Science Signaling
AAAS

` Protein Complex-Based Analysis Framework for High-Throughput Data Sets

Arunachalam Vinayagam, Yanhui Hu, Meghana Kulkarni, Charles Roesel, Richelle Sopko, Stephanie E. Mohr and Norbert Perrimon (26 February 2013) *Science Signaling* **6** (264), rs5. [DOI: 10.1126/scisignal.2003629]

Science Signaling (ISSN 1937-9145) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005.

SYSTEMS BIOLOGY

Protein Complex–Based Analysis Framework for High-Throughput Data Sets

Arunachalam Vinayagam,[1]* Yanhui Hu,[1,2]† Meghana Kulkarni,[1]†‡ Charles Roesel,[2,3] Richelle Sopko,[1] Stephanie E. Mohr,[1,2] Norbert Perrimon[1,2,4]*

Analysis of high-throughput data increasingly relies on pathway annotation and functional information derived from Gene Ontology. This approach has limitations, in particular for the analysis of network dynamics over time or under different experimental conditions, in which modules within a network rather than complete pathways might respond and change. We report an analysis framework based on protein complexes, which are at the core of network reorganization. We generated a protein complex resource for human, *Drosophila*, and yeast from the literature and databases of protein-protein interaction networks, with each species having thousands of complexes. We developed COMPLEAT (http://www.flyrnai.org/compleat), a tool for data mining and visualization for complex-based analysis of high-throughput data sets, as well as analysis and integration of heterogeneous proteomics and gene expression data sets. With COMPLEAT, we identified dynamically regulated protein complexes among genome-wide RNA interference data sets that used the abundance of phosphorylated extracellular signal–regulated kinase in cells stimulated with either insulin or epidermal growth factor as the output. The analysis predicted that the Brahma complex participated in the insulin response.

INTRODUCTION

The analysis of data sets from genome-scale screens typically involves raw data processing, such as calculating *z* scores and fold changes, where genes are given a score and identified as "hits." Because these screens output hundreds of genes, it is standard practice to identify the enrichment for a group of genes that are part of particular functional categories or pathways (*1*). The advantage of such analysis is that it is less prone to the inherent false positives and false negatives associated with the data. For example, in RNA interference (RNAi) screens, a gene might be considered a false negative due to ineffective knockdown or as a false positive due to off-target effects; however, it is less likely that an entire group of genes could be falsely classified. Further, analyses based on gene enrichments improve confidence in the results by placing them in biological context and helps generate new hypotheses. About 70 different enrichment analysis tools have already been developed, most of which use Gene Ontology (GO) (*2*) or pathway databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (*3*) to group functionally related genes (*1*).

Although GO and pathway annotations are useful, they can be either too specific or too broad in the context of network dynamics. For example, annotations from the KEGG MAPK (mitogen-activated protein kinase) pathway spans from the membrane receptor complexes that receive a signal to the nuclear transcription factor complexes that constitute the signal readout. It is difficult to identify changes in response to stimuli over time because these changes are likely to affect only a subset of pathway components. In contrast to pathways, protein complexes are the functional units of proteome organization, and their dynamic assembly is fundamental to induce cellular responses to different internal and external cues (*4*). Thus, for data sets that include multiple conditions or time points, a protein complex–based analysis might be preferable because it could reveal network dynamics that are missed in other types of analyses. Moreover, the individual protein complexes that participate in a signaling pathway assemble in different compartments and at different times, and some, but not all, complexes associated with a pathway might integrate signals from other pathways. Thus, to understand how cells reorganize at a systems level, we must be able to visualize and study the dynamics of protein complexes.

Recently, genome- or proteome-scale data sets have been generated under different conditions and time points with an objective of capturing the dynamics of the biological system (*5*–*10*). To efficiently analyze the network dynamics of these data sets, there is a need for analysis tools for data related to protein complexes. Even the most commonly used enrichment analysis tools, including the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (*11*) and gene set enrichment analysis (GSEA) (*12*), do not support complex-based analysis, mainly due to the lack of availability of comprehensive protein complex resources. For example, the existing protein complex databases either focus on a specific organelle or cover only a few protein complexes for a single species (*13*–*15*). Further, the current analysis tools do not support direct comparison and visualization of dynamic data sets. Hence, there is a need for both a comprehensive complex-based resource and a tool that uses the resource to analyze dynamic high-throughput data sets. Moreover, such a complex-based analysis is not restricted to dynamic data sets but could also be used for the analysis of single data sets.

To fill this gap, we developed a framework for the analysis of high-throughput data sets at the level of protein complexes (Fig. 1). Because the currently available databases of complex information underrepresent the full picture, we first generated comprehensive protein complex resources for *Homo sapiens* (human), *Drosophila melanogaster* (fly), and *Saccharomyces cerevisiae* (yeast). Using the protein complex resources as back-end annotations,

[1]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA. [2]Drosophila RNAi Screening Center, Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [3]Bioinformatics Program, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA. [4]Howard Hughes Medical Institute, Boston, MA 02115, USA.
*To whom correspondence should be addressed. E-mail: vinu@genetics.med.harvard.edu (A.V.); perrimon@receptor.med.harvard.edu (N.P.)
†These authors contributed equally as second authors.
‡Present address: Belfer Institute, Dana-Farber Cancer Institute, Boston, MA 02115, USA.

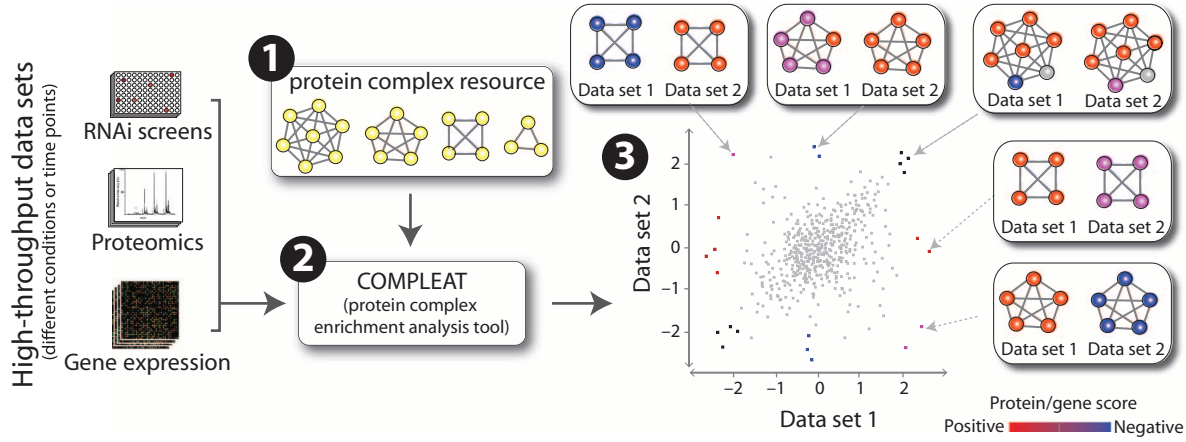

Fig. 1. Schematic representation of the protein complex–based analysis framework. The framework handles a variety of high-throughput data sets, including RNAi screen, proteomics, and expression data sets. The three major components of the framework are (1) the protein complex resource, (2) the complex enrichment analysis tool, and (3) data visualization, including the visualizations that facilitate comparison of multiple data sets.

we developed the protein Complex Enrichment Analysis Tool (COMPLEAT) and created a Web interface (http://www.flyrnai.org/compleat) that is freely available to the research community. We applied COMPLEAT to the analysis of genome-wide RNAi data sets that measured extracellular signal–regulated kinase (ERK) activity as represented by the presence of phosphorylated ERK (pERK) in cells stimulated with either insulin or epidermal growth factor (EGF) (*5*, *6*). Using pERK abundance as a measure, we identified insulin stimulus–dependent regulation of the Brahma protein complex and demonstrated experimentally that it is essential for mediating the insulin response.

