Supplementary material for *"Pathway redundancy and protein essentiality revealed in the S. cerevisiae interaction networks"*

A: Comparison between dense and connected pathway models

In order to compare the dense pathway and the connected pathway models, we executed the model finding algorithm under both assumptions with identical parameters. **Table S1** shows that by using connected pathways we construct more between-pathway explanations of GIs (3765 vs. 3117), while maintaining the significant functional content of BPMs. The dense pathway models are slightly better in terms of enrichment for cellular compartments (74% vs. 69%), but this comes at the cost of incorporating lower density of GIs between the pathways of BPMs (51% vs. 85%). Since the model score proposed in (Kelley and Ideker, 2005) sums the contributions from the GI and the PI subnetworks, some of the high-scoring models recovered using that score contain very dense pathways, but may be very poor in GIs. This problem is avoided in our approach, as the models are scored solely based on the number of GIs between the pathways. In terms of functional coverage, we found no clear advantage to either method (**Fig. S1**). Due to the significantly better GI coverage of the connectivity approach we used it in all subsequent analysis.

B: Comparison between our model and congruence-based method

The model-based approach, such as the one used here, and congruence scores, based on the similarity of GI profiles, were both shown to be powerful in functional annotation of genes and processes (Kelley and Ideker, 2005; Ye et al., 2005). While the congruence-based method is capable of identifying pathways with coherent genetic data, it does not generate hypotheses about buffering between pathways. This is evident in BPM 104 presented in **Figure S5**. The Csm1/Lsr4 pathway was outlined using the congruence score in (Pan et al., 2006). Our modeling suggests specific buffering between Csm1/Lrs4 and transcription-related complexes CTK1 and Nu4A, which was not recognized previously.

C: Comparison of complex coverage in the PPI network

In order to compare the functional content of the BPMs to that of modules obtained directly from the PI network, we executed the MCODE algorithm (Bader and Hogue, 2003) for detection of complexes on our PI network. We used the MCODE plug-in for Cytoscape (Shannon et al., 2003) with default parameters and obtained 35 complexes**.** We then applied the TANGO algorithm (Shamir et al., 2005) with the same parameters used to analyze our set of BPMs, in order to test functional enrichment in the complexes. As elaborated in the main text, 46.3% of the annotated complexes were enriched in at least one BPM in our analysis. The MCODE complexes covered 53.1% of the complexes. Given the current low coverage of the GIs (<15% of *S. cerevisiae* genes), it is encouraging that a large portion of the known complexes can be found (at least in part) in our models, which are based primarily on GIs.

D: Additional analysis of the essentiality of the pivot nodes

We validated that the enrichment of essential proteins in the collection of pivot proteins is robust to the parameters of pivot selection (**Fig. S3**). The essential pivots tend to have closer functions to their BPMs: when using GO semantic similarity (Lord et al., 2003) as a distance function, the average distance between the pivot and the proteins within its model is much smaller for an essential pivot than for a non-essential one ($p = 8.13 \cdot 10^{-15}$ using rank-sum test). Examples of essential pivots that are linked to BPMs that consist almost entirely of non-essential genes are shown in **Figs. 2B-D**.

E: Analysis of physiological properties of the BPM genes

In addition to the analysis of mRNA half-lives and the number of phosphorylation sites which is described in the main text, we analyzed the protein abundance (Ghaemmaghami et al., 2003) and the codon adaptation index (CAI) (Sharp and Li, 1987). As can be seen in **Table S2**, we did not find a statically significant difference for those parameters that could not be explained either by essentiality or high degrees of the proteins involved.

