Supplementary Materials: Assessing the role of multi-protein complexes in determining phenotype

Nolwenn LeMeur [1,**] and Robert Gentleman [2]
[1]EHESP, 35042 RENNES Cedex, FRANCE
[2]Genentech, Bioinformatics and Computational Biology Department,
1 DNA Way- South San Francisco, CA 94080, USA

Intro

This document supplements the paper entitled Assessing the role of multi-protein complexes in determining phenotype in which we promote the concept that phenotypic is related to high levels cellular organization untis, such as multi-protein complexes. We propose computational methods and present the use of R packages [1; 2] to disentangle the multi-protein complexe contribution to disease phenotype in *Saccharomyces cerevisiae*.

Understanding regulatory mechanisms and sensitivity of cellular organizational units in complex biological systems is an important challenge. In medicine, in particular, it will lead to greater understanding of the processes involved in some diseases. In that context, we have demonstrated the importance of multi-protein complexes in synthetic lethality and characterized some of the biological mechanisms involved [3]. Other studies also suggest that some control of phenotype can be usefully attributed to multi-protein complexes rather than genes or pathways [4–8] and hence may help provide elucidation of the underlying roles or mechanisms that directly control changes in phenotype. In the long term, in the case of disease phenotype, knowledge of organizational units involved in the disease regulator mechanisms will enable us to identify biological targets for drug therapy and improve the specificity and efficacy of those drugs.

The challenge of understanding cellular regulatory mechanisms by cellular organizational units is difficult due to the size of the underlying biological network and the heterogeneous nature of the control mechanisms involved [9; 10]. Indeed, many genes are pleiotropic and their product play many roles in the cell. It may then not be clear which of those different functions is directly related to the change in phenotype [6; 11]. Moreover, epistasis can mask the phenotypic effect of a gene, obscuring the relationship between gene and phenotype [10]. Tools are therefore needed to identify which function of a gene relates to a disease phenotype. More generally, systems biology approaches are now required to understand the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of that system.

Datasets

We studied the following phenotypic datasets:

- **Essential genes in YPD** media by Giaever *et al.*

  For the cellular organizational units we used with:

  - ScISI: database of yeast multi-protein complexes available in the *ScISI* R package, downloadable from the Bioconductor website

*to whom correspondence should be addressed: Nolwenn.LeMeur-Rouillard@ehesp.fr

– KEGG: Kyoto Encyclopedia of Genes and Genomes that provides pathway database [12].

We note that for the present time every interactions (fitness growth or complex menbership) are represented by discrete value. As an example, for fitness data we set to 0 a gene with no effect or an alleviating effect, and we set to 1 a gene whose deletion result in growth defect or lethal phenotype.

Protein complexes and phenotypes

Coverage

Table S1 shows the number of genes tested in the different experiment and represented in our interactome. We note that overall the coverage is reasonnable for statistical computating. However we ill guther down see that per experimetal condition some coverage are small and this might influence the computation and validity of statistical results.

	Phenotype	Interactome
1	1101	695
2	184	100
3	803	312
4	4723	1016

Table S1: Coverage of conditionnally essential genes in our interactome. Each row corresponds to a specific growth condition. Phenotype: number of conditionnally essential genes estimated per condition; Interactome: number of conditionnally essential genes present in the interactome. Essential: essential genes in rich medium; HI: haploinsufficient; .

Essential genes

The list of *S. cerevisiae* rich media essential genes was obtained from the *Saccharomyces* Genome Database [13]. Among the 4,918 verified open reading frames (ORFs) believed to compose *S. cerevisiae* genome (source: www.yeastgenome.org - last updated April 2011) 1,101 are classified as essential genes [14]. One can access this dataset from the *SLGI* R package, available from the Bioconductor Project [2].