RESULTS

Generation of comprehensive protein complex resources

We generated comprehensive protein complex resources for humans, *Drosophila*, and yeast by combining two different approaches: (i) We performed systematic identification of protein complex data reported in the literature, and (ii) we predicted protein complexes on the basis of protein-protein interaction (PPI) networks (Fig. 2A). To compile protein complex data for humans from the literature, we used Comprehensive Resource of Mammalian protein complexes (CORUM) (*15*), Proteins Interacting in the Nucleus database (PINdb) (*13*), protein complexes annotated by GO, and pathway modules and structural complexes from KEGG modules (*3*). For yeast, we used a manually curated catalog of protein complexes (denoted as CYC2008) (*14*), PINdb, and GO complexes. For *Drosophila*, we included complexes from GO and 556 protein complexes identified in an affinity purification mass spectrometry (AP-MS) pull-down study (*16*). We also mapped complexes in the human, *Drosophila*, and yeast data sets using the DRSC Integrative Ortholog Prediction Tool (DIOPT), an ortholog mapping tool (*17*) (table S1). In total, we compiled 3638, 3077, and 2173 literature-based protein complexes for humans, *Drosophila*, and yeast, respectively (Table 1). This collection includes both transient signaling complexes and stable complexes, such as proteasomes and ribosomes. Although the KEGG metabolic pathway modules are not necessarily physical complexes, they are included in our analysis because the metabolic pathways are underrepresented in these resources.

To predict protein complexes, we compiled experimentally identified PPIs for humans, *Drosophila*, and yeast by integrating PPI networks from major PPI databases, organism-specific databases, and high-throughput data sets (table S2). These integrated networks consist of 108,059 PPIs among 14,495 human proteins, 98,500 PPIs among 9373 *Drosophila* proteins, and 118,603 PPIs among 5729 yeast proteins (table S2). Next, we applied two different complex prediction tools, CFinder (*18*) and NetworkBLAST (*19*), that identify biologically meaningful protein complexes from PPI networks (*20*). CFinder is a clique percolation method to identify protein complexes from a single PPI network. NetworkBLAST is a network alignment tool for identifying conserved protein complexes. In the case of the CFinder analysis, we further filtered the PPI networks using coexpression values (for humans and *Drosophila*) or colocalization information (for yeast) to remove low-confidence PPIs. We did not apply the same filters for the NetworkBLAST analysis because false-positive interactions are unlikely to be reproduced across species (*21*). Together, we identified 6251 human complexes, 3639 *Drosophila* complexes, and 5551 yeast complexes (Table 1 and table S3). Finally, we integrated both literature-based and predicted complexes to create a comprehensive protein complex resource, resulting in 9881, 6703, and 7713 complexes for human, *Drosophila*, and yeast, respectively (Table 1).

Almost 50% of the literature-based and predicted protein complexes are redundant, comprising complexes that either are subsets of other complexes or differ from other complexes by only a few components (Fig. 2B and table S4). Because protein complexes are not rigid or fixed structures, we intentionally preserved such redundancies. Comparison of literature-based and predicted complexes reveals a low overlap, suggesting that distinct complexes are captured using these different approaches (Fig. 2C). For example, in *Drosophila*, there is a 15% overlap at the complex level and a 46% overlap at the protein level, suggesting that the computational-based predictions expand the resource with a large number of new proteins (table S5). We also observed that few proteins are part of many complexes, meaning that most of the proteins belong to one or very few complexes (Fig. 2D), an observation consistent with the scale-free behavior of PPI networks (*22*). Our complex resources include 68% of yeast proteins and ~50% of human and *Drosophila* proteins (Fig. 2E and table S6). Additionally, our resources include 75 to 90% of highly conserved proteins because we took advantage of evolutionary conservation to increase the coverage for individual species (Fig. 2E).

We analyzed various features of the complexes in the compiled resources. As expected, the size distribution of the protein complexes shows

that they are smaller than the size of KEGG pathway annotations. For example, the median size of the human complexes is 8, compared with 52 for KEGG (Fig. 2F). We also analyzed co-citation of the members of a complex in the literature by mapping the genes to PubMed citations to assess the biological relevance of protein complexes. To assess their significance, we compared them with 1000 random sets containing the same number of proteins as those in the complexes. For our human complexes, 96% showed significant co-citation, a proportion that was comparable to that found for

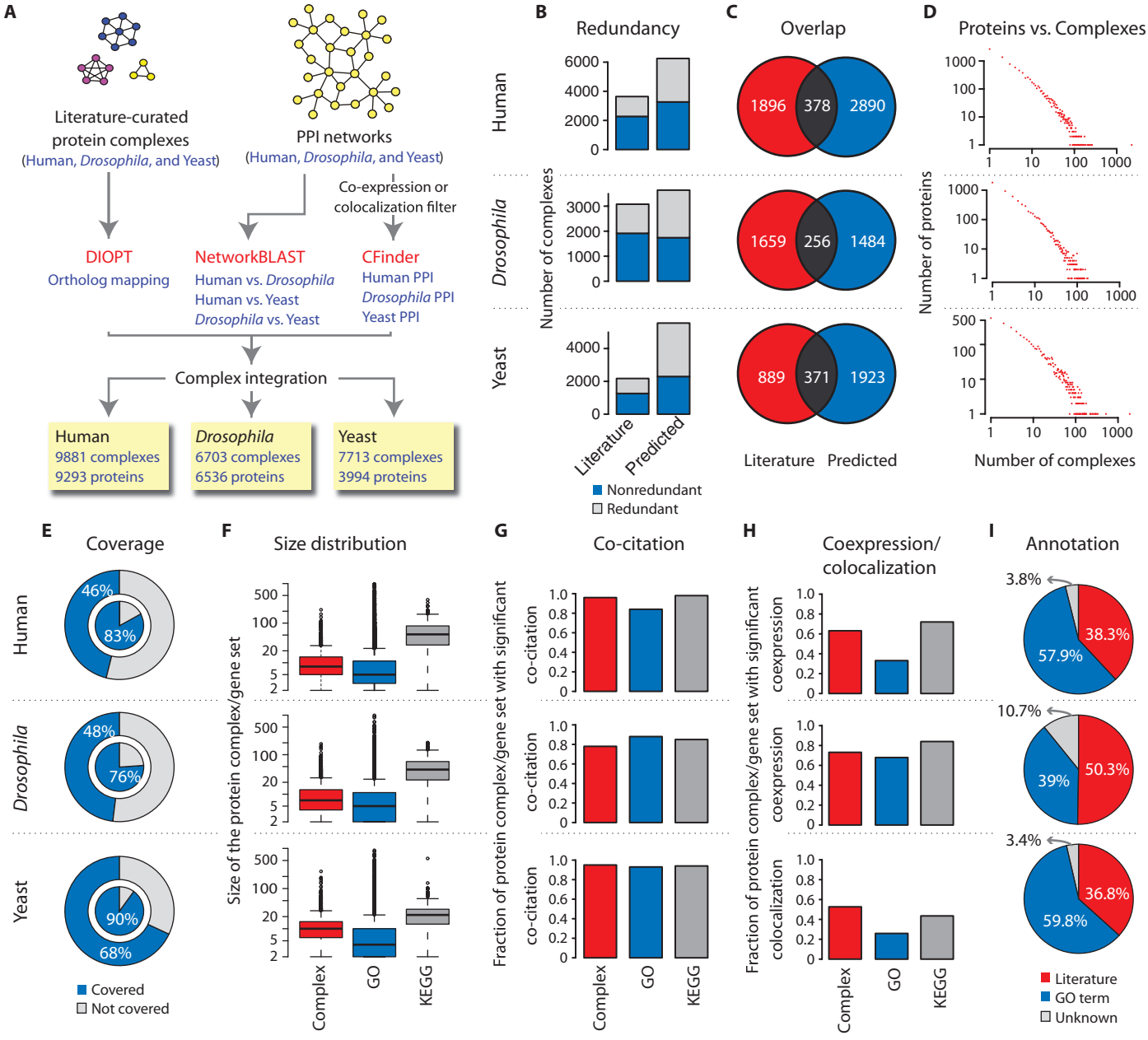

Fig. 2. Overview of the protein complex resources generated for humans, *Drosophila*, and yeast. (**A**) Schematic representation of the protein complex resource compilation. (**B**) Redundancies within the complex resource; blue bar corresponds to nonredundant complexes at 80% cutoff, and gray bar corresponds to redundant complexes. (**C**) Overlap between literature-based and predicted protein complexes. The overlap was computed at the nonredundant complex level (80% cutoff). (**D**) Distribution of the proteins against the number of complexes the protein belongs to. (**E**) The outer pie chart represents the percentage of the entire proteome covered by the complex resource, and the inner pie chart corresponds to highly conserved proteins. (**F**) Box plot showing the size distribution of the complex resource compared to the GO and KEGG pathways. (**G**) Bar graph of the significant fraction of the complexes and GO and KEGG annotations that are co-cited in the literature, compared to 1000 random sets. (**H**) Bar graph of the significant fraction of the complexes and GO and KEGG annotations that are coexpressed (human and *Drosophila*) or colocalize (yeast), compared to 1000 random sets. (**I**) Pie chart showing the contribution of different sources of annotations to the complex resources.