F: Significance filtering of BPMs

Kelley and Ideker (2005) generated a collection of random datasets that preserve the degree distributions in the original GI and PI networks, and accepted only models whose

score exceeded the 95th percentile of the highest model scores obtained on the randomized datasets. On our datasets we found that only a very small number of BPMs survived such filtering (12 and 5 models, when using the dense and the connected pathways, respectively). (Note that the number connectivity-based models is still larger than the density-based one under this aggressive filtering). The fact that the original Kelley-Ideker method - with the original significance filtering - also produces much fewer models on our dataset than in their original analysis (12 vs. 360) is probably due to the larger dataset that we used. The filtering method that we have employed is based on sampling of random groups matched in their size to the BPMs (see Materials and Methods). It produces more models, yet it provides meaningful models with clear biological relevance, as evident in the analysis of functional enrichment and of phenotypic coherence of the pathways.

References

Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737-741.

Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, **23**, 561-566.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275-1283.

Pan, X., Ye, P., Yuan, D.S., Wang, X., Bader, J.S. and Boeke, J.D. (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell*, **124**, 1069-1081.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y. and Elkon, R. (2005) EXPANDER--an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.

Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**, 1281-1295.

Ye, P., Peyser, B.D., Pan, X., Boeke, J.D., Spencer, F.A. and Bader, J.S. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol*, **1**, 2005 0026.

