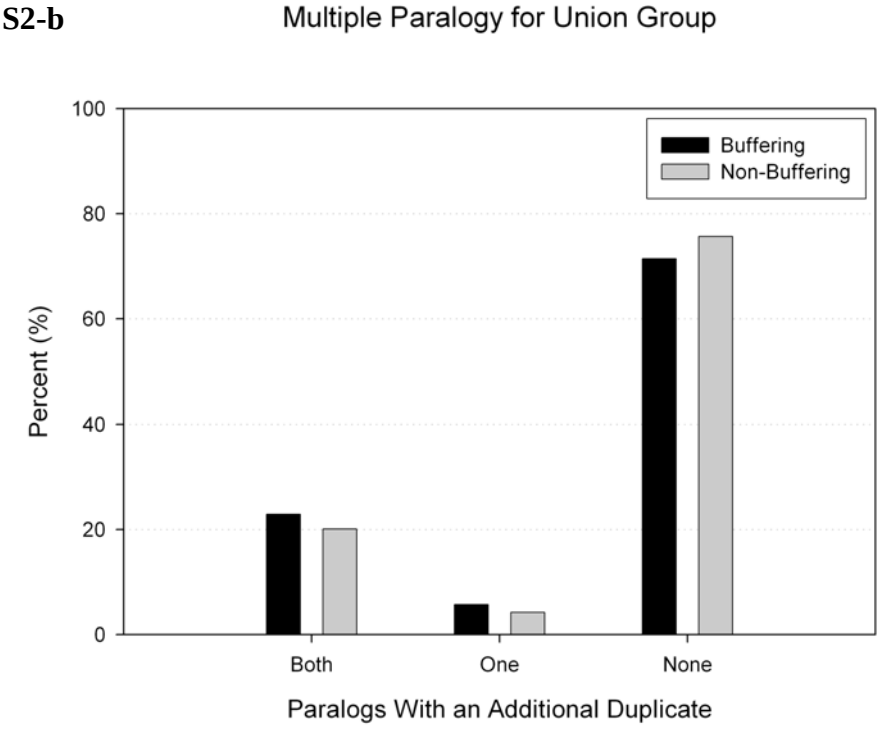


Figure S2 – The affect of multiple paralogy on epistasis. WGD paralogs were divided into three groups based on the presence of additional (non-WGD) paralogs. Paralogs were classified as being: Both (both members of a pair of WGD paralogs have additional duplicates in the *S.cerevisiae* genome), One (only one pair member has additional duplicates), or None (neither WGD paralog has additional duplicates). Depiction of epistasis as assessed by both RSA and GCA (a), or by either RSA or GCA (b) indicates no overall difference in the composition of WGD paralogs with additional, non-WGD, duplicates.

Supplemental References

Gavin, A.C., P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M.A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A.M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J.M. Rick, B. Kuster, P. Bork, R.B. Russell, and G. Superti-Furga. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature*.

Gu, Z., A. Cavalcanti, F.C. Chen, P. Bouman, and W.H. Li. 2002. Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol Biol Evol* **19:** 256-262.

Gu, Z., L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, and W.H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421:** 63-66.

Kellis, M., B.W. Birren, and E.S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428:** 617-624.

Krogan, N.J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J.M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St Onge, S. Ghanny, M.H. Lam, G. Butland, A.M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, and J.F. Greenblatt. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440:** 637-643.

Maere, S., K. Heymans, and M. Kuiper. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21:** 3448-3449.

Reguly, T., A. Breitkreutz, L. Boucher, B.J. Breitkreutz, G.C. Hon, C.L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O.G. Troyanskya, T. Ideker, K. Dolinski, N.N. Batada, and M. Tyers. 2006. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* **5:** 11.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* **12:** 85-94.

St Onge, R.P., R. Mani, J. Oh, M. Proctor, E. Fung, R.W. Davis, C. Nislow, F.P. Roth, and G. Giaever. 2007. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39:** 199-206.