Table 1. Summary of literature-based, predicted, and combined protein complexes for human, *Drosophila*, and yeast.

Organism	Literature		Predicted		Combined	
	Complexes	Proteins	Complexes	Proteins	Complexes	Proteins
Human	3638	7524	6251	6334	9881	9293
Drosophila	3077	5619	3639	3933	6703	6536
Yeast	2173	3280	5551	3366	7713	3994

KEGG (98%) and better than that found for GO (85%) (Fig. 2G and table S7). For *Drosophila*, the proportion of complex member co-citation was lower than it is for GO and KEGG, and in yeast, it was comparable to GO and KEGG (Fig. 2G).

We also analyzed the evidence of colocalization for yeast and coexpression for components of the human or *Drosophila* complexes. For this analysis, we removed complexes predicted by CFinder because CFinder uses networks enriched for coexpression or colocalized PPIs for complex prediction (Fig. 2A). For yeast, we benefited from a large-scale effort to determine the subcellular localization of proteins and analyzed evidence for colocalization of the members of specific complexes (*23*). Fifty-three percent of the complexes showed significant colocalization of their constituents, which was twofold higher than that in GO (26%) and 1.25-fold higher than that in KEGG (Fig. 2H and table S8). With respect to coexpression, 63% of the human genes encoding proteins in complexes were significantly coexpressed, which was comparable to the proportion in KEGG and twofold higher than that in GO (Fig. 2H and table S9). *Drosophila* complexes showed a similar enrichment of coexpressed pairs, and the fraction was significantly higher than that in GO. Together, these results indicate that the complexes that we compiled are more likely to be accurate and physiologically relevant than those identified in GO and KEGG, and these complexes represent an alternative resource for enrichment analysis.

Finally, we annotated the complexes on the basis of either literature annotation (for the literature-based complexes) or GO term enrichment. For literature-based complexes, we kept track of the complex nomenclature, purification method, references, and the species from which the complex was identified. The complexes for which such annotation is not available were annotated with up to five of the most informative GO terms enriched for complex members. For *Drosophila*, 50% of the complexes were annotated on the basis of the published literature, and 40% on the basis of GO term enrichment. Ten percent of the *Drosophila* complexes could not be annotated, suggesting either that they are previously unknown protein complexes with well-known protein components or that they contain unannotated components (Fig. 2I and table S10). Finally, each complex was associated with coexpression, colocalization, and co-citation information, and information about both known PPIs and interacting homologs in other organisms (known as interologs) was included.

Developing an interactive protein complex enrichment analysis tool

To analyze high-throughput data sets, we developed COMPLEAT. The tool handles complete high-throughput data without preselecting hits, although preselected hits could also be used as the input. First, individual protein or gene values from a data set are mapped to complexes. Next, the complexes are assigned scores by calculating the interquartile mean (IQM) of data points corresponding to individual protein components of the complex (see Materials and Methods; fig. S1). By assigning an IQM to each complex on the basis of the input data, COMPLEAT preserves the direction (stimulation or inhibition, or increased or decreased abundance) and the magnitude of changes associated with the individual components. For some complexes, the data corresponding to individual protein components within a complex include both positive and negative z scores or fold-change values, meaning that the complex is "incoherent," and IQM values of such incoherent complexes tend to be insignificant (*24*). Furthermore, COMPLEAT computes a P value to estimate the significance of complex scores as compared to 1000 random complexes of the same size. A key feature of COMPLEAT is that it enables comparison of multiple data sets. In such cases, the enrichment analysis is performed for each data set independently and the complex scores are compared. COMPLEAT also provides a Cytoscape-based visualization of the enriched complexes (*25*).

COMPLEAT (fig. S2) is accessible through a Web-based interface, where users can upload single or multiple data sets from small-scale or high-throughput studies. The tool accepts any type of data set that associates genes or proteins from human, *Drosophila*, or yeast with normalized values or scores, including z scores (from RNAi screens) or fold-change values (from gene expression analysis). COMPLEAT supports a number of commonly used identifiers, including Entrez gene identifier, UniProt identifier, and species-specific database identifiers (table S11). The tool calculates complex scores for each data set, and the results are visualized as an interactive scatter plot using iCanPlot (http://www.icanplot.org/) (*26*). In the case of multiple data sets (data from multiple conditions or time points), the user can choose which data sets to display on the x and y axes. In addition, the tool has a search box that allows the user to interactively query a complex or gene of interest, which then becomes highlighted in the scatter plot. The tool supports complex query functions, including search with Boolean operators "AND," "OR," and "NOT." Further, the user can restrict the search to specific fields like gene names, complex names, or complex resource (for example, to select literature-based complexes). Because the complex resource preserves redundant configurations of the complexes, the tool provides an option to hide redundant complexes and to select only nonredundant enriched complexes.

Moreover, using the interactive scatter plot, users can select complexes of interest (based on complex score or P value), and the network illustrations of the selected complexes are displayed on the same screen [using Cytoscape Web (*27*)]. When comparing multiple data sets, the tool displays the complexes from each data set side by side. Users can obtain more information about a given complex or proteins within the complex by clicking on that complex or gene in the visualized network. Finally, the tool provides the option to save the enriched complexes as a table, along with associated values, and can export scatter plots as well as Cytoscape visualizations of selected complexes as image files.

Analysis of genome-wide, cell-based RNAi screens to identify dynamic regulation of protein complexes

To demonstrate the usefulness of COMPLEAT, we used the tool to analyze the dynamic regulation of protein complexes after either insulin or EGF stimulation. We analyzed the results from four genome-wide, cell-based RNAi screens in *Drosophila* cells aimed at identifying components of the ERK signaling pathway (*5*, *6*). RNAi screens, measuring the abundance of pERK normalized to total ERK as the output, were performed in Schneider 2 receptor-plus (S2R+) cells in the absence (baseline) of either EGF or insulin stimulation and at 10 min after addition of either EGF or insulin. With pERK as the phenotypic readout, we identified complexes that behaved consistently across the baseline and stimulus conditions (common complexes) and others that showed dynamic changes (dynamic complexes) (tables S12 to S15).

Analysis of the baseline versus EGF-stimulated data sets revealed 11 common complexes and 184 dynamic complexes (Fig. 3, A and B, tables