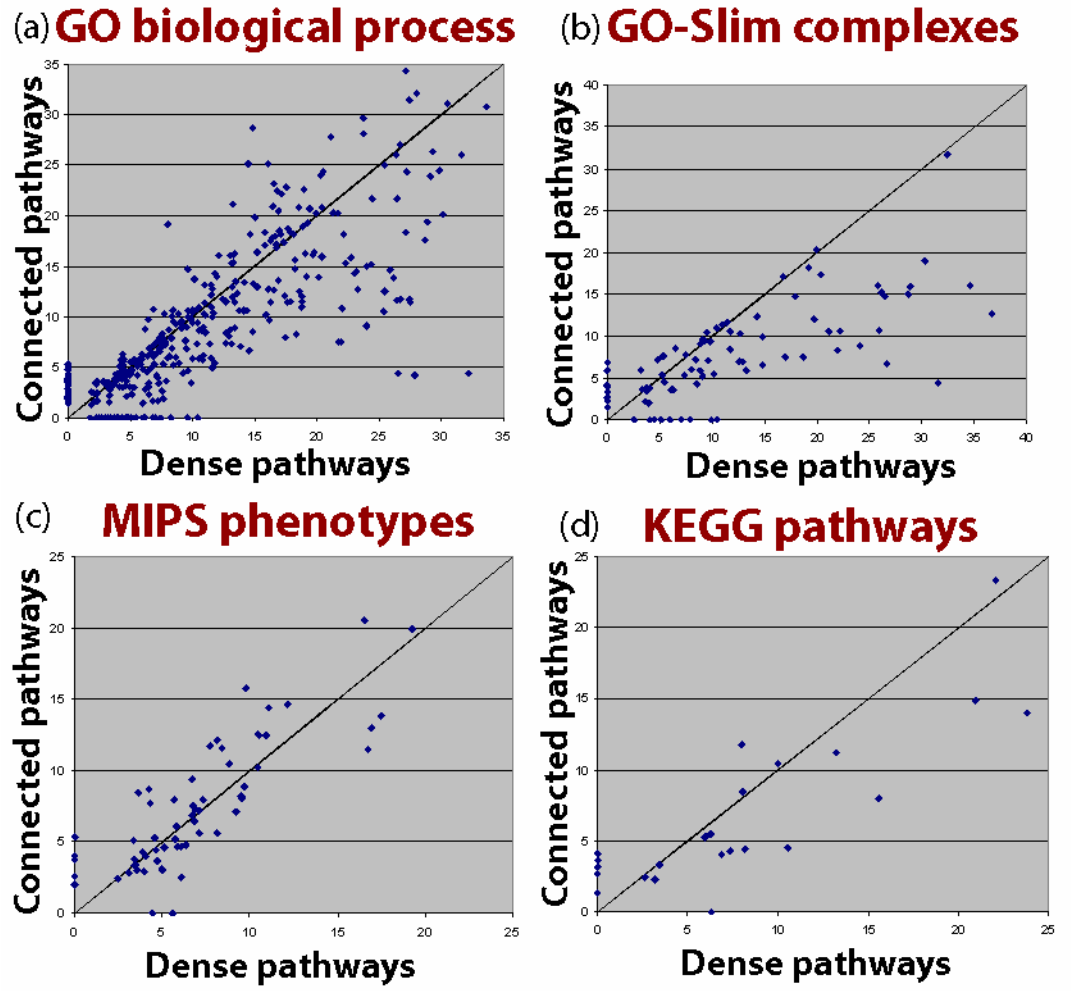

Figure S1:
Functional enrichment in the density-based and the connectivity-based models. A comparison between the dense pathways (Kelley and Ideker, 2005) and the connected pathways (this study) models in terms of the enrichment for diverse functional annotations. Each dot gives the enrichment of one functional category (logarithm base 10 of p-value enrichment) in the set of BPMs produced under each model. (a) Level 7 of the GO "biological process" ontology; (b) GO-slim molecular complexes (obtained from SGD) (c) Deletion phenotypes obtained from MIPS (d) KEGG molecular pathways. Using a sign-test, we found an advantage for using dense pathways only for GO-Slim complex annotations.

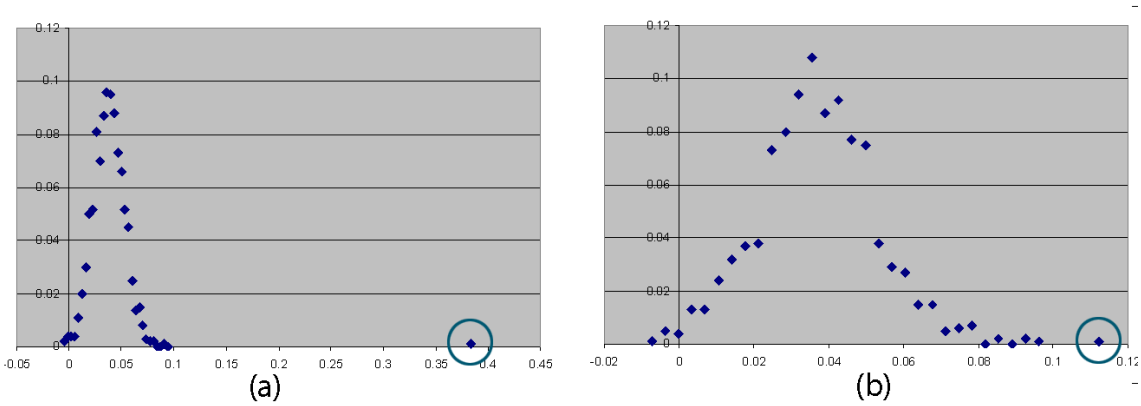

Figure S2: Significance of phenotype coherence. Empirical assessment of the significance of the phenotype coherence within **(a)** and between **(b)** pathways that belong to the same BPM. The density plots represent the phenotype coherence scores (Pearson correlation between fitness patterns) obtained in 1,000 randomized samples of collections of connected pathways with sizes matching those of the pathways within the BPM collection. X-axis: score, y-axis: number of samples with that score. The scores obtained in the BPM collection are circled.

