Sciencexpress Research Article

A Protein Interaction Map of *Drosophila melanogaster*

L. Giot,[1]* J. S. Bader,[1]*† C. Brouwer,[1]* A. Chaudhuri,[1]* B. Kuang,[1] Y. Li,[1] Y. L. Hao,[1] C. E. Ooi,[1] B. Godwin,[1] E. Vitols,[1] G. Vijayadamodar,[1] P. Pochart,[1] H. Machineni,[1] M. Welsh,[1] Y. Kong,[1] B. Zerhusen,[1] R. Malcolm,[1] Z. Varrone,[1] A. Collis,[1] M. Minto,[1] S. Burgess,[1] L. McDaniel,[1] E. Stimpson,[1] F. Spriggs,[1] J. Williams,[1] K. Neurath,[1] N. Ioime,[1] M. Agee,[1] E. Voss,[1] K. Furtak,[1] R. Renzulli,[1] N. Aanensen,[1] S. Carrolla,[1] E. Bickelhaupt,[1] Y. Lazovatsky,[1] A. DaSilva,[1] J. Zhong,[2] C. A. Stanyon,[2] R. L. Finley Jr.,[2] K. P. White,[3] M. Braverman,[1] T. Jarvie,[1] S. Gold,[1] M. Leach,[1] J. Knight,[1] R. A. Shimkets,[1] M. P. McKenna,[1] J. Chant,[1‡] J. M. Rothberg[1]

[1]CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511, USA. [2]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, 540 East Canfield Avenue, Detroit, MI 48201, USA. [3]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA.

*These authors contributed equally to this work.
†Present address: Department of Biomedical Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21215, USA.
‡To whom correspondence should be addressed. E-mail: jchant@curagen.com

Drosophila melanogaster **is a proven model system for many aspects of human biology. Here we present a two-hybrid–based protein-interaction map of the fly proteome. A total of 10,623 predicted transcripts were isolated and screened against standard and normalized complementary DNA libraries to produce a draft map of 7048 proteins and 20,405 interactions. A computational method of rating two-hybrid interaction confidence was developed to refine this draft map to a higher confidence map of 4679 proteins and 4780 interactions. Statistical modeling of the network showed two levels of organization: a short-range organization, presumably corresponding to multiprotein complexes, and a more global organization, presumably corresponding to intercomplex connections. The network recapitulated known pathways, extended pathways, and uncovered previously unknown pathway components. This map serves as a starting point for a systems biology modeling of multicellular organisms including humans.**

Transactions between proteins provide the mechanistic basis for much of the physiology and function of all organisms. Comprehensive analysis of the proteome of any organism presents an extraordinary challenge. The development of genome-scale protein-interaction maps is a powerful first step towards addressing this challenge and provides the framework upon which a systems-biology understanding of cells and organisms can be developed.

Yeast two-hybrid is a facile method that captures a significant fraction of meaningful protein-protein interactions and complexes (*1*). Two-hybrid can be applied in high throughput mode across the entire proteome of an organism to produce a comprehensive protein-protein interaction map (*2*, *3*). Given the value of the *Drosophila* system as a model for human biology, disease, and development, we capitalized upon the recently available *Drosophila* genome sequence and predicted transcriptome (*4*) to build a genome-scale protein interaction map. This map and its analyses are presented here.

Cloning of the transcriptome. To begin building the map, a high throughput effort was mounted to isolate cDNAs representing each predicted transcript of the genome (Figure 1). These efforts employed pooling and full-genome cloning in concert for maximum representation and normalization of the *Drosophila* proteome, with the concomitant drawback of possibly identifying non-biologically-relevant interactions between proteins not simultaneously present *in vivo*. Primers were designed to the 5' and 3' ends of 14,202 open reading frames predicted by release 1 or 2 of the genome sequence (*4*, *5*). The PCR template was a pool of cDNA libraries from embryonic, larval, pupal, and adult stages. PCR product was obtained from 12,278 reactions. These products were cloned into both DNA-binding domain (bait) and DNA-activation domain (prey) two-hybrid vectors (see supplementary methods). Clones whose inserts matched the predicted size, whose 5′ and 3′ ends matched the predicted sequence, and which did not self-activate the reporter system as baits were used further: 11,282 total (9647 both bait and prey; 976 bait only; 659 prey only).