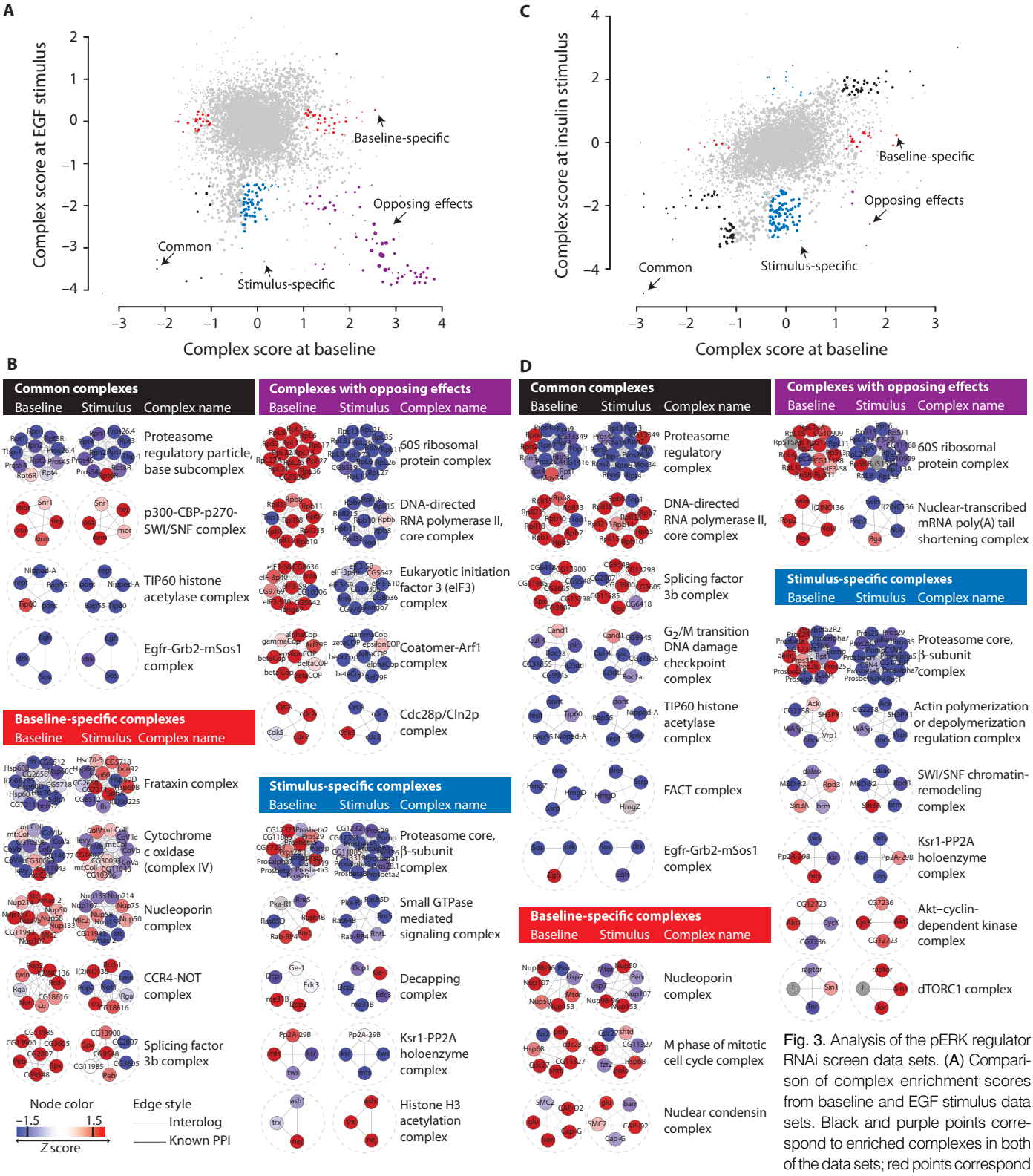

Fig. 3. Analysis of the pERK regulator RNAi screen data sets. (**A**) Comparison of complex enrichment scores from baseline and EGF stimulus data sets. Black and purple points correspond to enriched complexes in both of the data sets; red points correspond to significant complexes at baseline; and blue points correspond to stimulus-specific complexes. (**B**) Examples of protein complexes for each dynamic class from the baseline and stimulus (EGF) analysis. (**C**) Comparison of complex enrichment scores from the baseline and insulin stimulus data sets [color code same as in (A)]. (**D**) Examples of protein complexes for each dynamic class from the baseline and insulin stimulus analysis.

S16 and S17, and figs. S3 to S7). Analysis of baseline versus insulin-stimulated cells identified 84 common complexes and 110 dynamic complexes (Fig. 3, C and D, tables S18 and S19, and figs. S8 to S11). Among the common complexes found in the EGF data sets was the EGF receptor (EGFR) complex, consisting of EGFR, drk (a homolog of human GRB2), and Sos (son of sevenless homolog), which binds EGF and activates downstream signaling. In addition, the *Drosophila* TORC1 complex, consisting of Tor, Raptor, Lobe, and Sin1 (*28*), and the Akt1-CDK (cyclin-dependent kinase) complex, which includes core components of the insulin pathway, were among the dynamic complexes found only under the insulin stimulus condition. Among all of the dynamic complexes, we identified complexes that regulate pERK only in the presence or absence of stimulus, meaning that most complex members scored similarly only in the baseline (baseline-specific) or in the stimulus condition (stimulus-specific), and others that regulate pERK in both conditions but for which the stimulus appears to act as a switch between positive and negative regulation. For example, the Ksr1 (kinase suppressor of Ras)–PP2A (protein phosphatase 2) holoenzyme complex scored positively under the EGF stimulus condition (Fig. 3B), suggesting that the assembly or disassembly of the Ksr-PP2A complex can be potentially regulated by EGF signaling. Ksr1 is a conserved scaffold that facilitates signal propagation through the MAPK pathway and PP2A is a critical regulator of Ksr1 (*29*). Further, our analysis indicated that the coatomer-Arf1 (ADP-ribosylation factor 1) complex, which mediates transport between cellular compartments by coated vesicles and is regulated by MAPK in mammalian systems (*30*), acted as a negative regulator of pERK at baseline but a positive regulator in response to EGF (Fig. 3B), suggesting that EGF signaling changes the activity of this complex. Finally, all members of the nucleoporin complex, implicated as a scaffold for signal propagation (*31*), scored negatively only at baseline, suggesting that EGF signaling potentially regulates this complex.

Validation of COMPLEAT's prediction of the Brahma complex in the insulin response

In addition to the core components of the insulin pathway, we identified the Brahma complex (also called SWI/SNF chromatin-remodeling complex, a transcriptional regulator that activates many transcription factors or regulates global chromatin remodeling to facilitate transcription), only when the cells were stimulated with insulin, suggesting that this complex plays a role in the insulin response (Figs. 3D and 4A). A role for Brahma in insulin signaling was supported by the observation that 2 (Moira and MBD-R2) of 11 Brahma complex components were phosphorylated within 10 min after insulin treatment (Fig. 4A and table S20). Further, ribosomal S6 kinase 2 (S6KII), a core component of the insulin pathway, associated with the Brahma complex component Dalao at 10 min after stimulation (Fig. 4B), and this association was enhanced by insulin treatment.

To validate a role for the Brahma complex in insulin signaling, we examined its role in the regulation of cell cycle genes because insulin signaling regulates both growth and cell proliferation (*32*–*37*). Overexpression of the Brahma complex component *dalao* in S2R+ cells reduced the amount of *cyclin D* mRNA expression (Fig. 4C). Further, consistent with a role for the Brahma complex downstream of insulin signaling, inhibition of *cyclin D* expression by ectopic expression of *dalao* was relieved when cells were treated with insulin (Fig. 4C). Finally, to test the requirement of the Brahma complex in vivo, we analyzed whether it regulated muscle mass and nuclear and nucleolus sizes in fly larval muscles, a process under the control of insulin signaling (*38*, *39*). Knockdown of Brahma complex components, *brahma*, *dalao*, and *moira*, by RNAi in larval muscle resulted in larger muscle fibers, nuclei, and nucleoli (Fig. 4D), which is consistent with an increase in insulin signaling (*38*).

DISCUSSION

GO and pathway annotations have been the most common back-end annotation sources for high-throughput data mining (*1*). To complement these resources, we have created comprehensive protein complex resources for human, *Drosophila*, and yeast. In building such resources in parallel for three species, we took advantage of evolutionary conservation, increasing coverage for each individual species. In the case of literature-curated complexes, 40% of human and 80% of *Drosophila* complexes were mapped from other species. Spurious interactions in the PPI networks are of great concern for complex prediction approaches, and we addressed this issue at the stage of data collection and by filtering the PPIs with additional data sets. The poor overlap between literature-curated and predicted complexes in the complex resources is mainly due to the fact that the resources capture different proteins. Currently, both resources are complementary, and the overlap will improve as more PPIs are identified. The current resource covers almost 70% of the yeast proteome and half of the human and *Drosophila* proteomes, and we expect this coverage to improve in the future as new data become available from ongoing PPI mapping projects, including studies that map interactomes across multiple conditions, species, or time points (*9*, *16*, *40*–*44*). To handle the dynamic or alternative forms of protein complexes, such as "core-complex" with different "attachments" (*43*), we preserved all possible configurations of protein complexes reported in the complex databases and those by the prediction tools. Comparison of the biological relevance of the protein complexes with GO and KEGG pathway annotations, for example, in the context of coexpression, colocalization, and co-citation, reveals the high quality of the resource.

