Supplemental Data

A Protein Complex Network of *Drosophila melanogaster*

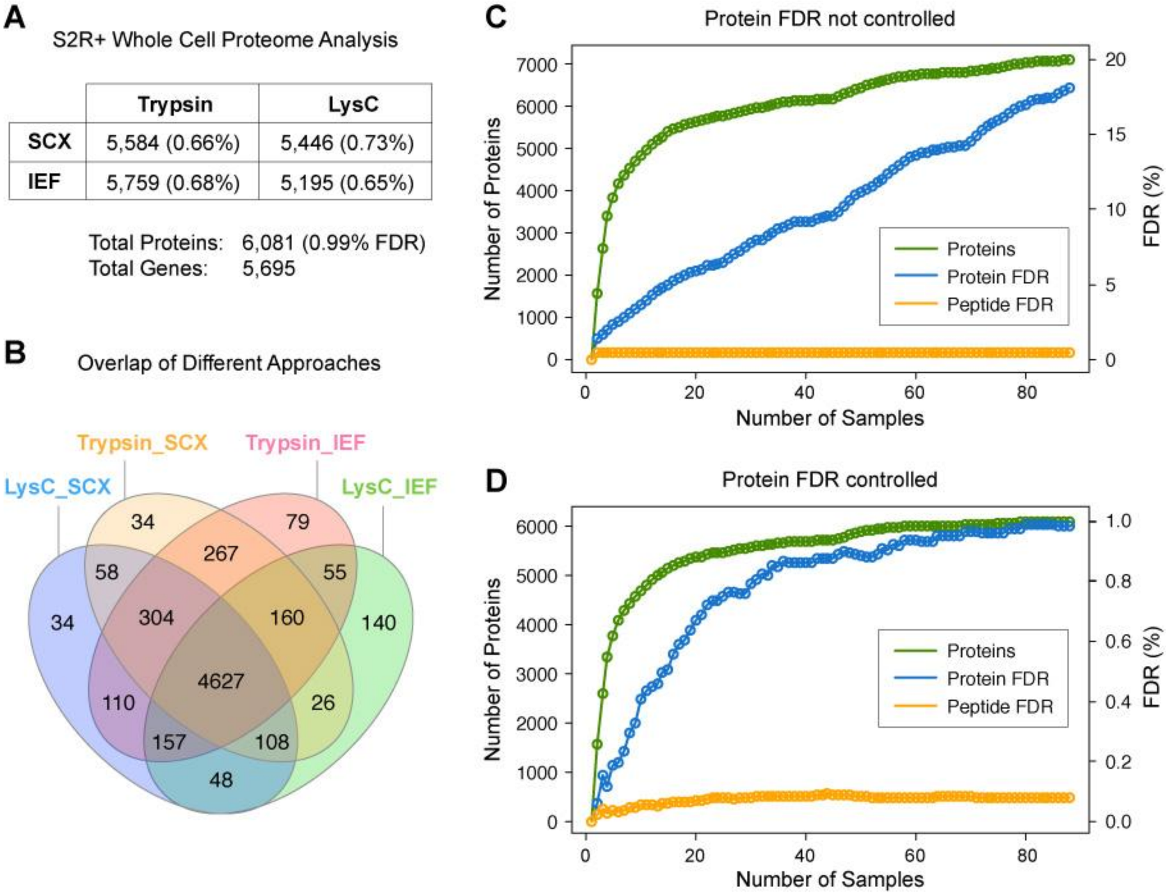
K. G. Guruharsha, J. -F. Rual, B. Zhai, J. Mintseris, P. Vaidya, N. Vaidya, C. Beekman, C. Wong, D. Y. Rhee, O. Cenaj, E. McKillip, S. Shah, M. Stapleton, K. H. Wan, C. Yu, B. Parsa, J. W. Carlson, X. Chen, B. Kapadia, K. VijayRaghavan, S. P. Gygi, S. E. Celniker, R. A. Obar, and S. Artavanis-Tsakonas.

Table of Contents

I. Supplemental Figures

Figure legends for each supplemental figure are found with the corresponding figures.

Figure S1



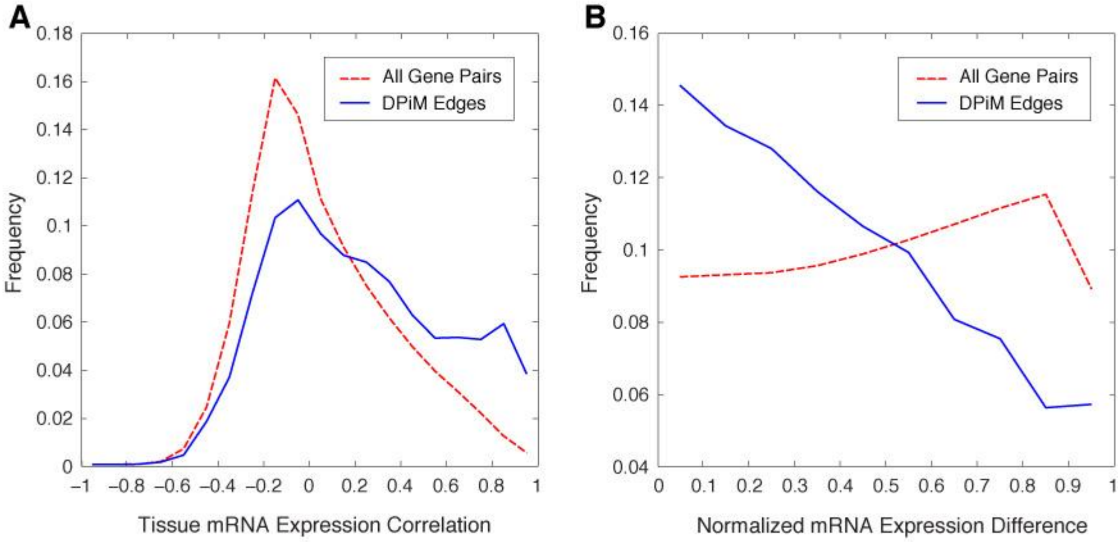
	Trypsin	LysC
SCX	5,584 (0.66%)	5,446 (0.73%)
IEF	5,759 (0.68%)	5,195 (0.65%)