	Observed	Expected	Size	Odds	P-value (adj)	P-value	Description
GO:0005732	42	21.59	56	5.03	1.87e-05	1.98e-08	small nucleolar ribo...
GO:0005666	17	6.55	17	Inf	3.84e-05	8.11e-08	DNA-directed RNA pol...
MIPS-410.30	16	6.17	16	Inf	6.74e-05	2.13e-07	Pre-replication comp...
apCompGavin2002: 228	18	7.32	19	29.43	8.87e-05	3.75e-07	-
GO:0005656	15	5.78	15	Inf	1.06e-04	5.61e-07	pre-replicative comp...
MIPS-360	28	13.88	36	5.77	2.37e-04	1.50e-06	Proteasome
MIPS-410.35	18	7.71	20	14.70	2.84e-04	2.40e-06	Replication complex
apCompGavin2002: 231	18	7.71	20	14.70	2.84e-04	2.40e-06	-
MIPS-510.120	13	5.01	13	Inf	4.07e-04	3.87e-06	RNA polymerase III
apCompGavin2002: 224	14	5.78	15	22.76	1.35e-03	1.43e-05	-
GO:0046540	22	10.79	28	6.00	1.40e-03	1.63e-05	U4/U6 x U5 tri-snRNP...
apCompGavin2002: 203	11	4.24	11	Inf	2.10e-03	2.66e-05	-
apCompGavin2002: 50	19	9.25	24	6.20	3.72e-03	5.36e-05	-
apCompGavin2002: 12	16	7.32	19	8.68	3.72e-03	5.50e-05	-
GO:0000172	10	3.85	10	Inf	4.39e-03	6.96e-05	ribonuclease MRP com...
GO:0005847	13	5.78	15	10.54	7.83e-03	1.70e-04	mRNA cleavage and po...
GO:0005669	13	5.78	15	10.54	7.83e-03	1.70e-04	transcription factor...

MIPS-360.10.10	13	5.78	15	10.54	7.83e-03	1.70e-04	20S proteasome
apCompGavin2002: 43	13	5.78	15	10.54	7.83e-03	1.70e-04	-
GO:0005849	9	3.47	9	Inf	7.83e-03	1.82e-04	mRNA cleavage factor...
GO:0005655	9	3.47	9	Inf	7.83e-03	1.82e-04	nucleolar ribonuclea...
apCompGavin2002: 205	9	3.47	9	Inf	7.83e-03	1.82e-04	-
GO:0005681	22	11.95	31	3.99	9.41e-03	2.29e-04	spliceosomal complex

Table S2: Multi-protein complexes associated with Essentiality (P-value<0.01). Observed: number of essential genes in the complex; Expected: expected number of essential genes in the complex; Size: total number of genes in the complex; Odds: odds ratios; P-value (adj): adjusted P-value of the Hypergeometric test (bonferroni correction); P-value: P-value of the Hypergeometric test; Description: annotation of for the given protein complex. Note that when the multi-protein complex is entirely composed of essential genes (Observed = Size) the odds ratio are infinite (Inf).

Haploinsufficient genes

The list of haploinsufficient genes was extracted from Deutschbauer *et al.* [5] who found that 184 *S. cerevisiae* genes were haploinsufficient for growth in Yeast extract/Peptone/Dextrose (YPD). The haploinsufficient dataset is included in the *PCpheno* R package, available from the Bioconductor Project [2].

	Observed	Expected	Size	Odds	P-value (adj)	P-value	Description
MIPS-130	7	0.44	8	128.11	7.69e-06	1.01e-08	Chaperonine containi...
GO:0005732	16	3.11	56	7.92	7.69e-06	1.62e-08	small nucleolar ribo...
GO:0005832	7	0.61	11	31.97	1.14e-04	3.61e-07	chaperonin-containin...
GO:0005665	7	0.67	12	25.56	1.96e-04	8.28e-07	DNA-directed RNA pol...
MIPS-510.40.10	7	0.72	13	21.29	3.24e-04	1.71e-06	RNA polymerase II
apCompGavin2002: 223	6	0.67	12	18.05	2.79e-03	1.77e-05	-
GO:0000176	6	0.72	13	15.47	4.24e-03	3.14e-05	nuclear exosome (RNa...