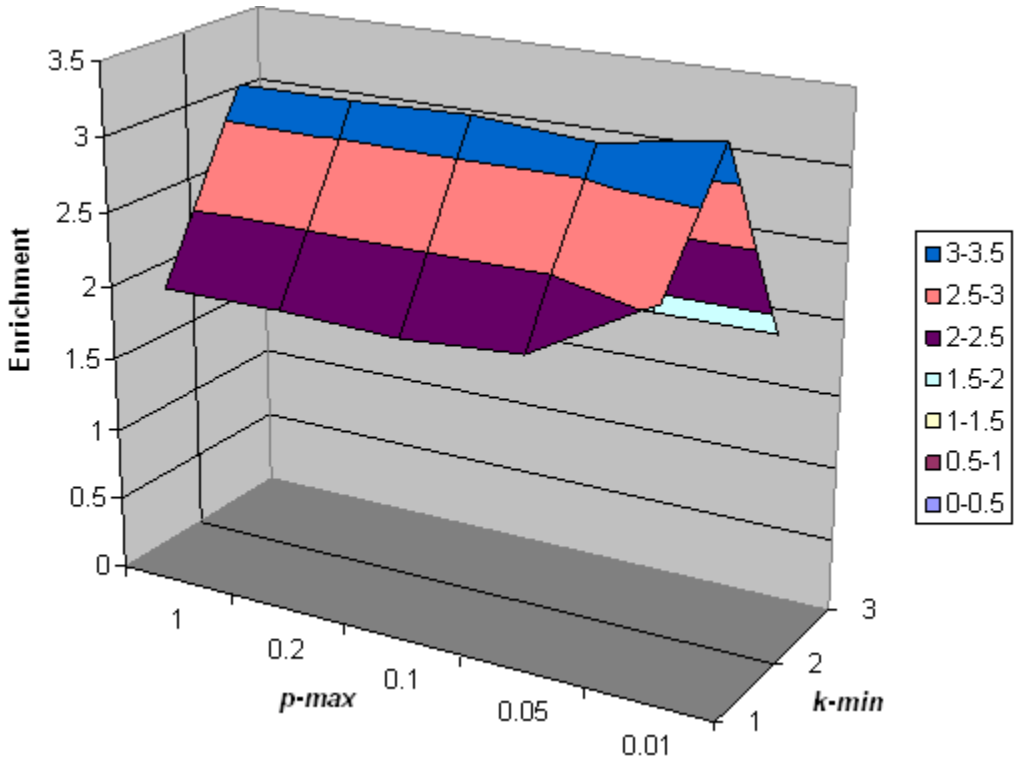

Figure S3: Robustness of pivot selection parameters. The enrichment of essential genes within the group of pivot nodes, as function of the parameters used for pivot selection. Enrichment is calculated as the observed number of essential genes divided by the expected number.