Total Proteins: 6,081 (0.99% FDR)
Total Genes: 5,695

Supplemental Figure S1. S2R+ Whole Cell Lysate Proteome Analysis

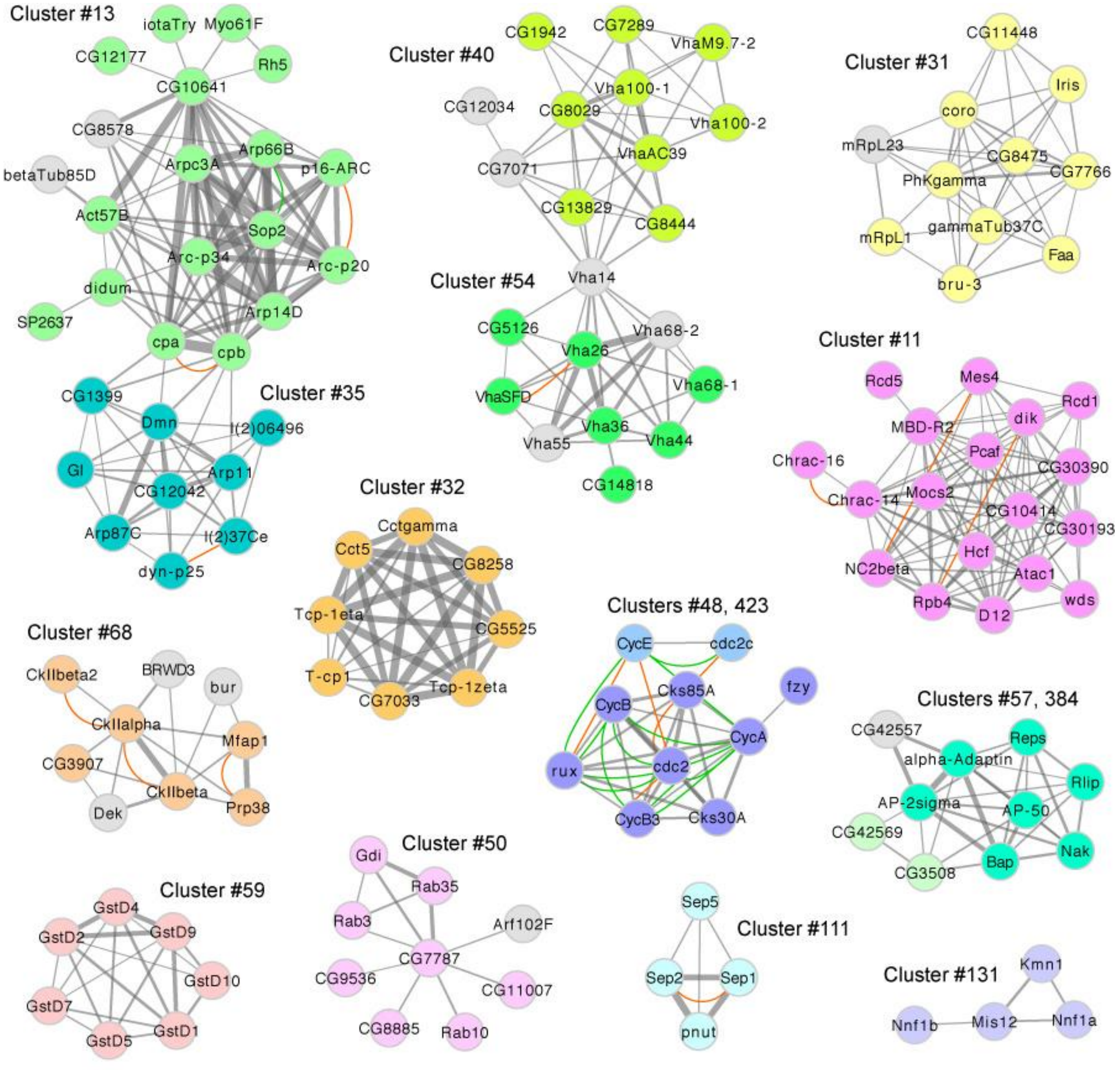
(A) By combining four methods (two proteases and two fractionation procedures), 6,081 proteins (5,695 genes) were identified with 0.99% FDR in whole cell lysates of S2R+ cells. The number of proteins identified by each method is listed with corresponding FDR. (B) Venn diagram of protein overlap of different methods. (C) Combining 86 LC-MS/MS runs requires controlling both peptide- and protein-level FDRs. The peptide FDR was set to 1% for each fraction, but the protein FDR was uncontrolled, which approached 20%. (D) All accepted proteins from Panel C were scored and then filtered to 1% FDR based on reversed protein sequences. The cumulative plot shows a greatly reduced peptide FDR and protein FDR that approaches 1%. The total proteins identified from this whole cell lysate analysis of S2R+ cells was used in the overlap analysis shown in Figure 1C or the main text.

Figure S2



Supplemental Figure S2. Co-regulation of mRNAs corresponding to DPiM interacting proteins. Microarray gene expression profiling data from FlyAtlas encompassing information from 26 *Drosophila* tissues (Chintapalli et al., 2007) was used. (**A**) Distribution of correlation coefficients for mRNAs corresponding to interacting proteins in DPiM compared to all gene pairs. The DPiM distribution is skewed to the right, indicating that transcripts of interacting proteins tend to be co-expressed more frequently than random pairs. This analysis is analogous to that presented in Figure 2B of main text. **(B)** Normalized absolute mRNA expression difference corresponding to DPiM interactors compared to all gene pairs. This analysis suggest that their stoichiometry tends to be more tightly regulated to maintain similar levels of expression, compared to random gene pairs. This analysis is analogous to that presented in Figure 3D of the main text.

Figure S3



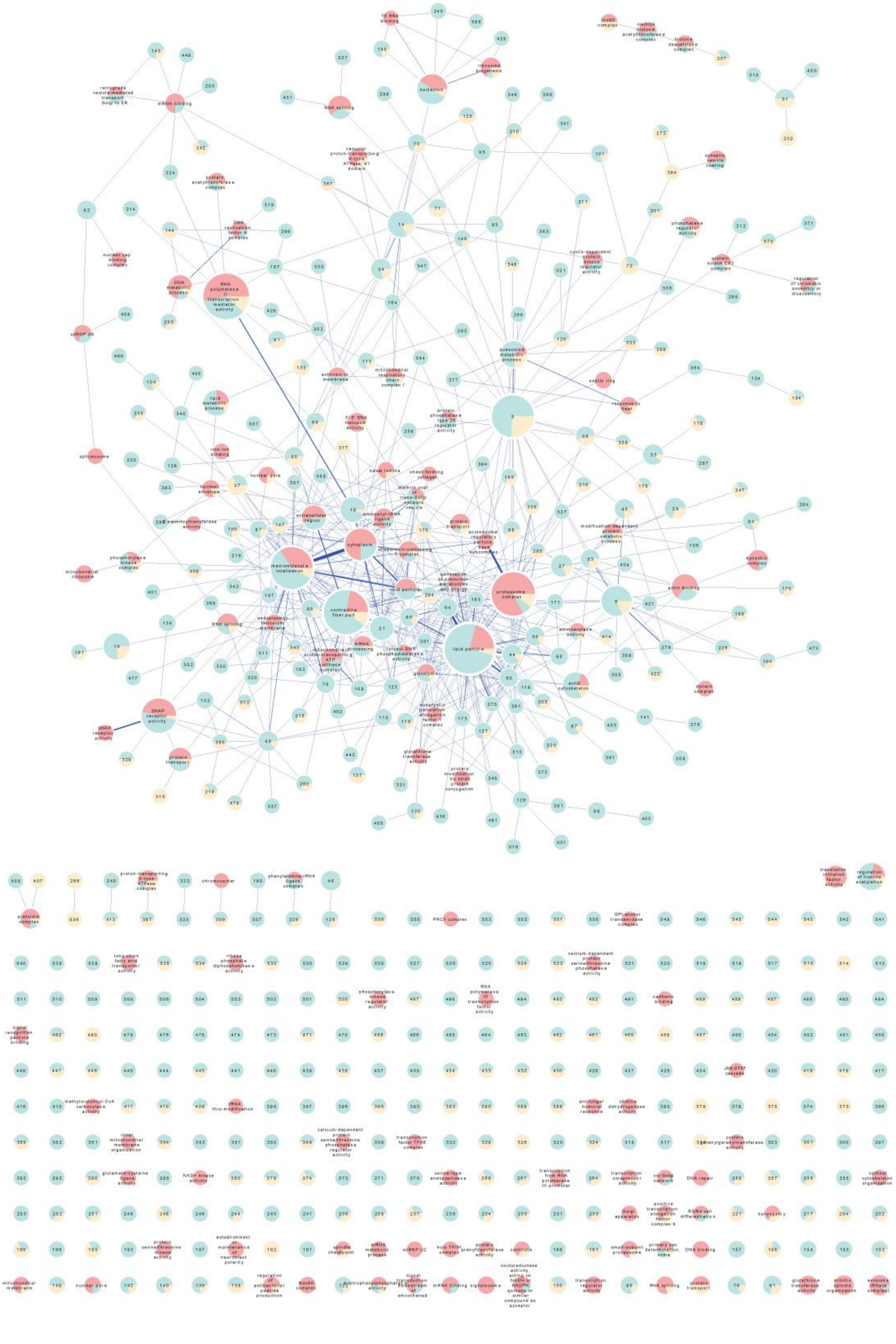
—— DPiM HGSCore
- - - - DPiM sub-threshold
—— Genetic interactions
—— Physical interactions

Cluster #11	regulation of histone acetylation
Cluster #13	regulation of actin polymerization/ depolymerization
Cluster #31	phosphorylase kinase complex
Cluster #32	chaperonin-containing T-complex
Cluster #35	dynactin complex
Cluster #40	vacuolar proton-transporting V-type ATPase, V0 domain
Cluster #48	cyclin-dependent protein kinase regulator activity
Cluster #50	protein transport
Cluster #54	vacuolar proton-transporting V-type ATPase, V1 domain
Cluster #57	vesicle coating
Cluster #59	glutathione transferase activity
Cluster #68	protein kinase CK2 complex
Cluster #111	septin ring
Cluster #131	Ndc80 complex
Clusrer #384	no significant terms
Cluster #423	JAK-STAT cascade

Supplemental Figure S3. Selected examples of complexes from DPiM with different subunits colored according to clusters defined in the map. Proteins shown in grey are not part of the computed cluster. The layout of each complex is guided by the interaction strength. The thickness of the grey line connecting the subunits is proportional to the HGSCore of interaction in DPiM. Additional physical evidence (red lines) and genetic evidence (green lines) from literature are also shown, with line thickness proportional to number of sources supporting it; similar to Figure 4 of the main text.