Construction of a draft map. Two strategies were performed for two-hybrid screening. First, individual bait fusions were screened against two cDNA libraries (cDNA screen). Second, individual bait fusions were screened against a pool of the 10,306 preys (collection screen). Screening was

performed by the mating method (*3*) (see supplementary methods).

Following screening, prey sequences were obtained from 63,093 diploid clones. Sequences were matched to predicted transcripts and coding domain sequences from Berkeley *Drosophila* Genome Project Genome Annotation Release 3.1 (see supplementary methods for details). Over 90% of the prey sequences matched at least one transcript or coding sequence. We constructed a locus-based map, rather than a transcript-based map because many sequences could not be unambiguously assigned to splice variants (*e.g.* if an interaction occurred through a domain shared by two or more variants). Prey sequences were mapped successfully for 53,834 diploids, corresponding to unique interactions involving 7048 of the 13,656 protein-coding gene loci. An additional 34 interactions were identified between protein-coding genes and predicted non-protein-coding genes, 33 with RNA genes and 1 with a transposable element. The entire set of 20,439 interactions is available (Supplementary table S7).

Automated confidence scoring of two-hybrid interactions. An important aspect of genome-scale data is the assignment of confidence metrics to data points. To provide a uniform basis for assessing the confidence of two-hybrid or other interaction data types, we developed a systematic statistical approach. Statistical model building incorporated experimental data, which had been stored from screening, and topological criteria, including measures of local clustering (*6*–*8*).

Two training sets were generated for modeling, one by manual annotation and a second by an automated method. Self-interactions were excluded from both. The manual training set was generated as following. An expert biologist reviewed the list of interactions on the basis of the names of the proteins in each interaction pair. High confidence interactions were those published previously and generally accepted to be correct, or those involving two proteins of the same complex. Low confidence interactions were those highly unlikely to occur *in vivo*, such as an apparent interaction between a nuclear and an extracellular protein. High and low confidence assignments were made purely on the basis of the identities of the proteins in each pair, such that statistics from screening could be used to predict interaction confidence.

The automated training set containing both positive and negative examples was generated by comparing the *Drosophila* interactions with physical interactions identified in yeast through a systematic immunoprecipitation-mass-spectroscopy-based approach (*9*, *10*). Positive examples were interacting proteins whose yeast orthologs had reported interactions (Supplementary table S3, S4). Negative examples were *Drosophila* interactions whose yeast orthologs were a distance of 3 or more protein-protein interaction links apart, because pairs of yeast proteins selected at random have a mean distance of 2.8 links. The final positive training set contained 129 examples (70 manual, 65 automated, 6 common to both), and the final negative training set contained 196 examples (88 manual, 112 automated, 4 common to both).

A generalized linear model was fit to the training set using a stepwise procedure to eliminate statistically redundant or non-informative variables. Significant predictors included the number of times each interaction was observed in either the bait/prey or prey/bait orientation, the number of interaction partners of each protein, the local clustering of the network, and the gene region (5′ UTR / CDS / 3′ UTR). While apparent reading frame of the prey relative to the activation domain was a significant predictor on its own, other predictors mask its contribution; retaining reading frame does not improve the final model. The dividing surface between high-confidence and low-confidence was designed to be 0.5 (Fig 2A,B). The fully cross-validated false-positive and false-negative rates for the training set were 16% and 21% (see supplemental methods).