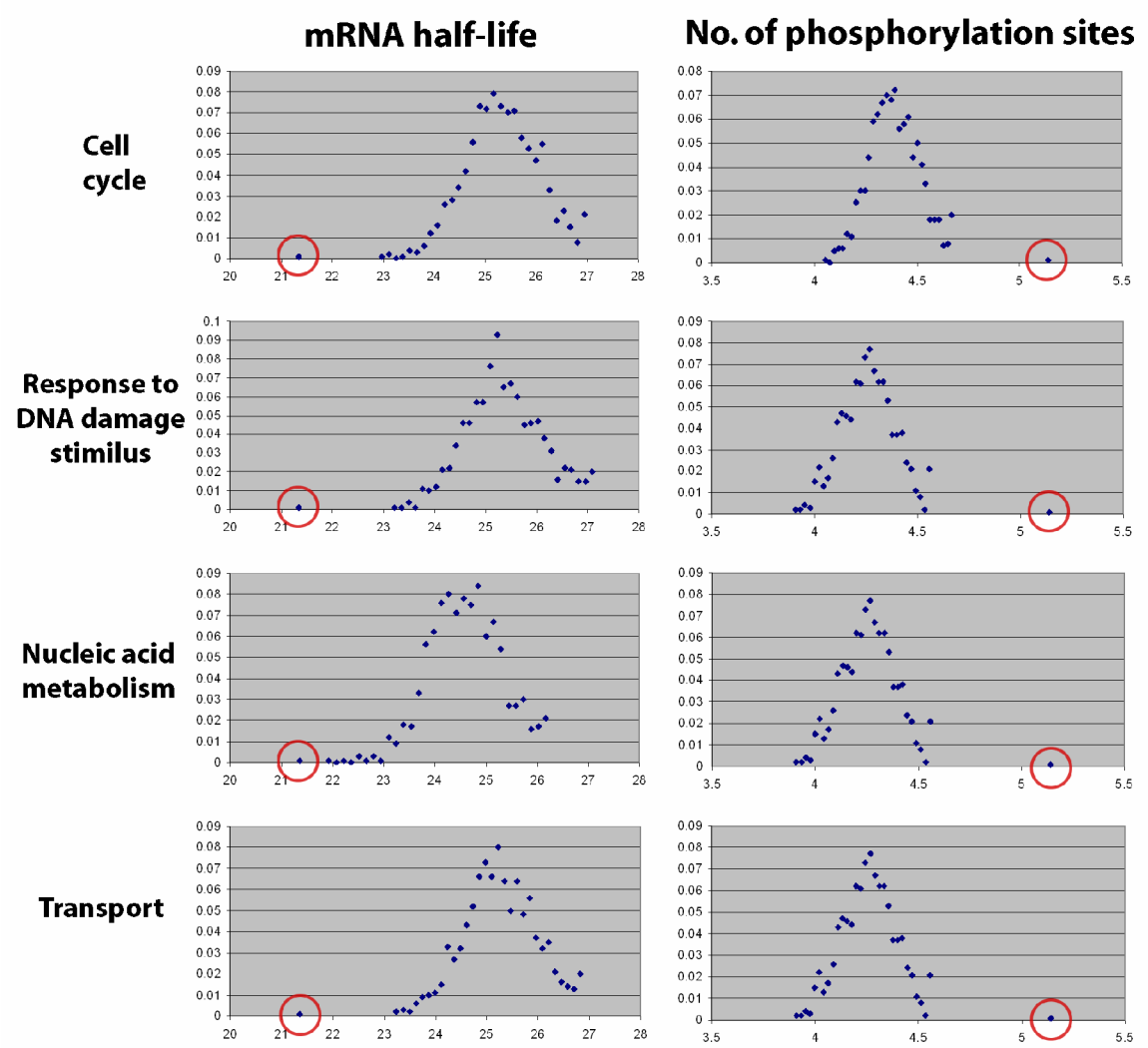

Figure S4: Assessment of the results reported for mRNA half-life and number of phosphorylation sites for BPM genes. In each test a random set of 1000 genes was sampled with the same number of genes from a specific GO biological process category as in the BPM gene set. In each plot the X-axis represents the attribute value (mRNA half-life or number of phosphorylation sites) and the Y-axis the fraction of the samples with that value. The values obtained for the BPM genes are circled.

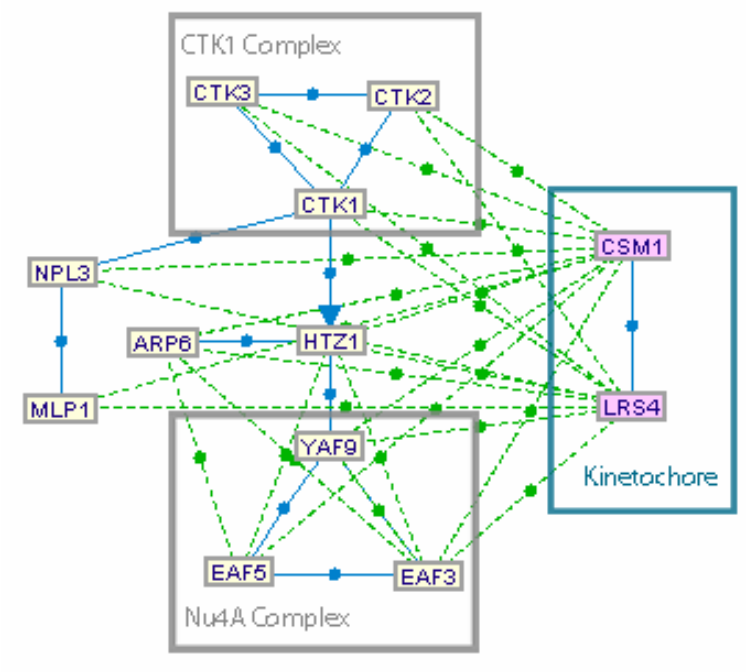

Figure S5: BPM 104. This model shows specific buffering between a kinetochore subunit Csm1/Lsr4 and parts of the CTK1 and Nu4A complexes. See Figure 2 for coloring legend.