Supplemental Figure S4. Inter-complex interactions in DPiM. Nodes are proportional to the cluster size and edges proportional to the sum of inter-cluster HGSCores. Only edges where the sum is greater than twice the minimum DPiM HGSCore are retained for clarity. Where GO term enrichment exists (multiple hypothesis testing-adjusted P<0.01), the nodes are labeled with the most significant term; otherwise cluster number is shown instead. The pie chart for each node is made up of 3 wedges: fraction of genes matching the most significantly enriched GO term (pink); fraction of genes having some GO term annotation but not matching the most significantly enriched GO term in the cluster (cyan) and fraction of genes lacking any GO term annotation (yellow); related to Figure 5 of the main text.

Figure S4



7S RNA binding
nucleolus
ribosome biogenesis
RNA splicing
siRNA binding
retrograde vesicle-mediated transport Golgi to ER
vacuolar proton-transporting V-type ATPase, V1 domain
protein acetyltransferase complex
DNA replication factor A complex
nuclear cap binding complex
DNA metabolic process
RNA polymerase II transcription mediator activity
snRNP U6
Ino80 complex
H4/H2A histone acetyltransferase complex
histone deacetylase complex
synaptic vesicle coating
phosphatase regulator activity
protein kinase CK2 complex
regulation of chromatin assembly or disassembly
cyclin-dependent protein kinase regulator activity
queuosine metabolic process
septin ring
response to heat
extrinsic to membrane
mitochondrial respiratory chain complex I
lipid metabolic process
protein phosphatase type 2A regulator activity
3'-5' DNA helicase activity
iron ion binding
spliceosome
nuclear pore
nuclear envelope
basal lamina
sheet-forming collagen
clathrin coat of trans-Golgi network vesicle
extracellular region
aminoacyl-tRNA ligase activity
protein transport
proteasome regulatory particle, base subcomplex
modification-dependent protein catabolic process
palmitoyltransferase activity
phosphorylase kinase complex
mitochondrial ribosome
macromolecule localization
cytoplasm
chaperonin-containing T-complex
generation of precursor metabolites and energy
lipid particle
proteasome complex
dynactin complex
actin binding
contractile fiber part
RNA splicing
endoplasmic reticulum membrane
aminoacylase activity
mitochondrial proton-transporting ATP synthase complex
mRNA processing
tyrosyl-DNA phosphodiesterase activity
glycolysis
actin cytoskeleton
dynein complex
eukaryotic translation elongation factor complex
SNAP receptor activity
SNAP receptor activity
protein transport
glutathione transferase activity
protein modification by small protein conjugation
proton-transporting V-type ATPase complex
chromocenter
phenylalanine-tRNA ligase complex
translation initiation factor activity
regulation of histone acetylation
prefoldin complex
PRC1 complex
GPI-anchor transamidase complex
long-chain fatty acid transporter activity
ribose phosphate diphosphokinase activity
calcium-dependent protein serine/threonine phosphatase activity
phosphorylase kinase regulator activity
RNA polymerase III transcription factor activity
cadherin binding
signal recognition particle binding
JAK-STAT cascade
methylcrotonoyl-CoA carboxylase activity
tRNA thio-modification
antifungal humoral response
choline dehydrogenase activity
inner mitochondrial membrane organization
calcium-dependent protein serine/threonine phosphatase regulator activity
transcription factor TFIIE complex
protein geranylgeranyltransferase activity
glutamate-cysteine ligase activity
NADH kinase activity
serine-type endopeptidase activity
transcription from RNA polymerase III promoter
transcription corepressor activity
cis-Golgi network
DNA repair
cortical cytoskeleton organization
Golgi apparatus
positive transcription elongation factor complex b
R7 cell differentiation
karyogamy
protein serine/threonine kinase activity
establishment or maintenance of neuroblast polarity
spindle checkpoint
miRNA metabolic process
snRNP U2
holo TFIIH complex
protein prenyltransferase activity
centriole
small-subunit processome
primary sex determination, soma
DNA binding
mitochondrial membrane
nuclear pore
regulation of antibacterial peptide production
Ndc80 complex
nitrophenylphosphatase activity
signal transduction downstream of smoothened
mRNA binding
signalosome
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor
transcription regulator activity
RNA splicing
protein transport
glutathione transferase activity
mitotic spindle organization
exosome (RNase complex)

II. Supplemental Tables

All supplemental tables, except table S7 (provided below), are in Microsoft Excel and provided as downloadable files.

Supplemental Table S1. Mass spectrometry data from 3,488 coAP-MS experiments used for the DPiM analysis. Proteins observed due to LC/MS carry-over have been removed as described in Experimental Procedures. (Microsoft Excel file; related to Figure 1 of the main text).

Supplemental Table S2. List of proteins identified by combining four methods for whole cell proteome analysis of S2R+ cells. A total of 6,081 proteins from 5,695 genes were identified with 0.99% FDR. The unique and total peptides identified for each protein are also listed. (Microsoft Excel file; related to Figure 1 of the main text).

Supplemental Table S3. A total of 209,912 pairs of co-purifying proteins were observed among 4,927 *Drosophila* proteins in our data set. HGSCores of all these interactions are listed. The 10,969 high-confidence co-complex membership interactions (0.05% FDR) involving 2,297 *Drosophila* proteins are highlighted in blue. Additional evidence from DroID is also listed for each interaction; (Microsoft Excel file) related to Figure 2 of the main text.

Supplemental Table S4. MCL-derived 556 putative *Drosophila* protein complexes in DPiM. A total of 153 protein clusters are enriched for features such as GO terms, KEGG pathways or Pfam/InterPro domains (multiple hypothesis testing -adjusted P<0.01). Enriched terms for each cluster along with p values are shown; (Microsoft Excel file) related to Figure 2 of the main text.

Supplemental Table S5. Validation of selected DPiM interactions in Human Embryonic Kidney-293F cells. List of pairwise interactions found as bait-prey interactions in DPiM mass spectrometry data, and corresponding human ortholog affinity purification in HEK-293F cells; (Microsoft Excel file) related to Figure 3 of the main text.

Supplemental Table S6. Proteasome subunit classification and mass spectrometry data matrix from 32 individual proteasome subunit bait experiments. Total peptide counts from two replicate affinity purification experiments are shown for each classified proteasome subunit; (Microsoft Excel file) related to Figure 4 of the main text.

Supplemental Table S7. Comparison of protein domains in the eIF3 complex and CSN complex in yeast, fly and human. In the eIF3 complex (Table S6A), many domains remains unchanged and conserved, PCI expands from one in yeast to four in flies and humans, while a few protein domains are seen in only in flies and humans, probably reflecting metazoan-specific functions. In the CSN complex (Table S6B), there is a linear growth of PCI domains and two novel domains are seen in Human, while the CSN proteins containing these novel domains (CSN1a, CSN1b and CSN8) are not part of the complex in DPiM; related to Figure 6 of the main text.

Table S7A. Eif3 Complex

Domain	Yeast	Fly	Human
WD40	4	5	4
PCI	1	4	4
eIF-3c_N	2	1	2
eIF2A	1	1	1
RRM_1	2	2	2
eIF3g	1	1	1
Pfam-B_4134	1	1	1
eIF3_subunit	1	1	1
Pfam-B_2213	1	1	-
Mov34	-	2	2
eIF-3_zeta	-	1	1
Paf67	-	1	1
PCI_Csn8	-	1	1
Pfam-B_2060	-	1	1
eIF3_N	-	1	1

Table S7B. CSN complex

Domain	Yeast	Fly	Human
PCI	2	4	6
Pfam-B_8170	1	1	1
Mov34	1	2	2
RPN7	-	?	1
PCI_Csn8	-	?	1

III. Supplemental Experimental Procedures

Construction of the pMK33-C-FLAG-HA Acceptor Vector

The pMK33-C-TAP vector (Veraksa et al., 2005) was modified to include splice acceptor and donor sites and prokaryotic promoter sequences. The vector was cut with *BamHI* and *SpeI*. The splice acceptor site and prokaryotic promoter sequences (shown below) were TA-cloned into pCR2.1. The product was digested with *BamHI* and *SpeI* and the following restriction fragment inserted ligated into the *BamHI* and *SpeI* ends of pMK33-C-TAP vector.