To validate the biological relevance of the statistical model, we examined GO annotations for pairs of interacting proteins binned according to confidence scores (Fig. 2C). The confidence score for an interaction correlates strongly with the depth in the hierarchy that two proteins share an annotation. The correlation increases steeply for confidence scores of 0.5 and higher, supporting the choice of 0.5 as the threshold for high-confidence interactions.

A refined high-confidence map. Applying the statistical model to the entire data set, we obtained a high confidence map of 4780 unique interactions involving 4679 proteins (Fig. 1). The dominant effect of the confidence scores is to remove highly connected proteins whose interactions may be non-specific (Fig. 2A,B): while only 23% of the interactions are retained in the high confidence network, 66% of the proteins are retained. Based on the classification accuracy for the training set, we infer that filtering has effected a 3.4-fold enrichment in the fraction of biologically-relevant interactions in the high-confidence subset, with 40% of the retained interactions likely to be biologically relevant (see supplemental methods). The resulting network consisted of a giant connected cluster (3039 proteins, 3659 interactions) and 565 smaller clusters (2.8 proteins, 2.0 interactions per cluster on average).

The distribution of interactions per protein decays faster than the power law predicted by a “rich-get-richer” model of scale-free networks (the probability that a recently evolved protein establishes a connection to a second protein is proportional to the number of existing interaction partners of the second protein) (Figure 2D). This rapid decay suggests

that highly connected proteins may be suppressed in biological networks, and supports a previous observation that connections between highly-connected proteins are also suppressed (*11*).

Enriched and depleted protein and interaction classes. Proteins were classified according to a reduced set of GO categories (Supplementary table S1) and Pfam domains (release. 8.0). We first identified protein classes significantly enriched or depleted in the high-confidence network (Supplementary table S5). Enriched classes relate primarily to DNA metabolism, transcription, and translation. Depleted classes are primarily plasma membrane proteins, including receptors, ion channels, and peptidases. Enrichment and depletion of specific classes may be due to technical biases of the two-hybrid assay.

We then classified each interaction according to its corresponding pair of protein classes to identify class-pairs that are enriched in the network. Rather than using a contingency table (*12*), we used a randomization method to calculate statistical significance (see supplementary methods). Enriched class-pairs involving structural domains (Pfam annotations) may represent binding modules and could provide the biological rules for building multi-protein complexes. We identified 67 pairs of Pfam domains enriched with a p-value of 0.05 or better after correcting for multiple testing (Supplementary table S6). These include known domain-pairs (F-box/Skp1, p-value = 9×10^{-20}; LIM/LIM-binding, p-value = 5×10^{-8}; actin/cofilin, p-value = 2×10^{-7}) as well as domain-pairs involving domains of unknown function (DUF227/DUF227, p-value = 9×10^{-5}; cullin/DUF298, p-value = 0.0003). An additional 88 domain-pairs are significant at p = 0.05 before correcting for multiple testing and may represent additional biologically relevant binding patterns.

Properties of the high-confidence protein interaction network. Protein networks are of great interest as examples of small world networks (*13–15*). Small world networks exhibit short-range order (two proteins interacting with a third protein have an enhanced probability of interacting with each other) but long-range disorder (two proteins selected at random are likely to be connected by a small number of links, as in a random network).

Small-world properties arise in part from the existence of hub proteins, those having many interaction partners. Hubs are characteristic of scale-free networks, and the Drosophila network resembles a scale-free network in that the distribution of interactions per protein decays slowly, close to a power law (Fig. 2D). To determine the signature of biological organization in small-world properties beyond what would be expected of scale-free networks in general, we calculated properties for both the actual Drosophila network and an ensemble of randomly rewired networks with the same distribution of interactions per protein as the original network. We considered only the giant connected component to avoid ill-defined mathematical quantities.