Using the complex resource as a foundation, we developed COMPLEAT, an enrichment analysis tool. About 70 enrichment analysis tools are currently available that can be broadly classified as tools facilitating singular enrichment analysis (SEA), GSEA, or modular enrichment analysis (MEA) (*1*). Most of these tools depend on GO or pathways as the back-end annotation data. COMPLEAT is unique with respect to back-end annotation because it uses our newly compiled protein complex resources and incorporates many useful features from other tools. COMPLEAT is flexible in handling high-throughput data because the tool accepts complete lists, similar to GSEA, as well as preselected hit lists, similar to SEA and MEA. COMPLEAT integrates experimental values (for example, *z* score or fold change) into the enrichment calculations, similar to recently reported GSEA tools (*1*). Indeed, the COMPLEAT complex scores directly reflect the experimental values of individual genes. A major limitation that is consistent across all the SEA, GSEA, and MEA tools is that a few highly changing (ranking) genes drive the enrichment calculation (*1*). We handled this issue by calculating the IQM of the values. Like the median, the IQM is robust to outliers because the lowest 25% and the highest 25% of scores are ignored. Like the mean, the IQM takes into account a much broader distribution because values from 50% of the complex members are included. The complex scores preserve the sign from the data set, such that the score directly indicates that a given complex is under- or overexpressed and a positive or negative regulator (depending on the type of data analyzed). Furthermore, the score also enriches the complexes with members that have coherence scores (*24*).

Instead of a long list of enriched annotation terms as output, COMPLEAT provides visualization of the data within a comprehensive data mining environment. For example, it supports interactive querying systems, where the user can interactively optimize thresholds to select complexes and query for a specific complex or gene of interest. In addition, the network-based visualization of the complexes helps to visualize individual gene scores in the context of known protein complexes, which helps generate specific hypotheses and design follow-up experiments.

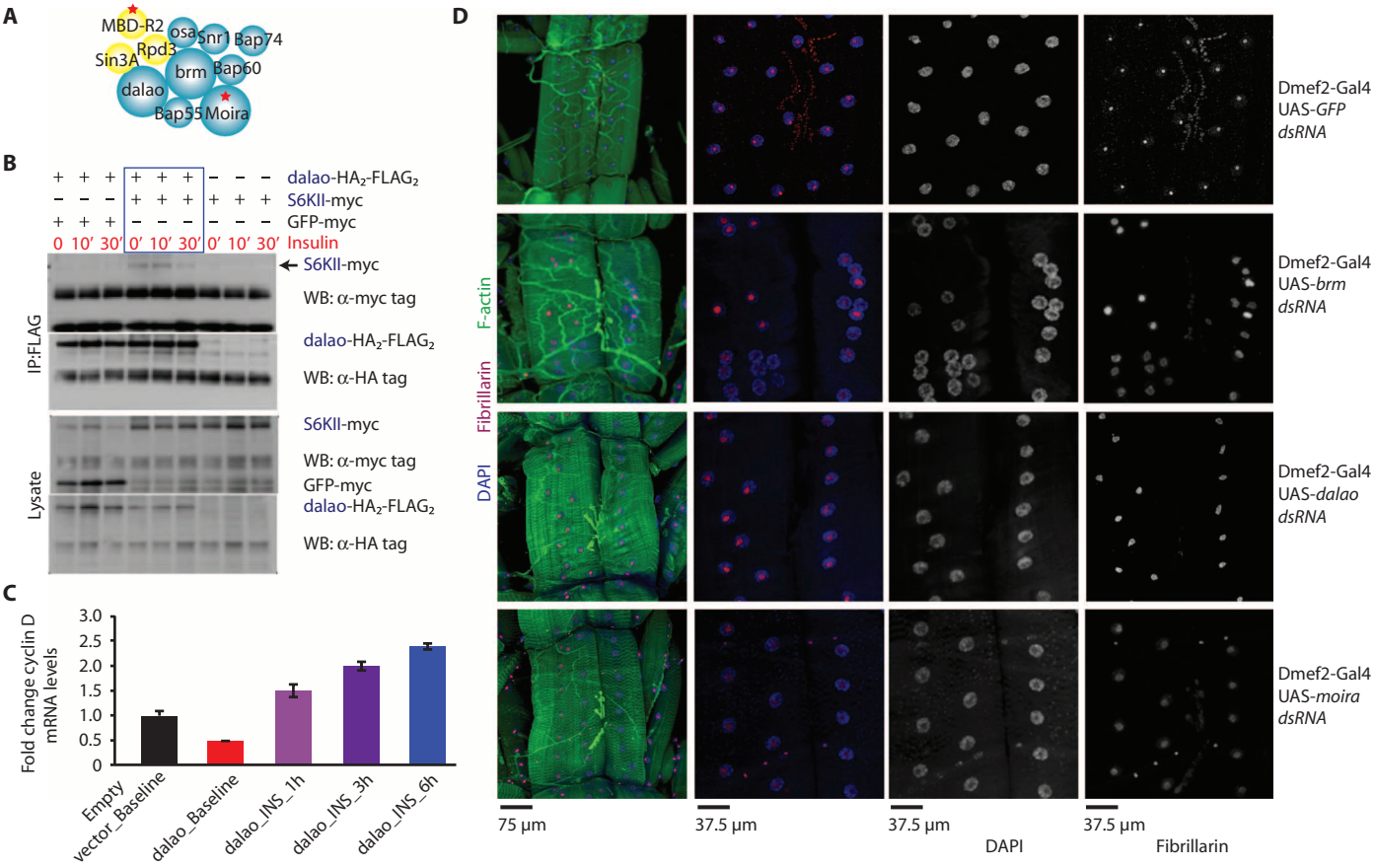

Fig. 4. Functional validation of the Brahma protein complex. (**A**) Schematic representation of Brahma protein complex members; core components and associated proteins according to COMPLEAT are shown in blue and yellow, respectively (*50*, *51*). The dynamic phospho-regulated proteins in response to insulin stimulus are indicated with a red star. (**B**) Coimmunoprecipitation for the FLAG tag followed by Western blotting for the myc and hemagglutinin (HA) tags on lysates from S2R+ cells cotransfected with FLAG and HA double-tagged Dalao protein and either myc-tagged S6KII or myc-tagged GFP. The effect of insulin on the physical interaction between Dalao and S6KII was measured at 10 and 30 min after treatment. One representative blot is shown from two independent experiments. (**C**) Comparison of fold change expression in cyclin D mRNA in untreated cells, cells overexpressing *dalao*, and cells overexpressing *dalao* and stimulated with insulin for 1, 3, and 6 hours, measured by real-time quantitative reverse transcription PCR (RT-PCR). Error bars indicate SD (n = 3). (**D**) Nuclei, nucleoli, and muscle fibers are larger in L3 *Drosophila* larvae in which the Brahma-containing SWI/SNF complex was knocked down by RNAi. Phalloidin (green) and 4′,6-diamidino-2-phenylindole (DAPI; blue) stained the body wall muscle fibers and nuclei, respectively, and fibrillarin (red) stained the nucleoli. Data shown are representative of experiments from five larvae.

We used COMPLEAT to directly compare z scores from pERK regulators identified in the absence or presence of stimulus (EGF or insulin) and distinguished two kinds of dynamic complexes. We identified complexes regulating pERK in the presence or absence of stimulus, where the stimulus may trigger complex assembly or disassembly. We also identified complexes in which the stimulus appeared to act as a switch between positive and negative regulation. In the latter case, the stimulus may regulate the output of the complex, for example, whether the complex functions as an activator or inhibitor, rather than assembly or disassembly. Thus, using COMPLEAT, we have analyzed multiple data sets to identify dynamic protein complexes based on pERK abundance. These analyses allow the generation of specific hypotheses that can then be validated experimentally, as we showed in the case of the Brahma complex.