5'-
GGATCCATAACTTCGTATAGCATACATTATACGAAGTTATAGATCCAATATTATTGAAGCATT
TATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATA
GGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTGGATCTCGAGCTCAAGCTT
CGAATTCAGGGTTTCCTTGACAATATCATACTTATCCTGTCCCTTTTTTTTCCACAGCTACC
GGTCGCGACTAGT-3'

The resulting vector was modified to accept any of a number of protein tags. We removed the C-TAP and *actin 5c* sequences by digesting with *SpeI* and *NotI*. The actin 5c element was added back using a *SpeI*/*NotI* fragment from pMK33-N-TAP and ligating the two components resulting in a vector called pMK33-C-TAG. The FLAG-HA sequence used in this study was inserted at the *SpeI* site. The following custom primers were used to generate the FLAG-HA sequence:

Primer 1:
5'-CTAGTGACTACAAAGACGATGACGACAAGGTCAAACTTTACCCATACGATGTTC
CAGATTACGCTGCTGCTTAGA-3'

Primer 2:
5'-CTAGTCTAAGCAGCAGCGTAATCTGGAACATCGTATGGGTAAAGTTTGACCTTGT
CGTCATCGTCTTTGTAGTCA-3'

The resulting Acceptor Vector was named pMK33-C-FLAG-HA

Construction of the FLAG-HA expression clone set

We transferred ORFs from the BDGP *Drosophila* melanogaster expression-ready clone set to the pMK33-C-FLAG-HA Acceptor Vector for expression of FLAG-HA fusion proteins (Yu et al., 2011). For recombination reactions, 200 ng of expression-ready Donor clone and pMK33-C-FLAG-HA Acceptor Vector were recombined in a final volume of 10 μl for 15 minutes at 25°C in a thermal cycler in the presence of *Cre* recombinase (0.2 units) and recombinase buffer supplemented with BSA (0.1 mg/ml) (Clontech #631614) according to Clontech manual PT3460-1. *Cre* recombinase was inactivated by incubating the reaction at 70°C for 10 minutes. From this reaction, 5 μl was transformed into chemically competent TAM-1 cells in a 96-well plate format (Active Motif #11096) and selected for Chloramphenicol resistance. Each clone was sequence verified to check for target mismatches using BigDye Terminator v3.1 ready reaction mix (Applied Biosystems #4337457) and the sequencing primer: 5'-GCCAATGTGCATCAGTTGTGGTC-3'. Sequencing samples were analyzed on a conventional capillary electrophoresis instrument (*e.g.*, ABI 3730/3730xl DNA Analyzer). Glycerol stocks were generated and stored for each isolate. A collection of 7,120 C-terminus FLAG-HA epitope-tagged clones representing 6,500 unique genes were thus generated and named "Universal Proteomics Resource". Clones can be obtained from the *Drosophila* Genome Research Center: https://dgrc.cgb.indiana.edu/vectors/store/infusion.html.

Analysis of the Universal Proteomics Resource clone set indicates that the median of the expression (FPKM calculated from the modENCODE RNA-seq data) distribution for the cDNAs is 65.22 whereas the median of the FPKM distribution for annotated transcripts is 34.98. The median of the length distribution for cDNAs is 1,134 bp whereas the median of the length distribution for annotated genes is 1,482 bp.

Transfection and recombinant expression of bait proteins in cell culture

Plasmid DNA was prepared using the PureLink™ HQ Mini Plasmid Purification Kit (Invitrogen: K2100-01) or QIAprep Spin Miniprep Kit (Qiagen: 27106). Individual clones were transiently transfected into a 54 ml culture of *Drosophila* S2R+ cells (Yanagawa et al., 1998) at 1x 10^6 cells per ml density. Routinely, 12-15 μg of DNA was used with 300 μl of Effectene (Qiagen) following the manufacturer's protocol. Twenty-four hours after transfection, expression of the tagged protein was induced with 0.35 mM $CuSO_4$. This level of $CuSO_4$ is 50% of the concentration used in a standard protocol (0.7 mM) (Bunch et al., 1988) and has been tested to induce a low-to-medium level of recombinant protein expression for a majority of representative clones. Transfection efficiency and level of bait protein expression in each cell line was analyzed by immunofluorescence using anti-HA antibody (Roche Applied Science, Clone 3F10). Twenty four hours after the addition of $CuSO_4$ to each culture, cells were washed twice with Phosphate Buffered Saline buffer (PBS) at room temperature and whole-cells were triturated in 5 ml of Lysis Buffer (25 mM NaF, 1 mM Na_3VO_4, 50 mM Tris pH 7.5, 1.5 mM $MgCl_2$, 125 mM NaCl, 0.2% IGEPAL, 5% glycerol, Complete™ Tablets) with gentle rotation at 4°C for 30 minutes.

Co-affinity purification of bait proteins along with interacting partners

Prior to binding to affinity resin, lysates were quick-thawed, then clarified by passage through a 0.45 micron PVDF filter (Millipore, Inc.). The entire process of purification was carried out at 4°C to preserve the integrity of protein complexes and prevent potential degradation. Initially, cleared cell lysates were bound overnight to 75 µl of immunoaffinity resin [Clone HA-7 agarose, Sigma-Aldrich #A2095, cross-linked as per (Veraksa et al., 2005)]. The immunoaffinity resin (along with captured proteins) was transferred to AcroPrep 96 Filter Plates, 1 mL, 1.2 µm Supor membrane (PALL #5065). Unbound proteins were removed by washing thrice with Lysis Buffer followed by four additional washes with PBS to remove traces of detergent from lysis buffer washes. We found this detergent removal step to be essential for the subsequent mass spec analysis. The bound complexes were released from the resin by competition with the synthetic HA peptide YPYDVPDYA (250 µg/ml, Biosynthesis, Inc.) in PBS. Three successive elutions (200 µl, 30 mins each at room temperature) were carried out and the eluates combined for further processing.

Liquid Chromatography-Mass Spectrometry (LC-MS/MS)

Eluate samples were prepared for LC-MS/MS by precipitating with cold 20% trichloroacetic acid (TCA). The resultant protein precipitates were washed once with 10% TCA followed by four washes with ice-cold acetone to remove excess HA peptide present in the eluate, and then air-dried. Dried protein samples are subjected overnight to in-solution trypsin digestion (in 50 mM ammonium bicarbonate). Tryptic peptides were analyzed by LC-MS/MS. Peptides were separated across a 45-min gradient from 10% to 35% (v/v) acetonitrile in 0.1% (v/v) formic acid in a (125 µm × 18 cm) C_{18} microcapillary column (Magic C18AQ, 5 µm particles, 200 Å pore size, Michrom Bioresources) and analyzed on-line on an LTQ XL™ (Thermo Fisher). For each cycle, one full MS scan was followed by ten MS/MS spectra on the linear ion trap XL from the ten most abundant ions.

S2R+ Whole Cell Lysate Proteome Analysis

Sample preparation, LC-MS/MS Database Searching and Filtering were performed as described previously (Torres et al., 2010; Villen and Gygi, 2008) with minor changes. Isoelectric focusing (IEF) was done according to (Chick et al., 2008). Protein identifications from four experimental conditions i.e. two proteases, Trypsin and LysC, and two fractionation methods, Strong Cation

Exchange (SCX) and Isoelectric Focusing (IEF) were combined to generate a list of proteins identified in the S2R+ proteome.

Pulldown Data Analysis and Filtering

LC-MS/MS data from 4,273 co-AP/MS experiments was searched with SEQUEST (Eng et al., 2008) against a database of *D. melanogaster* proteins derived from FlyBase version 5.23. To assess the quality of the data, the searches were performed against a concatenated forward/reverse database as described previously (Elias and Gygi, 2007). Each resulting data set was filtered based on XCorr, deltaCorr, charge and peptide length to a target peptide False Discovery Rate (FDR) of 5% using Linear Discriminant Analysis (LDA) to distinguish between forward and reverse hits as described previously (Huttlin et al., 2010). Data sets were further filtered to reduce the protein-level FDR. After all peptides were grouped with their corresponding proteins, proteins were scored based on their summed peptide LDA probabilities. The sorted lists were filtered based on reversed protein hits to maximally contain only 1% protein false positives.

Criteria for inclusion in subsequent analysis primarily focused on the observation of bait peptides. The three order of magnitude variation in *Drosophila* protein length and peculiarities of tryptic cleavage site distribution for each protein clearly have great effect on the probability of observing peptides by LC/MS. To estimate these effects, we pre-calculated the number of theoretical 6-35 residue-long tryptic peptides for each bait protein. To compensate for length and tryptic cleavage effects, as well as for low transfection efficiency of S2R+ cells, we retained for analysis, all runs containing at least one bait peptide and a small number of runs containing no bait peptides, if the bait proteins for those runs were predicted to have fewer than 5 potentially observable tryptic peptides. The final data set included identifications from 3,488 affinity purifications.