The distribution of the shortest path between pairs of proteins is peaked at 9-10 protein-protein links (Fig. 3A). A logistic-growth mathematical model for the probability that the shortest path between two distinct proteins has ℓ links is $(N-1)^{-1}K'(\ell;N,J)$, where $K(\ell;N,J)=N/\left[1+(N-1)J^{-\ell}\right]$ is the number of proteins within ℓ links of a central protein and the symbol $'$ indicates differentiation with respect to ℓ, $K'(d;N,J)=N(N-1)(\ln J)J^{-\ell}/\left[1+(N-1)J^{-\ell}\right]^2$. While this model fits the ensemble of random networks, the fit to the actual network is less adequate.

Small-world properties of biological networks may reflect biological organization, and hierarchical organization has been used to describe the properties of metabolic networks (*6*). We tested the ability of a simple, two-level hierarchical model to describe the properties of the Drosphila protein interaction network. The lower level of organization in this model represents protein complexes, and the high level represents interconnections of these complexes. In this case, the probability Pr(ℓ) that the shortest path has ℓ links is

$$\Pr(\ell)=(N_1N_2-1)^{-1}\left[K'(\ell;N_2,J_2)+K'(\ell;N_1,N_2J_1)+\int_0^{\ell}dxK'(x;N_1,N_2J_1)K'(\ell-x;N_2,J_2)\right]$$

where N_1 is the number of cluster, N_2 the number of proteins per cluster, and J_1+J_2 is the number of neighbors per protein, with J_2 within the same cluster and J_1 in other clusters.The 2-level model provides an improved fit to the distance distribution for the observed network, although the improvement is not significant at p = 0.05 (χ^2 decreases by 2.118 with 2 and 19 df, *p*-value = 0.16; see supplementary methods for fitting parameters).

Within multi-protein complexes, enhanced connectivity should yield loops of interacting proteins, which in the network form triangles, squares, pentagons, etc. An excess of loops, a signature of clustering, is observed in the *Drosophila* network (Fig. 3B).

Quantifying the enhancement of loops provides another route to extracting parameters for a hierarchical model of network organization. For the 2-level model in which proteins are organized into N_1 complexes with N_2 proteins per complex, with J_1 between-complex links and J_2 within-complex links per protein, the number of loops is

$$(\#\text{ loops of perimeter } L)=(J_1+J_2)^L/2L+N_1\left(J_2^L/2L\right)\exp\left(-L^2/2N_2\right)$$

Loops are enhanced until the perimeter of the loop is on the order of the square root of the number of proteins in a typical complex. For the actual network, the 2-level model

provides a significantly better fit than the 1-level model ($p = 4.5\times10^{-5}$); for the random network, the fits are indistinguishable ($p = 0.996$).

The 2-level models based on the distribution of shortest paths and the distribution of closed loops give differing estimates of the number of within-complex neighbors per protein (0.8 vs. 2.2), between-complex neighbors per protein (0.1 vs. 0.8), and the number of proteins in each complex (7 vs. 40). This difference arises in part because we employed a continuous model for the shortest path distribution and a discrete model for the loop distribution. The difference may also arise because the shortest path distribution depends on long-range connectivity in the network, the closed loops distribution depends on short-range connectivity, and properties of finite, small-world networks, such as the effective dimensionality, are known to depend on the distance scale measured. Thus, while the evidence for hierarchical organization in the network is highly significant, it may be premature to establish a direct, quantitative connection between parameters of the mathematical model and the composition of real protein complexes.

In summary, the statistical analysis shows that the *Drosophila* network is a small-world network that displays two levels of organization: local connectivity potentially representing interactions occurring within multi-protein complexes and more global connectivity potentially representing higher order communication between complexes.

Global views of the protein interaction map. Two global views of protein interaction network are illustrated: a protein class/human-disease-protein view (Fig. 4A) and a subcellular localization view (Fig. 4B). In both panels, interaction lines are color coded according to predicted confidence score.