Table S3: Haploinsufficiency can be attributed to some multi-protein complexes. These complexes (curated and predicted) present an over-representation of haploinsufficient genes (p-value<0.01). Observed: number of haploinsufficient genes in the complex; Expected: expected number of haploinsufficient genes in the complex; Size: total number of genes in the complex; Odds: odds ratios; P-value (adj): adjusted p-value of the Hypergeometric test (bonferroni correction); P-value: p-value of the Hypergeometric test; Description: fullname. Note that when the multi-protein complex is entirely composed of haploinsufficient genes (Observed = Size) the odds ratio are infinite (Inf).

Stressful conditions and protein complexes (Dudley et al., 2005)

Dudley *et al.*[6] created a collection of gene-deletion mutants to determine genes that contribute to a particular phenotype in 21 different environmental conditions. Table S3 presents the gene coverage between stress condition data and the interaction.

	Phenotype	Interactome
benomyl	34	19
CaCl2	180	88
CAD	83	45
Caff	208	105
cyclohex	164	78
DTT	5	1
EtOH	75	51
FeLim	35	17

HU	87	52
HygroB	264	108
lowPO4	34	10
MPA	11	6
NaCl	57	28
Paraq	36	22
pH3	16	8
rap	119	51
Sorb	8	2
UV	32	22
YPGal	41	20
YPGly	206	76
YPLac	159	52
YPRaff	31	16

Table S4: Coverage of Dudley's conditionnally essential genes in our interactome. Each row corresponds to a specific growth condition. Phenotype: number of conditionnally essential genes estimated per condition; Interactome: number of conditionnally essential genes present in the interactome. benomyl: 15ug/ml benomyl,microtubule function; CaCl2: 0.7M calcium chloride, divalent cation; CAD: 55uM Cadmium, heavy metal; Caff: 2mg/ml Caffeine; cyclohex: 0.18ug/ml cycloheximide, protein synthesis; DTT: unknown; EtOH YPD + 6% Ethanol; FeLim: irion limited,nutrient limited condition; HU: 11.4mg/ml Hudroxyurea, DNA replication and repair; HygroB: 50ug/ml hygromycin B, aminoglycosides; lowPO4: low phosphate, nutrient limited condition; MPA: 20ug/ml mycophenolic acid, transcriptional elongation; NaCl: 1.2M sodium chloride, general stress condition; Paraq: 1mM paraquat, oxidative stress; pH3: low pH, general stress condition; rap: 0.1ug/ml rapamycin, protein synthesis; Sorb: 1.2M sorbitol, general stress condition; UV: 100J/m2 ultra-violet, DNA replication and repair; YPGal 2% galactose, carbon source; YPGly 3% glycerol, carbon source; YPLac 2% lactate, carbon source; YPRaff 2% raffinose, carbon source.

Table S4 shows the results of the graph theory approach and the Hypergeometric test applied to each of the condition. The first two columns indicates the number of genes that were identified as sensitive by Dudley *et al.*

	Dudley et al (2005)	Interactome	p.value	nb.C 0.01	nb.C 0.05
cyclohex	164	79	0	0	6
FeLim	35	17	0	3	3
MPA	11	6	0.001	0	2
Paraq	36	22	0.001	3	5
YPGal	41	20	0.002	0	1
YPRaff	31	16	0.002	2	4
HU	87	52	0.003	0	5
CaCl2	180	88	0.007	2	7
YPGly	206	76	0.008	3	4
UV	32	22	0.009	1	1
EtOH	75	51	0.012	-	-
YPLac	159	52	0.014	-	-
CAD	83	45	0.027	-	-
lowPO4	34	10	0.037	-	-
pH3	16	8	0.052	-	-
rap	119	51	0.071	-	-
HygroB	264	109	0.145	-	-
Caff	208	105	0.192	-	-
NaCl	57	29	0.244	-	-
benomyl	34	19	0.594	-	-
DTT	5	-	-	-	-
Sorb	8	-	-	-	-