Despite controlling false discovery rates for each run, combining such a large number of data sets always leads to higher FDR as the true positive hits are more likely to be the same from run to run than false positive hits. We therefore excluded from the combined data set any protein hits obtained from a single peptide identification in a single run. After all the filtering, the combined data set contained 2,770,552 total peptides at 0.007% FDR identified from 4,927 proteins at 0.8% FDR, This implies an average 2x10^{-6} % peptide FDR and 2x10^{-4} % protein FDR per pulldown experiment.

Correction for Column Carry-over Between LC/MS Experiments

Running thousands of LC/MS experiments back-to-back inevitably results in situations where peptides of a highly abundant protein (usually one used as bait) are not completely eluted off the column and show up in subsequent experiment(s) in lower numbers. Most algorithms for identification of specific interactions from affinity purification data, including the one described here, rely on distinguishing – based on observed frequency – those proteins that occur rarely with respect to some model or background frequencies. Thus when a protein that is rarely observed in the data set is detected as a carry-over contaminant, it is likely to result in high-scoring false positive interactions. Some previous studies have dealt with this issue manually, relying on human curation and/or via duplicate runs (Breitkreutz et al., 2010; Sowa et al., 2009). Here we employ a statistical approach to deal with this problem.

We start by selecting batches of experiment-protein pairs where a protein was detected with sequentially decreasing number of Total Spectrum Counts (TSCs). The first instance in each batch was required to have at least 20 TSCs because we reasoned, based on previous observations, that a smaller number is unlikely to lead to carry-over. For each batch of *n* consecutive observations we compute a probability composed of two components:

$$P(Carry-over) = P(\sum_{i=1}^{n} TSC_i) \times P(Run\ of\ Length\ n)$$

The first component estimates the probability of observing a batch of experiments with a certain sum of TSCs or greater. Let us treat the set of TSCs across all experiments as a random variable X {X$_i$…X$_n$} with mean *μ* and standard deviation *σ*. According to the central limit theorem, the distribution of sample sums drawn from X tends to be approximately normal, with distribution parameters easy to estimate from data. We can thus compute the probability of observing a given sum of TSCs from a batch as an argument of the error function:

$$\sum X \sim N\left(n \times \mu, \sqrt{n} \times \sigma\right)$$

$$P(\sum_{i=1}^{n} TSC_i >= a) = \frac{2}{\sqrt{\pi}} \int_a^{\infty} e^{-x^2} dx = erfc(x)$$

The second component estimates the probability of observing the putative contaminant in a consecutive run of experiments of length *n*, given the frequency of its identification in the overall data set. Probability of a run of at least *r* consecutive successes in *k* independent trials where the probability of success at any one trial is *p* was computed following Villarino (Villarino, 2005):

$$z_k = \beta_{k,r} - p^r \beta_{k-r,r}$$

$$\beta_{k,r} = \sum_{l=0}^{\left\lfloor \frac{k}{r+1} \right\rfloor} (-1)^l \binom{k-lr}{l} \left(qp^r\right)^l, \text{ where } q = (1-p)$$

In the limiting case where the protein is commonly observed and a run of n sequential observations is likely, the probability of contamination is estimated from abundance as measured by TSCs. In the other limiting case where the protein observation overall is rare and the sum of TSCs in all experiments is equal to or approaches that observed in the batch, the contamination probability is given by likelihood of observing such rare proteins consecutively in a batch. The joint probability calculated above was further corrected for multiple hypothesis testing to ensure < 0.01 FDR thus defining a set of protein observations deemed to be the result of carry-over contamination. Contaminants were then taken into account and subtracted from the overall data set prior to final data analysis.

Protein complex map generation and HGSCore method

After LC/MS carry-over correction, data from the selected 3,488 affinity purification experiments were used to create a protein interaction network by scoring each protein's probability of interaction based on a hypergeometric distribution error model (Hart et al., 2007) but modified to incorporate TSC information. For each affinity purification, we derived Normalized Spectral Abundance Factors (NSAF) (Zybailov et al., 2006), thus correcting for differences in protein length. NSAF values were transformed, setting the smallest value to 1, calculating the square root of each and keeping the integer value. The square root step compressed the range of the values and made the subsequent probability calculation feasible for the whole range. The resulting transformed, normalized TSCs, referred to as T_N, were used to convert the data into the matrix model representation (Bader and Hogue, 2002). Traditionally, the matrix model represents counts of affinity purification experiments where proteins i and j are identified. Instead of binary co-occurrence we wanted to capture the quantitative aspect of the data represented by T_N. We reasoned that for each pair of proteins, the determination of specificity of interaction between them hinges on the smaller of the two T_N values. Thus, replacing the occurrence observations with sum of min(T_N) values, we can calculate the hypergeometric probability of observing an interaction between i and j given the background spectral abundance:

$$P(\sum \min(T_N) > k \mid n, m, N) = \sum_{x=k}^{\min(n,m)} P_{hygeo}(x \mid n, m, N)$$

$$P_{hygeo}(x \mid n, m, N) = \frac{\binom{n}{x}\binom{N-n}{m-x}}{\binom{N}{m}}$$

where

$$k = \sum \min(T_N) \text{ for experiments with } \mathrm{T}_{\mathrm{N;\,i}} > 0 \text{ and } \mathrm{T}_{\mathrm{N;\,j}} > 0$$

$$n = \sum \min(T_N) \text{ for experiments with } \mathrm{T}_{\mathrm{N;\,i}} > 0$$

$$m = \sum \min(T_N) \text{ for experiments with } \mathrm{T}_{\mathrm{N;\,j}} > 0$$

$$N = \sum \min(T_N) \text{ for all experiments}$$

The final scores among all pairs were calculated as follows:

$$HGSCore_{i,j} = -\log(P_{hygeo;i,j})$$

Without a sufficiently large positive reference set available for fly protein interactions, we relied on simulations to set a cutoff for DPiM and estimate the False Discovery Rate. Similar to the procedure described earlier (Sowa et al., 2009), for each simulated affinity purification data set, we randomly sampled proteins from a distribution of total combined spectral counts until the number of unique proteins in the simulated data set equaled that in the original. Sampling then continued, but drawing only from the TSC distribution for the proteins already selected until the total spectral counts for the original experiment was reached. The resulting simulated data set thus closely resembles the original statistically, while allowing for randomized assortment of proteins among experiments. To derive a reliable estimate we created at least 10 simulated data sets for every version of the original, which was sufficient to reach convergence. Based on the simulated data sets, we defined an HGSCore cutoff of 61.537 for 5% FDR in the final DPiM data set. See Supplemental Table S3.

Prediction of protein complexes and enrichment analysis

The network of protein interactions filtered to 5% FDR was clustered using MCL (Enright et al., 2002) (inflation coefficient of 3.0 and pre-inflation factor of 0.01). For each cluster, we searched for statistically overrepresented Gene Ontology (GO) Terms (Ashburner et al., 2000), KEGG pathways, and Pfam domains, using DAVID (Huang da et al., 2009). Resulting p-values were further filtered to adjust for multiple hypothesis testing, keeping clusters based on a 0.01 Q-value cutoff (Storey and Tibshirani, 2003). Calculations of annotated genes were made based on GO annotations corresponding to FlyBase release 5.23. We considered all protein-coding genes with GO Evidence Codes other than ND and NR. Annotations with IEA Evidence Codes were considered separately where indicated. See Supplemental Table S4.

Overlap of HGSCore and other methods with DroID data sets

Seven of the eight DroID 5.1 data sets (i.e. excluding "genetic interactions") were used for this analysis. Only those gene pairs with both genes identified in our affinity purifications were considered. We counted the number of sources supporting each protein-protein interaction subset classified by DroID and defined four bins of pairwise interactions with increasing levels of confidence *i.e.,* supported by at most one, two, three or four independent sources and computed the overlaps with DPiM. In addition to the HGSCore method, the coAP-MS data set from 3,488 affinity purifications was also analyzed using other published scoring methods (Breitkreutz et al., 2010; Gavin et al., 2006; Hart et al., 2007; Sowa et al., 2009). We used the publicly available SAINT algorithm implementation (command line options were set to: 1000 10000 0.1 0 1 1) used by Breitkreutz et al. (Choi et al., 2011). Other algorithms were implemented exactly as described in respective publications. As the total number of potential protein interactions defined by each method was different, we compared the top 25,000 interactions reported from each of the methods with interactions listed in DroID. Significance of improvement in recall of DroID interactions was calculated using Chi-square. The same analysis was carried out on DPiM at 5% FDR to estimate sensitivity of our method.