Figure 4A is particularly relevant to understanding human disease and potential treatment. In Fig. 4A Protein discs are color-coded according to broad classes of molecular functions as taken from the Gene Ontology annotations (see legend; (*16*)). Many of these classes are suitable targets for the development of small molecule drugs. Drosophila proteins with sequence similarity to human disease proteins are denoted by a star outline (according to the Homophila database; (*17*)). The linkage of proteins altered in human disease to enzyme classes, some of which are druggable, provides insight into the potential development of therapeutics for human diseases such as cancer, heart disease, or diabetes. As shown in Fig. 4A, The homophila gene BCL6 (CG1832), a transcription factor involved in the pathogenesis of human B-cell non-Hodgkin lymphoma [(*17a*)] is connected to calcium-dependent phosphatases CanA1 and Pp2B-14D. CG1832 is connected via the calcium binding protein Eip63F-1. Perhaps BCL6 is regulated in a manner akin to the regulation of NFAT which is dephosphorylated thereby inducing its nuclear translocation [(*17b*)]. The results shown here raise the speculation that therapeutic intervention of calcineurin phosphatases therefore may be an attractive strategy to treat lymphomas and other cancer types. Given the proven utility of Drosophila as a model system, many of the linkages uncovered in this view should be examined for their conservation in human cells.

Figure 4B, a global analysis of protein interaction topology, shows proteins whose sub-cellular localizations are annotated in the Gene Ontology database along with their neighboring proteins. Overall the proteins were laid out according to three broad classes of subcellular localization: nucleus, cytoplasm, and plasma membrane/extracellular space.

Analysis of this subcellular localization view allows the prediction of the subcellular localization, and potential function, of proteins which have not been studied or annotated previously. In Fig. 4B, a local protein interaction network is enlarged which includes, several proteins annotated as nuclear (Srp54, su(w[a]), CG5343, CG11266, CG10689). Highly connected to these are several additional proteins whose localizations are not annotated (CG6843, CG31211, CG14104, CG10324, CG14490, CG14323). Analysis of their sequences using PSORT I & II (http:www.psort.org) indicated that four of the six proteins have >90% probability of being nuclear (CG6843, CG31211, CG14104, CG10324). CG14490 and CG14323 are not necessarily predicted to reside in the nucleus (30% and 10% predicted probabilities). However, they may represent nuclear proteins, which lack detectable signatures of nuclear localization or proteins that shuttle between compartments.

The analysis underlying the figure allows one to query the relative frequencies with which proteins interact with partners from the same or different compartments. The biological expectation is that interactions would be most frequent between proteins within the same compartment with interactions between compartments, which represent inter-compartment communication or protein shuttling, being more rare. As summarized in supplemental table S6 we observe strong enrichment of nuclear-nuclear, cytoplasm-cytoplasm, cytoskeleton-cytoskeleton, and endoplasmic reticulum-endoplasmic reticulum interactions. Inter-compartment interactions (e.g. nucleus-plasma membrane, extracellular-nucleus) tend to be depleted from the data set, consistent with the view that inter-compartment communication is a relatively rare regulatory event. While this global analysis meets with the expectation that interactions within a compartment would be observed more frequently than those between compartments, it is gratifying that this is seen quantitatively in the two-hybrid network generated by high throughput means. The two-hybrid network maintains a signature of cellular topology.

Local pathway views. The refined interaction map provides an opportunity to magnify and examine local interaction networks. Here we present five pathways in detail.

Transcription: Two transcription regulatory circuits involving the well-characterized co-repressors CtBP (c-Terminal Binding Protein) and Gro (groucho) are depicted in Figure 5A. The binding partners of the two co-repressors are largely non-overlapping which concurs with existing evidence that CtBP and Gro repressors independently mediate short and long-range transcriptional repression (*18*). CtBP interacts with a range of transcription factors including homeodomain, nuclear hormone receptor, and C2-H2 Zn-finger proteins, along with the NC2 alpha subunit of the basal transcriptional machinery. Each CtBP interactor has an identifiable variant of the known CtBP interaction motif. Gro interacts with a large group of homeodomain and helix loop helix domain proteins. Gro interactors are known to interact through C-terminal WRPW motifs or the engrailed homology 1 (eh1) domains (*19*, *20*). Each HLH protein shown interacting with Gro (Her, dpn, E(spl), HLHm3, HLHm5, HLHmdelta) possesses a C-terminal WRPW motif. Among the homeodomain interactors, three contain a recognizable eh1 domain (Invected, Unc-4 and Ladybird late [Lbl]). The previously unrecognized Lbl eh1-domain interacting with Gro may provide the basis for the Lbl-mediated repression of target genes, such as even-skipped in the embryonic mesoderm (*21*).