		Models generated using dense pathways	Models generated using connected pathways
No. of models		270	140
No. of genes		728	711
Average model size		11.82	16.02
GIs between pathways		3117	3765
PIs within pathways		2424	1709
GI density between pathways		0.51	0.85
PI density within pathways		0.79	0.75
Average no. of genes with no inter-pathway GI		3.89	0.0071
Average no. of genes with at most one inter-pathway GI		5.11	2.16
Functional enrichment in GO bp	BPMs	196/270 (72.6%)	100/140 (71.4%)
	pathways	177/540 (32.7%)	100/280 (35.7%)
Functional enrichment in GO cc	BPMs	201/270 (74.4%)	97/140 (69.3%)
	pathways	194/520 (37.3%)	94/280 (33.6%)

Table S1: Density-based vs. connectivity-based models. A comparison between the models identified using dense pathways and connected pathways. "GO bp"; "biological process" ontology in GO. "GO cc": GO "cellular compartment" ontology. Pathways and modules were considered enriched if they had a hypergeometric p-value < 0.05.

Protein abundance			Without control		Controlling for essentials		Controlling for PI degrees	
Group	No.	Average	Expected	*p*-value	Expected	*p*-value	Expected	*p*-value
All pivots	101	32364.94	11593.58	0.225	14683.10	0.88	20730.84	0.076
Essential pivots	59	7996.61	18524.74	0.695	17616.95	0.07	**21531.52**	**<0.001**
Non-essential pivots	42	66596.64	9945.70	0.160	10607.90	0.054	20843.31	0.764
BPM genes	704	12960.52	11955.80	0.130	12077.71	0.122	**15755.82**	**<0.001**

CAI			Without control		Controlling for essentials		Controlling for PI degrees	
Group	No.	Average	Expected	*p-value*	Expected	p-value	Expected	p-value
All pivots	124	182.74	163.63	0.227	171.51	0.680	203.61	0.136
Essential pivots	72	150.22	180.69	0.622	177.49	0.038	**196.31**	**<0.001**
Non-essential pivots	52	227.77	**160.10**	**0.007**	**160.91**	**0.002**	209.81	0.262
BPM genes	850	156.86	165.27	0.203	164.64	0.860	**180.85**	**<0.001**

mRNA half-life			Without control		Controlling for essentials		Controlling for PI degrees	
Group	No.	Average	Expected	*p-value*	Expected	*p-value*	Expected	*p-value*
All pivots	97	21.80	25.54	0.013	22.68	0.338	22.99	0.400
Essential pivots	54	22.43	19.31	0.155	19.56	0.304	20.03	0.178
Non-essential pivots	43	21.02	26.98	0.075	27.05	0.004	24.07	0.120
BPM genes	**658**	**21.35**	**26.21**	**$1.95 \cdot 10^{-9}$**	**25.38**	**<0.001**	**23.94**	**<0.001**

Phosphorylation sites			Without control		Controlling for essentials		Controlling for PI degrees	
Group	No.	Average	Expected	*p-value*	Expected	*p-value*	Expected	*p-value*
All pivots	103	5.17	4.18	0.285	4.32	0.220	4.86	0.976
Essential pivots	58	4.50	4.47	0.843	4.48	0.376	4.80	0.116
Non-essential pivots	45	6.02	4.12	0.009	4.14	0.010	4.82	0.074
BPM genes	**740**	**5.14**	**4.03**	**$6.25 \cdot 10^{-9}$**	**4.21**	**0.002**	**4.51**	**<0.001**

Table S2: The physiological properties of pivot proteins and of genes within BPMs. Essential and non-essential pivots are compared to the pool of all essential and non-essential genes, respectively. *p*-values without control are computed using Wilcoxon rank-sum test. Controlling for essential genes and for PI degrees is done as described in Materials and Methods. Significant figures are in bold.