Gene expression analysis

From the modEncode mRNA expression time course study across 30 points of male and female fly development (Graveley et al., 2011), we calculated all-against-all Pearson correlation coefficients for the 2,297 genes that are represented in the map. The entire distribution of coefficients was then compared to the coefficients corresponding to the subset of interacting pairs identified in DPiM.

Using the RPKM values from the RNA-Seq study of *Drosophila* S2R+ cells (Cherbas et al., 2011) as a proxy for gene expression levels, we calculated the normalized differences between absolute expression following Jansen et al. (Jansen et al., 2002).

Analogous calculations were also made using microarray gene expression data across 26 *Drosophila* tissues and the S2 cell line (Chintapalli et al., 2007).

Cross-species Validation of DPiM Interactions

Orthology mapping was done using InParanoid7 best reciprocal hits (Ostlund et al., 2010). Purifications of human baits were performed as described previously (Behrends et al., 2010) with minor modifications. ORF clones were obtained from the CCSB human ORFeome collection (Lamesch et al., 2007; Rual et al., 2004) and subcloned into the pHAGE-N-FLAG-HA vector by Gateway cloning. Proteins were expressed in 293F cells (Gibco #11625-019) after transient transfection of 4µg of DNA into 20ml of cells using Effectene (Qiagen). Cells were lysed in a total volume of 2ml of lysis buffer (50mM Tris-HCl pH7.5, 150 mM NaCl, 0.2% Nonidet P40, Roche Complete, EDTA-free protease inhibitor cocktail) and processed using our regular coAP-MS procedure. Mass spectrometry data were processed as described previously (Behrends et al., 2010).

Evolution of complexes

To look at conservation of complexes across evolution, we compared those predicted from DPiM with data available for yeast and humans. To ensure the best possible definition of complex membership, we used two independent data sets for each organism. For yeast, we relied on the consolidated experimental data sets from two large-scale studies (Gavin et al., 2006; Krogan et al., 2006), re-analyzed by Pu et al (Pu et al., 2007). In addition, we used the gold-standard manually curated data set known as CYC2008 (Pu et al., 2009) – an update to the MIPS data set (Mewes et al., 2004). For human data, we relied on two independent manually curated data sets – the REACTOME (Croft et al., 2011; Matthews et al., 2009) and CORUM (Ruepp et al., 2008) databases. The complexes were aligned and displayed in Cytoscape (Cline et al., 2007) with the help of DualLayout plugin (Sharan et al., 2005) as well as other tools made available by the Resource for Biocomputing, Visualization, and Informatics (RBVI) Center at UCSF.

IV. Supplemental References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bader, G.D., and Hogue, C.W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol *20*, 991-997.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68-76.

Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G.*, et al.* (2010). A global protein kinase and phosphatase interaction network in yeast. Science *328*, 1043-1046.

Bunch, T.A., Grinblat, Y., and Goldstein, L.S. (1988). Characterization and use of the Drosophila metallothionein promoter in cultured Drosophila melanogaster cells. Nucleic Acids Res *16*, 1043-1061.

Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B.D., Carlson, J.W., Landolin, J.M., Kapranov, P., Dumais, J.*, et al.* (2011). The transcriptional diversity of 25 Drosophila cell lines. Genome Res *21*, 301-314.

Chick, J.M., Haynes, P.A., Molloy, M.P., Bjellqvist, B., Baker, M.S., and Len, A.C. (2008). Characterization of the rat liver membrane proteome using peptide immobilized pH gradient isoelectric focusing. J Proteome Res *7*, 1036-1045.

Chintapalli, V.R., Wang, J., and Dow, J.A. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet *39*, 715-720.

Choi, H., Larsen, B., Lin, Z.Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S., Tyers, M., Gingras, A.C., and Nesvizhskii, A.I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat Methods *8*, 70-73.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B.*, et al.* (2007). Integration of biological networks and gene expression data using Cytoscape. Nat Protoc *2*, 2366-2382.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B.*, et al.* (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res *39*, D691-697.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods *4*, 207-214.

Eng, J.K., Fischer, B., Grossmann, J., and Maccoss, M.J. (2008). A fast SEQUEST cross correlation algorithm. J Proteome Res *7*, 4598-4602.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575-1584.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W.*, et al.* (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473-479.

Hart, G.T., Lee, I., and Marcotte, E.R. (2007). A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics *8*, 236.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villen, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. Cell *143*, 1174-1189.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. Genome Res *12*, 37-46.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G., Martin, P.*, et al.* (2007). hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics *89*, 307-315.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B.*, et al.* (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res *37*, D619-622.

Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V.*, et al.* (2004). MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res *32*, D41-44.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res *38*, D196-203.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S.J. (2007). Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics *7*, 944-960.

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S.J. (2009). Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res *37*, 825-831.

Rual, J.F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O.*, et al.* (2004). Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res *14*, 2128-2135.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stumpflen, V.*, et al.* (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res *36*, D646-650.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. (2005). Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A *102*, 1974-1979.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389-403.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.

Torres, E.M., Dephoure, N., Panneerselvam, A., Tucker, C.M., Whittaker, C.A., Gygi, S.P., Dunham, M.J., and Amon, A. (2010). Identification of aneuploidy-tolerating mutations. Cell *143*, 71-83.

Veraksa, A., Bauer, A., and Artavanis-Tsakonas, S. (2005). Analyzing protein complexes in Drosophila with tandem affinity purification-mass spectrometry. Dev Dyn *232*, 827-834.

Villarino, M.B. (2005). The Probability of a Run. arXiv *math/0511652*.

Villen, J., and Gygi, S.P. (2008). The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat Protoc *3*, 1630-1638.

Yanagawa, S., Lee, J.S., and Ishimoto, A. (1998). Identification and characterization of a novel line of Drosophila Schneider S2 cells that respond to wingless signaling. J Biol Chem *273*, 32353-32359.

Yu, C., Wan, K.H., Hammonds, A.S., Stapleton, M., Carlson, J.W., and Celniker, S.E. (2011). Development of expression-ready constructs for generation of proteomic libraries. Methods Mol Biol *723*, 257-272.

Zybailov, B., Mosley, A.L., Sardiu, M.E., Coleman, M.K., Florens, L., and Washburn, M.P. (2006). Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res *5*, 2339-2347.

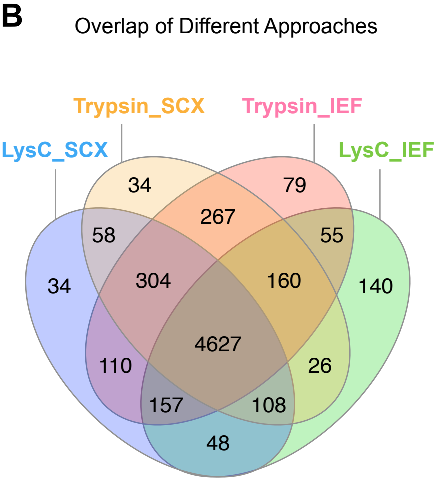
Supplemental Figure S1

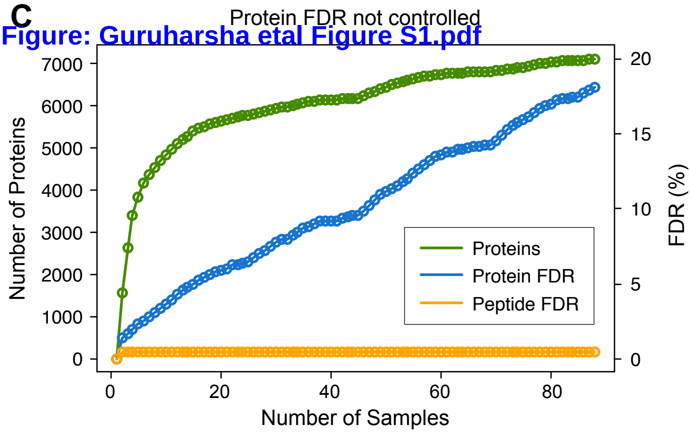
A

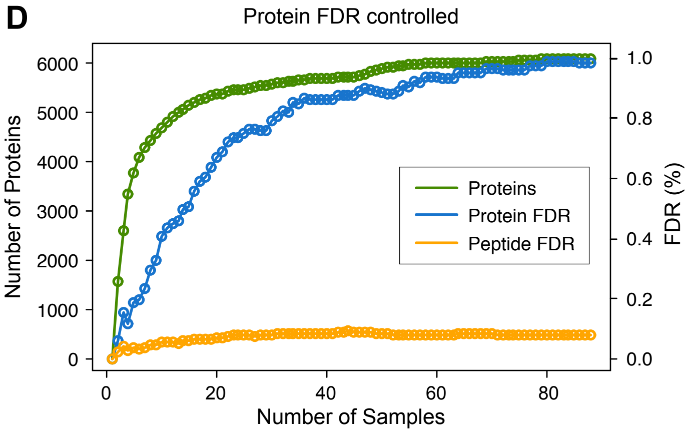
S2R+ Whole-cell Proteome Analysis

	Trypsin	LysC
SCX	5,584 (0.66%)	5,446 (0.73%)
IEF	5,759 (0.68%)	5,195 (0.65%)