Splicing: Figure 5B shows an extensive network of proteins involved in RNA metabolism. The network captures the regulation of sex determination from X:A ratio to the machinery responsible for the splicing of doublesex and fruitless mRNAs (*22*, *23*). Existing evidence indicates that both Tra-2 and Rbp1 are substrates of Doa kinase (*24*). Our pathway recapitulates known interactions and indicates a pivotal role of Rbp-1 connecting splicing machinery to the upstream components of the sex determination pathway (*25–27*). Three novel proteins (CG14323, CG6843, CG31211) are linked to splicing components through an extensive set of interactions. While these proteins have no recognizable RNA binding motif, the degree of high confidence connectivity with other splicing components suggests that they are complex members. The network also reveals the close association of G-patch domain proteins with splicing factors and RNA-binding proteins. The G-patch domain is a conserved motif found in a variety of eukaryotic RNA processing proteins (*28*, *29*).

Signal transduction: Signal transduction from the membrane to downstream cytoplasmic processes is illustrated in Fig. 5C. The network consists of kinases, adaptor proteins and downstream effectors. Two src isoforms are observed to bind adaptor proteins, drk, Socs36E and CG2079, that dock to phosphotyrosine via SH2/PTB domains and recruit other proteins via their SH3 domains. Within the Sevenless tyrosine kinase pathway, Drk is known to recruit dos. (*30*, *31*), while here drk potentially recruits CG13358 and Nek2 a serine/threonine kinase. A novel adaptor protein CG2079, possessing PTB and PH domains similar to those of the IRS (insulin receptor substrate protein) and DOK (downstream of kinases) family of adaptor proteins, interacts with two Src kinases Src64B and Src42A raising the possibility that CG2079 may link Insulin signaling to Src tyrosine kinases. Two recently identified mammalian proteins IRS5/DOK4 and IRS6/DOK5 bind Src kinases upon phosphorylation by insulin receptor (*32*). Two novel proteins that interact with the bifunctional adaptor proteins in the pathway are CG15022 and CG13358. Inspection of their sequences indicated that they both have poly-proline SH3-binding domains. Further down in the signal transduction pathway, we see recruitment of machinery controlling actin organization and vesicular trafficking.

Calcium regulation: Calcium regulates diverse signaling pathways by binding calmodulin and other calcium-binding proteins. Calmodulin and related proteins in turn transduce signal via effector proteins such as kinases and phosphatases. Figure 5D illustrates a network of calmodulins (Cam and And), novel calmodulin-like proteins (CG11165, CG31958, CG11638), calcium-binding proteins, and the calcineurin family of calmodulin-dependent ser/thr phosphatases. Two cell surface ion-channels inx2 and KCNQ interact with calmodulin proteins. Although regulation of inx2 and KCNQ by calcium has not been reported in Drosophila, their mammalian counterparts (connexin and KCNQ) are regulated by calcium (*33–35*). A potentially significant link between the tyrosine transporter hoe1 and two calcineurin phosphatases is shown. Mutation in human homolog of hoe1 causes ocular albinism.