Table S5: Dudley et al. (2005) environmental stress conditions. Each row corresponds an environmental stress condition. The first column indicates the number of mutants with growth defect in Dudley's experiment. The second column indicates the number of those deleted genes in the interactome. The third column presents the p-value obtained by the graph theory test. A p-value <= 0.01 indicates that those deleted genes are not randomly distributed in the multi-protein complexes of the interactome. The fourth and fifth columns indicate the number of multi-protein complexes involved at a FDR adjusted pvalue <= 0.01 and 0.05. The 22 environmental conditions listed are: benomyl: 15ug/ml benomyl,microtubule function; CaCl2: 0.7M calcium chloride, divalent cation; CAD: 55uM Cadmium, heavy metal; Caff: 2mg/ml Caffeine; cyclohex: 0.18ug/ml cycloheximide, protein synthesis; DTT: unknown; EtOH YPD + 6% Ethanol; FeLim: irion limited,nutrient limited condition; HU: 11.4mg/ml Hudroxyurea, DNA replication and repair; HygroB: 50ug/ml hygromycin B, aminoglycosides; lowPO4: low phosphate, nutrient limited condition; MPA: 20ug/ml mycophenolic acid, transcriptional elongation; NaCl: 1.2M sodium chloride, general stress condition; Paraq: 1mM paraquat, oxidative stress; pH3: low pH, general stress condition; rap: 0.1ug/ml rapamycin, protein synthesis; Sorb: 1.2M sorbitol, general stress condition; UV: 100J/m2 ultra-violet, DNA replication and repair; YPGal 2% galactose, carbon source; YPGly 3% glycerol, carbon source; YPLac 2% lactate, carbon source; YPRaff 2% raffinose, carbon source.

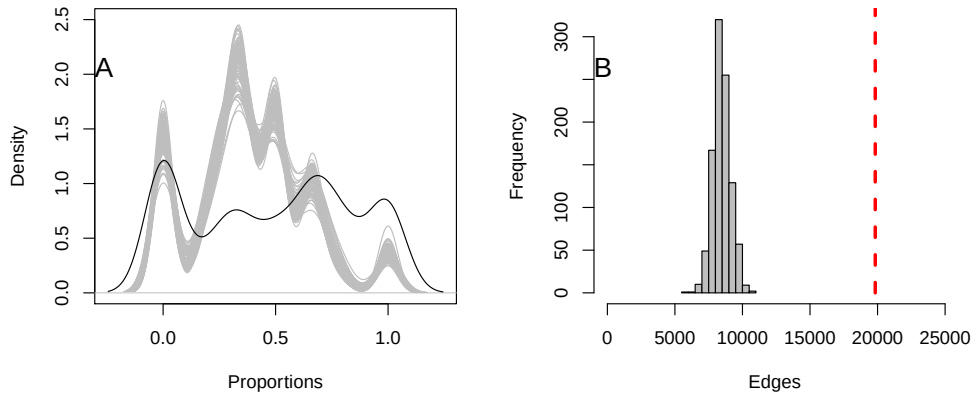

Figure S1: Essential genes random are not randomly distributed among multi-protein complexes. Panel A. Smoothed histograms of the proportion of genes per multi-protein complexes that are associated with a phenotype. The dark line represents the observed data and the light curves represent the permuted data. Only the first 50 simulated density estimates out of 1,000 permutations are displayed for visualization efficiency. Panel B. Distribution of the number of edges, under the null distribution (1,000 permutations) of genes randomly distributed in multi-protein complexes (grey histogram)compared to the number of observed edges, dashed line.

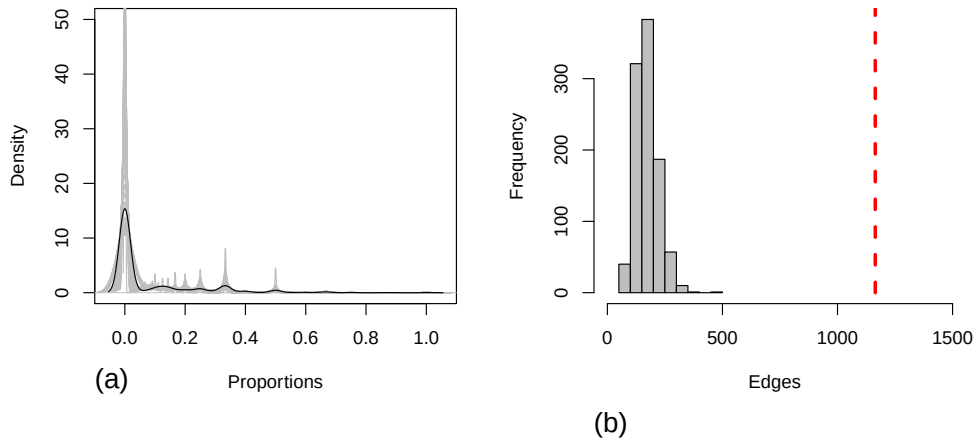

Figure S2: HI genes seem not randomly distributed among multi-protein complexes.