In summary, we developed a protein complex–based analysis tool that efficiently addresses a current limitation in high-throughput data analysis. The tool uses a comprehensive protein complex resource for back-end annotation and also incorporates several key features from various other tools. The tool provides a data-mining environment supported by network-based visualization and can be applied to analyze not only functional RNAi and overexpression screens but also results from genome-wide association studies and exome sequencing projects. COMPLEAT may prove useful for identifying human disease genes because different members of the same protein complex often lead to common disease phenotypes (*45*). Further, the tool enables direct comparison of multiple data sets and integration of heterogeneous data sets. Thus, COMPLEAT complements the existing enrichment analysis tools to provide a different dimension to the interpretation of high-throughput data.

MATERIALS AND METHODS

Compilation of literature-based complexes

Literature-based complex information was retrieved from databases such as CORUM, PINdb, CYC2008, GO, KEGG, and *Drosophila* AP-MS pull-down complexes (table S1). With the exception of protein complexes that are annotated by GO, all the other complexes were mapped across human, *Drosophila*, and yeast. We used DIOPT (http://www.flyrnai.org/diopt), an integrative ortholog prediction tool, to map orthologs among human, *Drosophila*, and yeast. DIOPT scores were used to select the best ortholog match in case of "one-to-many" ortholog relationships (DIOPT score cutoff ≥2). Only complexes consisting of two or more proteins are included in the complex resources. Complex annotations from the source databases, including complex name, purification method, and PubMed ID, are also included in the resources.

Applying CFinder to predict protein complexes

CFinder (http://www.cfinder.org/) was downloaded and implemented locally. We applied CFinder to identify protein complexes from human, *Drosophila*, and yeast PPI networks. We filtered the PPI networks using co-expression values or colocalization information to remove low-confidence PPIs. The coexpression values were used to filter human and *Drosophila* PPI networks, and only edges with Pearson correlation ≥0.2 were retained. For the yeast network, we used colocalization information and retained only those PPIs where the subcellular localization of both proteins is known and the proteins colocalize. We used the filtered human, *Drosophila*, and yeast PPI networks as input and ran CFinder with the default parameters. The outputs were analyzed and integrated with Perl scripts.

Applying NetworkBLAST to predict protein complexes

NetworkBLAST was used to identify evolutionarily conserved protein complexes by aligning two networks from different species. NetworkBLAST was downloaded (http://www.cs.tau.ac.il/~bnet/networkblast.htm) and implemented locally. We used stringent parameters to align human, *Drosophila*, and yeast networks. The complex density was set to 0.95, and false negatives to 0.2, 0.2, and 0.1 for human, *Drosophila*, and yeast networks, respectively. The outputs were analyzed and integrated with Perl scripts.

Literature co-citation

For each protein in the complex resource, we retrieved the corresponding literature from the National Center for Biotechnology Information (NCBI) gene resource (http://www.ncbi.nlm.nih.gov/). We only selected articles associated with 2 to 50 genes to eliminate high-throughput data that may be associated with false discovery. For all possible pairs of proteins in a protein complex, we extracted the pairs that are co-cited in the same publication(s). To assess the significance of co-citation, we computed the *P* value for each complex by comparing these results with the results obtained using a set of 1000 randomly generated protein complexes of the same size. For co-citation annotation of each complex, we ranked the articles based on the fraction of the pairs co-cited and selected the top 10 articles for display at COMPLEAT.

Coexpression

Coexpression data for human and *Drosophila* were downloaded from COXPRESdb (*46*). COXPRESdb provides weighted Pearson's correlation coefficients for gene pairs based on 4401 and 1102 microarray experiments corresponding to human and *Drosophila,* respectively. For all possible pairs of proteins in a protein complex, we extracted the coexpression values (weighted Pearson's correlation coefficients) and computed the average coexpression value. To assess the significance of the average co-expression value, we computed the *P* value for each complex by comparing the results with those obtained using a set of 1000 randomly generated protein complexes of the same size.

Colocalization

The localization annotation for yeast proteins was obtained from the UniProt database (http://www.uniprot.org/) and the Yeast GFP Fusion Localization Database (http://yeastgfp.yeastgenome.org/) (*23*). Subcellular localization annotations were consolidated and simplified. For example, "nucleolus" was included in the broader category "nucleus," and "ER to Golgi" was mapped to "ER" and "Golgi." Next, the proteins were annotated based on their localization to one or more of the following subcellular locations: peroxisome, nucleus, mitochondrion, lysosome, Golgi, ER, endosome, and cytoplasm. For all possible pairs of proteins in a protein complex, we extracted the pairs that colocalize to the same subcellular location. To assess the significance of colocalization, we computed the *P* value for each complex by comparing the results with the results obtained using a set of 1000 randomly generated protein complexes of the same size.

Complex scoring

Values from the input data were mapped to complex members and sorted highest to lowest. The complex score was computed as the IQM as follows:

$$C_{\mathrm{IQM}} = \frac{1}{(Q_3 - Q_1) + 1} \sum_{i=Q_1}^{Q_3} x_i$$

$$Q_1 = \frac{n}{4} + 1 \qquad Q_1 \in \mathbf{Z}$$

$$Q_3 = \frac{3n}{4} \qquad Q_3 \in Z$$

where n is the number of proteins in the complex, x_i is the score of ith protein in the complex, and Q_1 and Q_3 are the integers of the first and third quartiles, respectively.

Implementation of COMPLEAT

The COMPLEAT user interface was implemented as a collection of Java servlets, JavaScript, and Adobe Flash components. COMPLEAT integrates existing tools, including Cytoscape Web for complex visualization (http://cytoscapeweb.cytoscape.org/), iCanPlot for plotting scores (http://www.icanplot.org/), and WebFX for parameter adjustment sliders (http://webfx.eae.net/). The application is hosted on the Orchestra cluster supported by the Research IT Group (RITG) at Harvard Medical School. The complexes and associated annotations and relationships are maintained as flat files. The Java Servlets and Java Server Pages run within an instance of Tomcat6.0.18 on the Orchestra cluster. The Entrez gene identifiers, symbols, locus tags, and alias names for human, *Drosophila*, and yeast genes were retrieved from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/). For *Drosophila* genes, their FlyBase gene identifiers, CG numbers, symbols, and synonyms were also retrieved from FlyBase (ftp://ftp.flybase.net/releases/current/). Protein identifiers were retrieved from UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/). A program developed in-house automatically transfers and processes these files on a monthly basis.

Complex enrichment analysis of pERK RNAi data sets

We selected four RNAi data sets aimed at identifying components of the ERK pathway (*5*, *6*). Briefly, the first set of RNAi screens were performed in an S2R+ cell line expressing *Drosophila* EGFR (DER) from the metallothionein promoter (S2R+mtDER). The screens were performed in the absence of stimulus and at 10 min after treatment with Spitz-containing

conditioned medium (EGF in mammals), which activates the Rolled kinase and increases the abundance of phosphorylated Rolled (known as ERK or MAPK in mammals). The second set of RNAi screens were performed in S2R+ cell lines in the absence of stimulus and at 10 min after treatment with bovine insulin (Sigma; 25 μg/ml). The amounts of pERK and endogenous ERK were measured to determine the *z* scores for each RNAi experiment (*5*, *6*), which we used as input for the complex enrichment analysis. The complex enrichment analysis was performed independently for all four data sets. To study the dynamics of complex regulation by EGF, we compared the complex scores at baseline (S2R+mtDER cell) with that after 10-min treatment with EGF stimulus (S2R+mtDER cell). Similarly, to study insulin dynamics, we compared baseline (S2R+ cells) with 10-min insulin treatment (S2R+ cells).

We applied greedy algorithm to select nonredundant representative complexes as shown in Fig. 3, B and D, and used both *P* value and score cutoff to define a protein complex as "enriched" in a particular data set. For baseline data sets (S2R+mtDER and S2R+), we used a *P* value cutoff of 0.01 and an IQM cutoff of <−1 or >1. In case of stimulus data sets (both EGF and insulin), we used a *P* value cutoff of 0.01 and an IQM cutoff of −1.5 or 1.5. To define a protein complex as "not enriched" in a data set, we used a *P* value cutoff of >0.25 and low IQM values (values between −0.5 and 0.5). The complexes enriched in both the baseline and stimulus data sets are grouped as "common" complexes. The complexes enriched in both data sets but which have opposing IQM values are grouped as "opposing effects" (for example, negative score in baseline and positive score in stimulus condition, or vice versa). If the complexes are enriched in the baseline but not the stimulus data set, the complexes are grouped as "baseline-specific." Similarly, complexes enriched in the stimulus but not the baseline data set are grouped as "stimulus-specific."