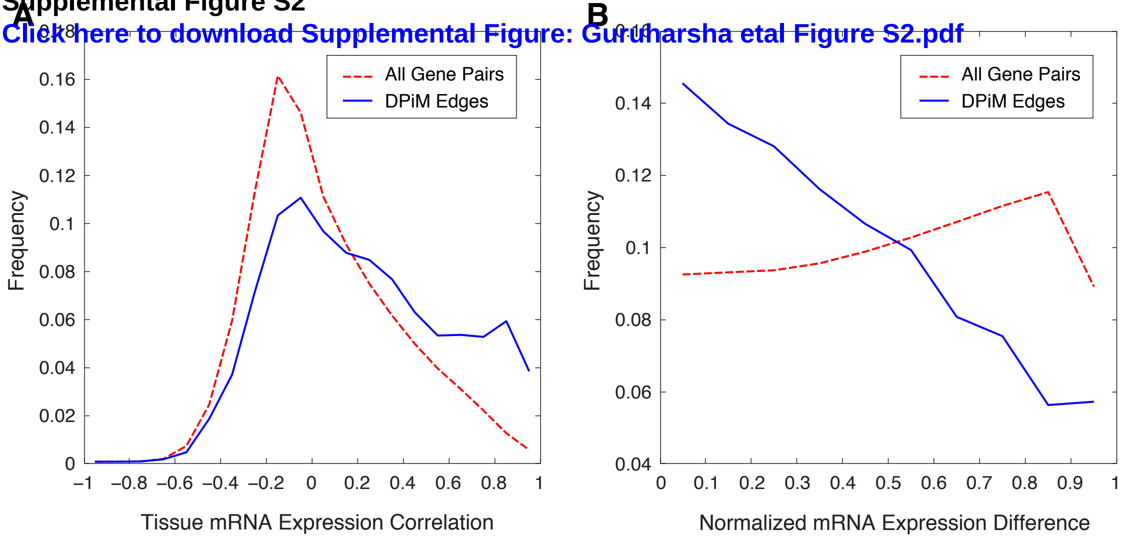
Total Proteins: 6,081 (0.99% FDR)
Total Genes: 5,695