More precisely the multi-protein complexes involved are:

```
---------Condition: cyclohex -----------
MIPS-230.20.10  ADA complex
ADA complex
apCompGavin2002: 5
GO:0000508  NA
GO:0016593  Cdc73/Paf1 complex
GO:0000119  NA
---------Condition: FeLim -----------
MIPS-220  H+-transporting ATPase, vacuolar
GO:0000220  vacuolar proton-transporting V-type ATPase, V0 domain
GO:0000221  vacuolar proton-transporting V-type ATPase, V1 domain
---------Condition: MPA -----------
MIPS-230.20.20  SAGA complex
SAGA complex
---------Condition: Paraq -----------
MIPS-220  H+-transporting ATPase, vacuolar
GO:0000220  vacuolar proton-transporting V-type ATPase, V0 domain
GO:0000814  ESCRT II complex
GO:0000221  vacuolar proton-transporting V-type ATPase, V1 domain
MIPS-90.30  ER assembly complex
---------Condition: YPGal -----------
MIPS-220  H+-transporting ATPase, vacuolar
---------Condition: YPRaff -----------
MIPS-220  H+-transporting ATPase, vacuolar
GO:0000220  vacuolar proton-transporting V-type ATPase, V0 domain
MIPS-90.30  ER assembly complex
apCompHo2002: 31
---------Condition: HU -----------
MIPS-510.40  RNA polymerase II holoenzyme
GO:0000119  NA
apCompHo2002: 8
apCompKrogan2004: 18
MIPS-510.40.20  Kornberg's mediator (SRB) complex
---------Condition: CaCl2 -----------
MIPS-220  H+-transporting ATPase, vacuolar
GO:0000815  ESCRT III complex
GO:0000814  ESCRT II complex
MIPS-260.70  Vps4p ATPase complex (Vps protein complex)
GO:0016593  Cdc73/Paf1 complex
apCompKrogan2004: 1
GO:0000221  vacuolar proton-transporting V-type ATPase, V1 domain
---------Condition: YPGly -----------
apCompGavin2002: 167
apCompGavin2002: 166
MIPS-220  H+-transporting ATPase, vacuolar
GO:0009353  mitochondrial oxoglutarate dehydrogenase complex
---------Condition: UV -----------
MIPS-510.180.10  Nucleotide excision repairosome
```

Stressful conditions and protein complexes (Giaever et al., 2002)

Giaever et al (2002) created a collection of gene-deletion mutants to determine genes that contribute to a particular phenotype in specific environmental conditions. This list is generated from a fitness analysis under six different experimental conditions. See Table S5 for details.

Only a few number of experimental conditions seems to associate phenotypic with multi-protein complexes (Tab. S6). This might partly be explained by the low coverage.

	Phenotype	Interactome
NaCl15a	325	52
NaCl15b	74	20
NaCl5a	148	19
NaCl5b	88	17
lysM5a	253	36
lysM5b	251	45
minimalC5a	103	18
minimalC5b	168	28
minimalPlus15a	92	19
minimalPlus15b	79	15
minimalPlus5a	116	19
minimalPlus5b	208	33
nystatin15a	33	8
nystatin15b	41	7
nystatin5a	147	40
nystatin5b	150	38
pH8g15a	200	49
pH8g20b	153	40
pH8g5a	112	27
pH8g5b	265	63
sorbitol15b	9	2
sorbitol20a	56	7
sorbitol5a	315	55
sorbitol5b	54	10
trpM5a	262	46
trpM5b	303	58
ypg15a	24	6
ypg15b	23	4
ypg5a	19	7
ypg5b	15	5