Clustering protein complexes to select nonredundant complexes

We used a greedy algorithm to cluster significant complexes and select nonredundant representatives. The complexes were sorted based on size (largest to smallest), and the largest complex was selected as representative of the cluster. A complex was assigned to an existing cluster if it was a subset or shared at least 80% similarity to the representative cluster. If a complex did not match an existing cluster, it became the representative complex for a new cluster. This process was iterated until all complexes were placed in appropriate clusters.

Coimmunoprecipitation and Western blotting

An expression construct for FLAG and HA double-tagged Dalao protein was cotransfected with either a myc-tagged S6KII construct or myc-GFP as a control in S2R+ cells. Twenty-four hours after transfection, cells were treated with copper sulfate to induce expression of the tagged proteins. Twenty-four hours later, cells were either untreated or stimulated with insulin for 10 and 30 min. Total cell lysates were prepared and immunoprecipitated with anti-FLAG M2 affinity gel (Sigma-Aldrich). Cell lysates and immunoprecipitated samples were separated by SDS–polyacrylamide gel electrophoresis and transferred onto a polyvinylidene difluoride membrane. The association between S6KII and Dalao was demonstrated by probing the membrane with antibody against the myc tag (Cell Signaling Technologies). Protein input blots were probed with the myc-tag antibody and an antibody against the HA tag (Roche Diagnostics, clone 3F10).

Quantitative RT-PCR

Drosophila S2R+ cells were transfected with either empty vector or an expression construct for FLAG and HA double-tagged Dalao protein. Twenty-four hours after transfection, cells were treated with copper sulfate to induce expression of the tagged protein. Twenty-four hours later, cells were untreated or stimulated with insulin for 1, 3, or 6 hours. Total RNA was prepared from cells with Trizol (Invitrogen), followed by the RNeasy kit (Qiagen). The iScript cDNA Synthesis kit (Bio-Rad) was used for complementary DNA (cDNA) synthesis, and quantitative RT-PCR was performed with the iQ SYBR Green Supermix (Bio-Rad). Rp49 was used as normalization reference. Relative quantitation of mRNA expression was calculated using the comparative C_T method. The primers used were rp49, 5′-ATCGGTTACGGATCGAACAA-3′ (forward) and 5′-GACAATCTCCTTGCGCTTCT-3′ (reverse), and *cyclin D*, 5′-GCCGAATGGATGATGGAA-3′ (forward) and 5′-CCATGTAATTTAATGCCAGTAATACG-3′ (reverse).

Fly stocks

Dmef2-Gal4 drives transgene expression in all body wall muscles. For transgene expression with the Gal4-UAS system (*47*), flies were reared at 25°C. Hairpin lines were obtained from the TRiP facility at Harvard Medical School [UAS-brm dsRNA (HMS00050) and UAS-moira (HMS01267)] or from the Vienna Stock Center [UAS-dalao (KK1044361)].

Histology, laser-scanning confocal microscopy, and image analysis

Wandering third instar larvae were dissected in ice-cold Ca^{2+}-free saline buffer [128 mM NaCl, 2 mM KCl, 4 mM $MgCl_2$, 1 mM EGTA, 35 mM sucrose, 5 mM Hepes (pH 7.2)] in a dissection chamber (*48*). Body wall muscles were fixed for 20 to 30 min in Ca^{2+}-free saline buffer containing 4% paraformaldehyde and 0.1% Triton X-100. After being washed, body wall muscles were incubated for 10 hours with DAPI (1 μg/ml) and Alexa 633–conjugated phalloidin (1:100) to visualize nuclei and F-actin, respectively. To examine biogenesis of nucleoli, an antibody against fibrillarin [EnCore Biotechnology (*49*)] was applied (1:100), followed by incubation with Alexa 555–conjugated secondary antibodies (Molecular Probes). Muscles VL3 and VL4 of abdominal segments 2 to 5 were imaged with a Leica TCS SP2 confocal laser-scanning microscope.

SUPPLEMENTARY MATERIALS

www.sciencesignaling.org/cgi/content/full/6/264/rs5/DC1

Fig. S1. Schematic representation of protein complex scoring.
Fig. S2. Snapshots of the COMPLEAT Web interface.
Fig. S3. Complex enrichment results of baseline and EGF stimulus.
Fig. S4. Baseline compared with EGF stimulus common complexes.
Fig. S5. Baseline compared with EGF stimulus dynamic complexes: opposing effects.
Fig. S6. Baseline compared with EGF stimulus: baseline-specific dynamic complexes.
Fig. S7. Baseline compared with EGF stimulus: stimulus-specific dynamic complexes.
Fig. S8. Complex enrichment results of baseline and insulin stimulus.
Fig. S9. Baseline compared with insulin stimulus: common complexes.
Fig. S10. Baseline compared with insulin stimulus: baseline-specific dynamic complexes.
Fig. S11. Baseline compared with insulin stimulus: stimulus-specific dynamic complexes.
Table S1. Compilation of literature protein complexes for humans, *Drosophila*, and yeast.
Table S2. PPI data sets used to construct integrated PPI networks for humans, *Drosophila*, and yeast.
Table S3. Predicted protein complexes for humans, *Drosophila*, and yeast.
Table S4. Redundancy in the protein complex resource.
Table S5. Overlap of the literature and predicted complexes at the protein level.
Table S6. Proteome covered by the protein complex resources.
Table S7. Comparison of protein complexes with GO and KEGG with respect to co-citation.
Table S8. Comparison of protein complexes with GO and KEGG with respect to protein colocalization.
Table S9. Comparison of protein complexes with GO and KEGG with respect to gene coexpression.
Table S10. Annotation of the protein complex resource.
Table S11. Gene or protein input identifiers supported by the COMPLEAT.
Table S12. Enriched protein complexes at baseline (mtDER-S2R+ cell line).
Table S13. Enriched protein complexes at EGF stimulus (mtDER-S2R+ cell line).
Table S14. Enriched protein complexes at baseline (S2R+ cell line).
Table S15. Enriched protein complexes at insulin stimulus (S2R+ cell line).

Table S16. Consistent protein complexes with respect to baseline versus EGF stimulus.
Table S17. Dynamic protein complexes with respect to baseline versus EGF stimulus.
Table S18. Consistent protein complexes with respect to baseline versus insulin stimulus.
Table S19. Dynamic protein complexes with respect to baseline versus insulin stimulus.
Table S20. Dynamic phosphosites changing in response to insulin treatment.