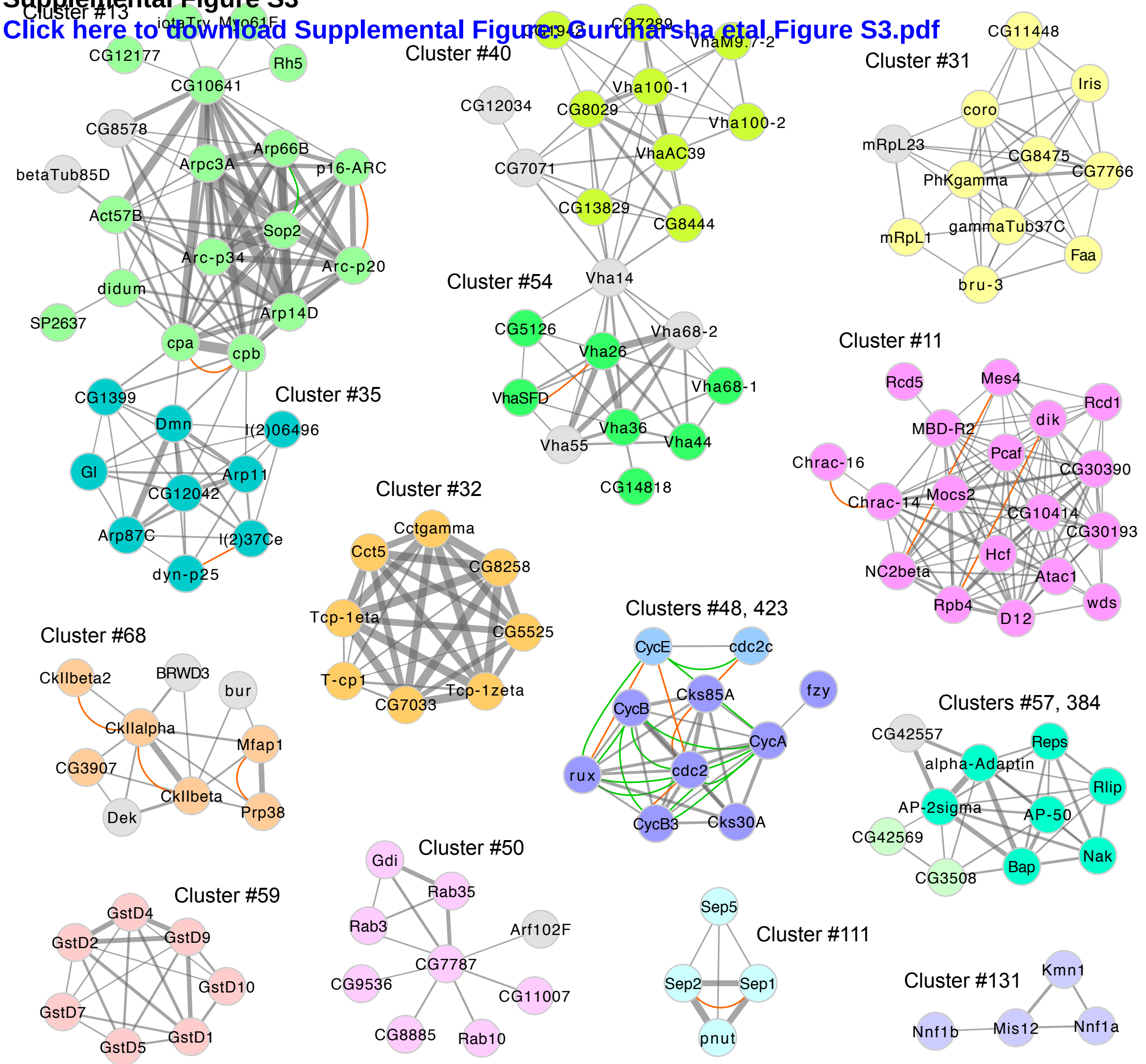




Supplemental Figure S2



Supplemental Figure S3

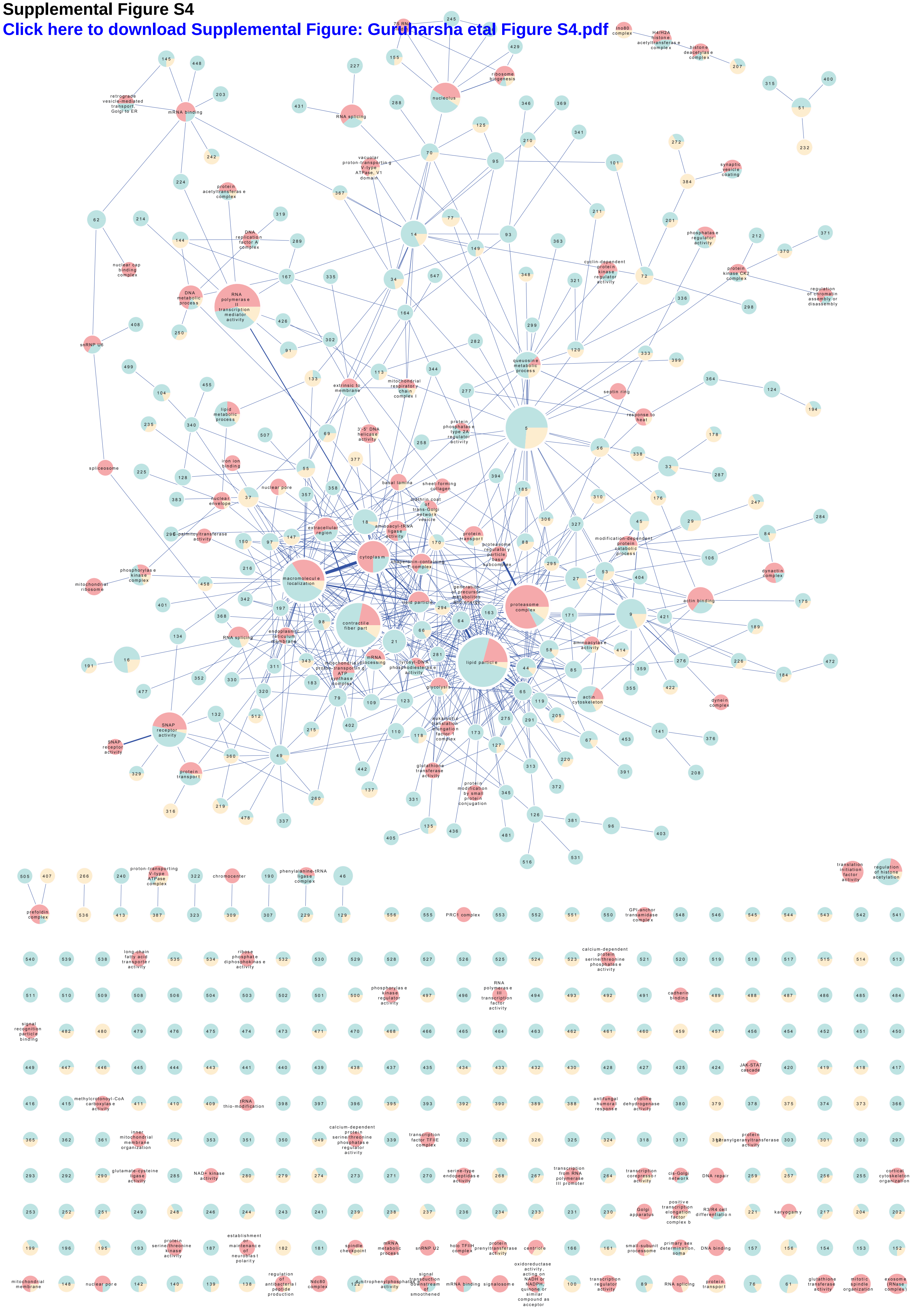


— DPiM HGSCore
- - - DPiM sub-threshold
— Genetic interactions
— Physical interactions

Cluster #11	regulation of histone acetylation
Cluster #13	regulation of actin polymerization/depolymerization
Cluster #31	phosphorylase kinase complex
Cluster #32	chaperonin-containing T-complex
Cluster #35	dynactin complex
Cluster #40	vacuolar proton-transporting V-type ATPase, V0 domain
Cluster #48	cyclin-dependent protein kinase regulator activity
Cluster #50	protein transport
Cluster #54	vacuolar proton-transporting V-type ATPase, V1 domain
Cluster #57	vesicle coating
Cluster #59	glutathione transferase activity
Cluster #68	protein kinase CK2 complex
Cluster #111	septin ring
Cluster #131	Ndc80 complex
Clusrer #384	no significant terms
Cluster #423	JAK-STAT cascade

Supplemental Figure S4

Click here to download Supplemental Figure: Guruharsha etal Figure S4.pdf



7S RNA
245
Ino80 complex
H4/H2A histone acetyltransferase complex
histone deacetylase complex
207
145
448
155
429
227
ribosome biogenesis
315
400
retrograde vesicle-mediated transport, Golgi to ER
203
mRNA binding
431
RNA splicing
288
nucleolus
346
369
51
242
125
341
210
232
70
272
vacuolar proton-transporting V-type ATPase, V1 domain
95
101
224
protein acetyltransferase complex
367
384
synaptic vesicle coating
77
211
62
214
319
201
DNA replication factor A complex
14
93
363
phosphatase regulator activity
212
371
144
289
149
370
nuclear cap binding complex
167
335
34
547
348
cyclin-dependent protein kinase regulator activity
72
protein kinase CK2 complex
regulation of chromatin assembly or disassembly
DNA metabolic process
RNA polymerase II transcription mediator activity
321
336
298
164
299
408
426
250
302
282
queuosine metabolic process
120
333
399
snRNP U6
91
499
113
344
364
133
extrinsic to membrane
mitochondrial respiratory chain complex I
277
septin ring
124
455
104
194
lipid metabolic process
response to heat
235
340
3'-5' DNA helicase activity
protein phosphatase type 2A regulator activity
5
69
178
507
258
56
338
377
iron ion binding
33
287
225
128
55
spliceosome
basal lamina
394
sheet-forming collagen
185
nuclear pore
358
357
176
383
nuclear envelope
37
310
247
clathrin coat of trans-Golgi network vesicle
306
284
327
29
45
18
extracellular region
aminoacyl-tRNA ligase activity
protein transport
84
296
S-palmitoyltransferase activity
modification-dependent protein catabolic process
150
97
147
170
88
proteasome regulatory particle, base subcomplex
106
cytoplasm
dynactin complex
chaperonin-containing T-complex
295
phosphorylase kinase complex
216
53
27
404
mitochondrial ribosome
458
macromolecule localization
generation of precursor metabolites and energy
actin binding
175
401
342
lipid particle
proteasome complex
197
294
163
171
9
421
368
64
98
contractile fiber part
66
189
134
RNA splicing
endoplasmic reticulum membrane
aminoacylase activity
414
21
58
16
281
lipid particle
226
472
191
mRNA processing
tyrosyl-DNA phosphodiesterase activity
276
343
44
mitochondrial proton-transporting ATP synthase complex
85
359
184
311
352
183
glycolysis
330
355
422
477
65
actin cytoskeleton
119
dynein complex
320
79
109
123
132
512
205
275
291
SNAP receptor activity
215
402
110
118
eukaryotic translation elongation factor 1 complex
173
453
141
376
SNAP receptor activity
67
360
49
127
220
442
391
208
329
protein transport
glutathione transferase activity
313
372
137
protein modification by small protein conjugation
316
219
260
331
345
478
337
126
381
135
96
436
403
405
481
516
531
505
407
266
240
proton-transporting V-type ATPase complex
322
chromocenter
190
phenylalanine-tRNA ligase complex
46
translation initiation factor activity
regulation of histone acetylation
prefoldin complex
536
413
387
323
309
307
229
129
556
555
PRC1 complex
553
552
551
550
GPI-anchor transamidase complex
548
546
545
544
543
542
541
540
539
538
long-chain fatty acid transporter activity
535
534
ribose phosphate diphosphokinase activity
532
530
529
528
527
526
525
524
523
calcium-dependent protein serine/threonine phosphatase activity
521
520
519
518
517
515
514
513
511
510
509
508
506
504
503
502
501
500
phosphorylase kinase regulator activity
497
496
RNA polymerase III transcription factor activity
494
493
492
491
cadherin binding
489
488
487
486
485
484
signal recognition particle binding
482
480
479
476
475
474
473
471
470
468
466
465
464
463
462
461
460
459
457
456
454
452
451
450
449
447
446
445
444
443
441
440
439
438
437
435
434
433
432
430
428
427
425
424
JAK-STAT cascade
420
419
418
417
416
415
methylcrotonoyl-CoA carboxylase activity
411
410
409
tRNA thio-modification
398
397
396
395
393
392
390
389
388
antifungal humoral response
choline dehydrogenase activity
380
379
378
375
374
373
366
365
362
361
inner mitochondrial membrane organization
354
353
351
350
349
calcium-dependent protein serine/threonine phosphatase regulator activity
339
transcription factor TFIIE complex
332
328
326
325
324
318
317
312
geranylgeranyltransferase
protein prenyltransferase activity
303
301
300
297
293
292
290
glutamate-cysteine ligase activity
285
NAD+ kinase activity
280
279
274
273
271
270
serine-type endopeptidase activity
268
267
transcription from RNA polymerase III promoter
264
transcription corepressor activity
cis-Golgi network
DNA repair
259
257
256
255
cortical cytoskeleton organization
253
252
251
249
248
246
244
243
241
239
238
237
236
234
233
231
230
Golgi apparatus
positive transcription elongation factor complex b
R3/R4 cell differentiation
221
karyogamy
217
204
202
199
196
195
193
protein serine/threonine kinase activity
187
establishment or maintenance of neuroblast polarity
182
181
spindle checkpoint
mRNA metabolic process
snRNP U2
holo TFIIH complex
protein prenyltransferase activity
centriole
166
161
small-subunit processome
primary sex determination, soma
DNA binding
157
156
154
153
152
mitochondrial membrane
148
nuclear pore
142
140
139
138
regulation of antibacterial peptide production
Ndc80 complex
122
4-nitrophenylphosphatase activity
signal transduction downstream of smoothened
mRNA binding
signalosome
oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor
100
transcription regulator activity
89
RNA splicing
protein transport
76
61
glutathione transferase activity
mitotic spindle organization
exosome (RNase complex)