Table S6: Coverage of Giaever's conditionnally essential genes in our interactome. Phenotype: number of conditionnally essential genes estimated per condition; Interactome: number of conditionnally essential genes present in the interactome. Each row corresponds to an environmental stress condition and different generation time (5, 15). The differents conditions are: ypg: yeast/peptone/galactose 5 gen. rep. a and b; sorbitol: 1.5M Sorbitol (sugar, osmotic stress); NaCl: 1M NaCl (salt, osmotic stress); lysM: lysine minus (lack of required AA); thM: threonine minus (lack of required AA); trpM: tritophanee minus (lack of required AA); minimalPlus: minimal + histidine/leuvine/uracile; minimalC: minimal complete; nystatin: Nystatin (antifungal drug); pH8: pH 8 (alkali stress).

	Giaever et al. (2002)	Interactome	p.value	nb.C 0.01	nb.C 0.05
pH8g15	225	56	0.002	3	6
nystatin15	46	8	0.007	1	3
pH8g5	275	66	0.032	-	-
ypg15	30	6	0.045	-	-
nystatin5	171	45	0.097	-	-
ypg5	23	8	0.098	-	-
minimalPlus15	93	19	0.155	-	-
minimalC5	183	29	0.269	-	-
sorbitol15	59	8	0.278	-	-
sorbitol5	356	62	0.433	-	-
NaCl5	175	27	0.56	-	-
trpM5	343	63	0.612	-	-

NaCl15	334	58	0.614	-	-
lysM5	304	48	0.664	-	-
minimalPlus5	262	42	0.843	-	-

Table S7: Some phenotypic changes induced in environmental stress conditions (Giaever et al. 2002) are tightly associated with multi-protein complexes. Each row corresponds to an environmental stress condition and different generation time (5, 15). The first column indicates the number of mutants with growth defect in Giaever's experiment. The second columns indicates the number of those deleted genes in the interactome. The third columns presents the p-value obtained by the graph theory test. A p-value < 0.01 indicates that those deleted genes are not randomly distributed in the multi-protein complexes of the interactome. The fourth and fifth columns indicate the number of multi-protein complexes involved at a FDR adjusted pvalue <= 0.01 and 0.05. The differents conditions are: ypg: yeast/peptone/galactose 5 gen. rep. a and b; sorbitol: 1.5M Sorbitol (sugar, osmotic stress); NaCl: 1M NaCl (salt, osmotic stress); lysM: lysine minus (lack of required AA); thM: threonine minus (lack of required AA); trpM: tritophanee minus (lack of required AA); minimalPlus: minimal + histidine/leuvine/uracile; minimalC: minimal complete; nystatin: Nystatin (antifungal drug); pH8: pH 8 (alkali stress).

More precisely the multi-protein complexes involved are:

```
---------Condition: nystatin15 -----------
GO:0000813  ESCRT I complex
MIPS-260.70  Vps4p ATPase complex (Vps protein complex)
GO:0000815  ESCRT III complex
---------Condition: pH8g15 -----------
MIPS-260.20  Clathrin-associated protein (AP) complex
GO:0030122  AP-2 adaptor complex
GO:0030121  AP-1 adaptor complex
GO:0005955  calcineurin complex
GO:0030123  AP-3 adaptor complex
MIPS-260.20.10  AP-1 complex
```

Pathways and phenotypes

As described in the article, we computed the two omnibus tests to evaluate whether there is an over-abundance of KEGG pathways with low or high proportions of genes associated with essentiality

References

[1] CRAN. The comprehensive r archive network. http://www.R-project.org.

[2] Bioconductor. Open source software for bioinformatics. http://www.bioconductor.org/.

[3] N. Le Meur and R. Gentleman. Modeling synthetic lethality. *Genome Biology*, 9(9):R135, 2008.

[4] AC. Gavin, M. Boesche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, JM. Rick, AM. Michon, CM. Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[5] AM. Deutschbauer, DF. Jaramillo, M. Proctor, et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, 169:1915–1925, 2005.