REFERENCES AND NOTES

1. D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
2. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
3. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
4. L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray, From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
5. A. Friedman, N. Perrimon, A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. *Nature* **444**, 230–234 (2006).
6. A. A. Friedman, G. Tucker, R. Singh, D. Yan, A. Vinayagam, Y. Hu, R. Binari, P. Hong, X. Sun, M. Porto, S. Pacifico, T. Murali, R. L. Finley Jr., J. M. Asara, B. Berger, N. Perrimon, Proteomic and functional genomic landscape of receptor tyrosine kinase and ras to extracellular signal–regulated kinase signaling. *Sci. Signal.* **4**, rs10 (2011).
7. M. Barrios-Rodiles, K. R. Brown, B. Ozdamar, R. Bose, Z. Liu, R. S. Donovan, F. Shinjo, Y. Liu, J. Dembowy, I. W. Taylor, V. Luga, N. Przulj, M. Robinson, H. Suzuki, Y. Hayashizaki, I. Jurisica, J. L. Wrana, High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307**, 1621–1625 (2005).
8. N. Bisson, D. A. James, G. Ivosev, S. A. Tate, R. Bonner, L. Taylor, T. Pawson, Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor. *Nat. Biotechnol.* **29**, 653–658 (2011).
9. T. Ideker, N. J. Krogan, Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
10. J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, M. Mann, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
11. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
12. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
13. P. V. Luc, P. Tempst, PINdb: A database of nuclear protein complexes from human and yeast. *Bioinformatics* **20**, 1413–1415 (2004).
14. S. Pu, J. Wong, B. Turner, E. Cho, S. J. Wodak, Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37**, 825–831 (2009).
15. A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H. W. Mewes, CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).
16. 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, S. Artavanis-Tsakonas, A protein complex network of *Drosophila melanogaster*. *Cell* **147**, 690–703 (2011).
17. Y. Hu, I. Flockhart, A. Vinayagam, C. Bergwitz, B. Berger, N. Perrimon, S. E. Mohr, An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* **12**, 357 (2011).
18. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
19. R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, T. Ideker, Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1974–1979 (2005).
20. J. Song, M. Singh, How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics* **25**, 3143–3150 (2009).
21. R. Sharan, T. Ideker, Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* **24**, 427–433 (2006).
22. A. L. Barabási, Z. N. Oltvai, Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
23. W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O'Shea, Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
24. M. Michaut, A. Baryshnikova, M. Costanzo, C. L. Myers, B. J. Andrews, C. Boone, G. D. Bader, Protein complexes are central in the yeast genetic landscape. *PLoS Comput. Biol.* **7**, e1001092 (2011).
25. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
26. A. U. Sinha, S. A. Armstrong, iCanPlot: Visual exploration of high-throughput omics data using interactive Canvas plotting. *PLoS One* **7**, e31690 (2012).
27. C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, G. D. Bader, Cytoscape Web: An interactive web-based network browser. *Bioinformatics* **26**, 2347–2348 (2010).
28. T. Glatter, R. B. Schittenhelm, O. Rinner, K. Roguska, A. Wepf, M. A. Jünger, K. Köhler, I. Jevtov, H. Choi, A. Schmidt, A. I. Nesvizhskii, H. Stocker, E. Hafen, R. Aebersold, M. Gstaiger, Modularity and hormone sensitivity of the *Drosophila melanogaster* insulin receptor/target of rapamycin interaction proteome. *Mol. Syst. Biol.* **7**, 547 (2011).
29. S. Ory, M. Zhou, T. P. Conrads, T. D. Veenstra, D. K. Morrison, Protein phosphatase 2A positively regulates Ras signaling by dephosphorylating KSR1 and Raf-1 on critical 14-3-3 binding sites. *Curr. Biol.* **13**, 1356–1364 (2003).
30. H. Farhan, M. W. Wendeler, S. Mitrovic, E. Fava, Y. Silberberg, R. Sharan, M. Zerial, H. P. Hauri, MAPK signaling to the early secretory pathway revealed by kinase/phosphatase functional screening. *J. Cell Biol.* **189**, 997–1011 (2010).
31. N. Xylourgidis, M. Fornerod, Acting out of character: Regulatory roles of nuclear pore complex proteins. *Dev. Cell* **17**, 617–625 (2009).
32. D. C. Goberdhan, N. Paricio, E. C. Goodman, M. Mlodzik, C. Wilson, *Drosophila* tumor suppressor PTEN controls cell size and number by antagonizing the Chico/PI3-kinase signaling pathway. *Genes Dev.* **13**, 3244–3258 (1999).
33. S. E. Scanga, L. Ruel, R. C. Binari, B. Snow, V. Stambolic, D. Bouchard, M. Peters, B. Calvieri, T. W. Mak, J. R. Woodgett, A. S. Manoukian, The conserved PI3'K/PTEN/Akt signaling pathway regulates both cell size and survival in *Drosophila*. *Oncogene* **19**, 3971–3977 (2000).
34. W. Brogiolo, H. Stocker, T. Ikeya, F. Rintelen, R. Fernandez, E. Hafen, An evolutionarily conserved function of the *Drosophila* insulin receptor and insulin-like peptides in growth control. *Curr. Biol.* **11**, 213–221 (2001).
35. O. Puig, M. T. Marr, M. L. Ruhf, R. Tjian, Control of cell number by *Drosophila* FOXO: Downstream and feedback regulation of the insulin receptor pathway. *Genes Dev.* **17**, 2006–2020 (2003).
36. M. A. Jünger, F. Rintelen, H. Stocker, J. D. Wasserman, M. Végh, T. Radimerski, M. E. Greenberg, E. Hafen, The *Drosophila* forkhead transcription factor FOXO mediates the reduction in cell number associated with reduced insulin signaling. *J. Biol.* **2**, 20 (2003).
37. C. J. Potter, H. Huang, T. Xu, *Drosophila Tsc1* functions with *Tsc2* to antagonize insulin signaling in regulating cell growth, cell proliferation, and organ size. *Cell* **105**, 357–368 (2001).
38. F. Demontis, N. Perrimon, Integration of insulin receptor/Foxo signaling and dMyc activity during muscle growth regulates body size in *Drosophila*. *Development* **136**, 983–993 (2009).
39. D. J. Hall, S. S. Grewal, A. F. de la Cruz, B. A. Edgar, Rheb-TOR signaling promotes protein synthesis, but not glucose or amino acid import, in *Drosophila*. *BMC Biol.* **5**, 10 (2007).
40. A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, E. E. Wanker, A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* **4**, rs8 (2011).
41. M. E. Sowa, E. J. Bennett, S. P. Gygi, J. W. Harper, Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009).
42. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E. E. Wanker, A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
43. A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, 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, G. Superti-Furga, Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
44. N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-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, J. F. Greenblatt, Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).

45. K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, S. Brunak, A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
46. T. Obayashi, K. Kinoshita, COXPRESdb: A database to compare gene coexpression in seven model animals. *Nucleic Acids Res.* **39**, D1016–D1022 (2011).
47. A. H. Brand, N. Perrimon, Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).
48. V. Budnik, M. Gorczyca, A. Prokop, Selected methods for the anatomical study of *Drosophila* embryonic and larval neuromuscular junctions. *Int. Rev. Neurobiol.* **75**, 323–365 (2006).
49. S. S. Grewal, L. Li, A. Orian, R. N. Eisenman, B. A. Edgar, Myc-dependent regulation of ribosomal RNA synthesis during *Drosophila* development. *Nat. Cell Biol.* **7**, 295–302 (2005).
50. A. J. Kal, T. Mahmoudi, N. B. Zak, C. P. Verrijzer, The *Drosophila* Brahma complex is an essential coactivator for the trithorax group protein zeste. *Genes Dev.* **14**, 1058–1071 (2000).
51. O. Papoulas, S. J. Beek, S. L. Moseley, C. M. McCallum, M. Sarte, A. Shearn, J. W. Tamkun, The *Drosophila* trithorax group proteins BRM, ASH1 and ASH2 are subunits of distinct protein complexes. *Development* **125**, 3955–3966 (1998).
52. D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, L. Stein, Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
53. S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, H. Hermjakob, The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
54. L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
55. L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, G. Cesareni, MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
56. N. Dephoure, S. P. Gygi, Hyperplexing: A method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Sci. Signal.* **5**, rs2 (2012).
57. J. Villén, S. P. Gygi, The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1630–1638 (2008).
58. E. L. Huttlin, M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villén, W. Haas, M. E. Sowa, S. P. Gygi, A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).

Acknowledgments: We thank R. Neumuller, I. T. Flockhart, C. Bergwitz, A. Samsonova, and S. Rajagopal for helpful suggestions for tool development and manuscript preparation. We also thank A. U. Sinha for the help with the iCanPlot software. **Funding:** This work was supported by P01-CA120964, R01-GM067761, and R01-DK088718. S.E.M. is supported in part by the Dana-Farber/Harvard Cancer Center (P30-CA06516). R.S. is supported by the Leukemia and Lymphoma Society. N.P. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** A.V. and Y.H. built the protein complex resources. A.V. developed COMPLEAT and analyzed the RNAi data. C.R. created the Web interface for COMPLEAT. M.K. performed experimental validation. R.S. generated phospho-data. All authors contributed to the manuscript preparation. A.V., S.E.M., and N.P. conceived and supervised the project. **Competing interests:** The authors declare that they have no competing interests.

Submitted 21 September 2012
Accepted 1 February 2013
Final Publication 26 February 2013
10.1126/scisignal.2003629
Citation: A. Vinayagam, Y. Hu, M. Kulkarni, C. Roesel, R. Sopko, S. E. Mohr, N. Perrimon, Protein complex–based analysis framework for high-throughput data sets. *Sci. Signal.* **6**, rs5 (2013).