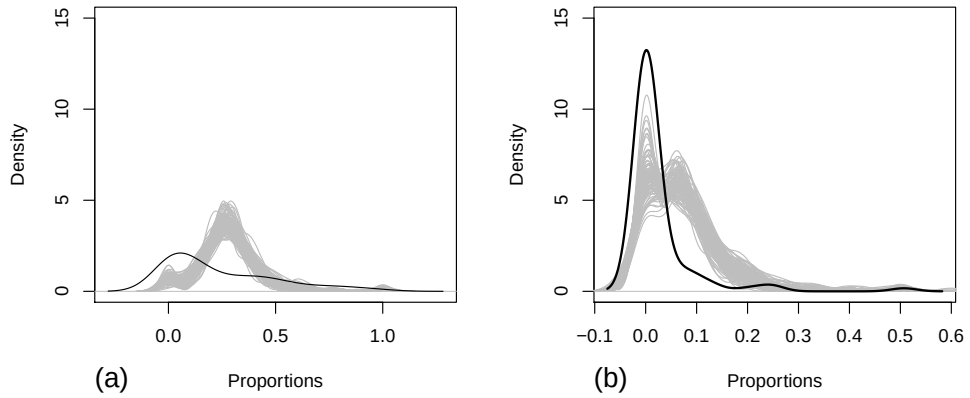

Figure S3: Essential and haploinsufficient genes are not well represented in KEGG. Smoothed histograms of the proportion of genes per multi-protein complexes that are associated with a phenotype. The dark line represents the observed data and the light curves represent the permuted data. Only the first 50 simulated density estimates out of 1,000 permutations are displayed for visualization efficiency. Panel A. Essential genes Panel B Haploinsufficient genes.

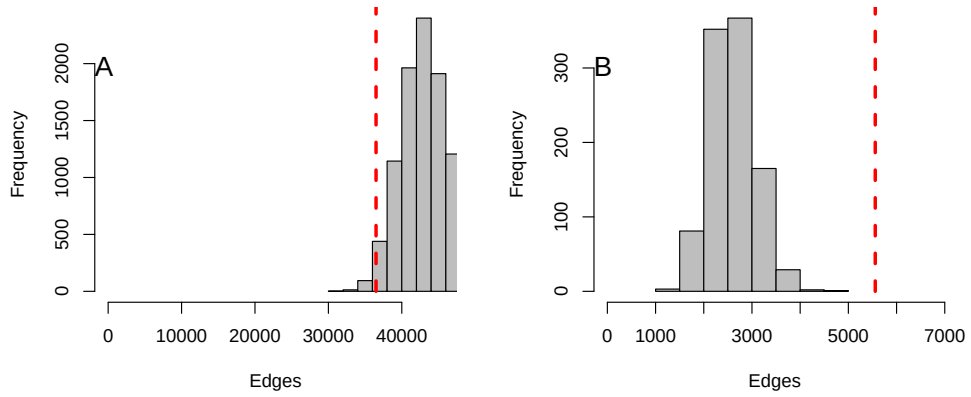

Figure S4: Essential and haploinsufficient genes are not well represented in KEGG. Distribution of the number of edges, under the null distribution (1,000 permutations) of genes randomly distributed in multi-protein complexes (grey histogram) compared to the number of observed edges, dashed line. Panel A. Essential genes. Panel B Haploinsufficient genes.

[6] AM. Dudley, DM. Janse, A. Tanay, R. Shamir, and G. McDonald Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology*, 1:E1–E11, 2005.

[7] V. Spirin, MS. Gelfand, AA. Mironov, and LA. Mirny. A metabolic network in the evolutionary context: multiscale structure and modularity. *PNAS*, 23:8774–8779, 2006.

[8] Magali Michaut, Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda J Andrews, Charles Boone, and Gary D Bader. Protein complexes are central in the yeast genetic landscape. *PLoS Computational Biology*, 7(2):e1001092, February 2011. PMID: 21390331.

[9] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101(9):2981–2986, 2004.

[10] M. Oti and HG Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, 2007.

[11] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, BarabÃąsi AL, Tavernier J, Hill DE, and Vidal M. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, 10 2008.

[12] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[13] R. Balakrishnan, K. R. Christie, M. C. Costanzo, et al. Saccharomyces genome database. 2006.

[14] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, et al. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, Jul 2002.