Cell cycle regulation: Fig 5E shows the network surrounding the Skp protein complex (SCF complex) that targets proteins to ubiquitin-mediated proteasomal degradation (*36*). Target proteins are recruited to the Skp complex by F-box proteins (*37–39*). Among the Skp proteins, only SkpA is reported in the literature to bind F-box proteins (*40*). Two F-box proteins Morgue and Slmb interact with SkpA in the pathway. Morgue associates with SkpA to mediate the ubiquitination of DIAP1 and target its degradation (*41*). Other significant target proteins in the pathway include Rca1, CG9790 (CDK regulator) and skl (sickle). Rca1 is known to regulate the level of cyclin-A during the cell cycle (*42*) and is reported to be an inhibitor of the anaphase-promoting complex (APC) (*43*). CG9790 gene is homologous to the CDK regulatory protein, Cks. Human Cks-1 is an accessory protein of the SCF complex required for ubiquitin ligation of the CDK inhibitor p27 (*44*, *45*). The Sickle (skl) protein is a recently described novel DIAP-

binding protein that induces apoptosis (*46*, *47*). The presence of skl in the Skp complex suggests that, like Morgue, it may target DIAP1 to degradation by the SCF complex. As shown in Fig. 5D, skl protein also interacts with several calmodulin-binding proteins (Fig. 5C). It is tempting to speculate that skl may regulate the half-life of these proteins as well. This network suggests that target proteins may also be recruited to the Skp complex via Skp-dimerization domain containing proteins and RNI domain proteins. Of the five RNI domain proteins in the network, the function of ppa in targeting the transcription factor paired to degradation has been reported (*48*). It is suggested that RNI domain proteins may function as accessory proteins of the SCF complex.

Diverse pathway examples: In Fig 5F we present a collage of 10 diverse networks from the dataset. Three of these pathways are described here with the others described in supplementary materials.

Innate immunity: The Imd pathway is a well-characterized Drosophila-signaling complex involved in innate immune response against gram-negative bacteria (*49*). The Imd-BG4-Dredd protein complex activates the transcription factor Relish by proteolytic cleavage. Their human orthologs RIP-FADD-Caspase-8, bind each other in the same order, suggesting that the organization of the two signaling complex is evolutionarily conserved. These components are connected intimately to the protein ubiquitination machinery via ari-2 and the E2 class of ubiquitin ligases (Ubc84D and UbcD10). A recent study has reported that ubiquitin pathway represses IMD signaling (*50*) by targeting the transcription factor Relish. Our findings suggest that the ubiquitin machinery may target the upstream components of the signaling complex as well.

EGF receptor localization: The Egf-veli-skf complex is similar to the well-characterized *C. elegans* protein complex of LET-23- LIN-2-LIN-7 involved in the polarized localization of the LET-23 receptor (EGF receptor) during vulval development (*51*). The veli protein is a PDZ domain protein (similar to Lin-7) that brings together the receptor and the skf protein. The latter is a guanylate kinase containing PDZ and SH3 domains (similar to Lin-2). Veli protein has been suggested to function in Drosophila nervous system. However, our pathway suggests the existence of a conserved protein complex that functions in EGF receptor localization.

Photoreceptor differentiation: The protein complex associated with Sina functions in Drosophila photoreceptor differentiation by down regulating the transcription repressor ttk (tramtrack) in a subset of photoreceptor cells in response to RAS signaling (*52*, *53*). Our pathway shows that the adaptor protein phyl (phyllopod) brings together Sina (E3 ligase) and ttk, resulting in the ubiquitination and degradation of the repressor protein. A recent biochemical analysis has identified two separate domains in phyl that bind Sina and ttk (*54*). A novel interactor in our pathway is rin (rasputin), a RasGAP protein that functions in eye development as a regulator of RAS signaling (*55*). In addition our pathway suggests a novel function of a yet uncharacterized Drosophila protein CG13030. The protein shares 45% amino acid identity to Sina, with a ring finger domain that is similar in organization to the Sina ring finger domain (C3HC4 type). Significantly, both the proteins share the same binding partners. Taking together, the results of pathway analysis and the domain organization of both proteins suggest that CG13030 may overlap in function with Sina protein as a novel regulator of photoreceptor differentiation.

The genome scale network introduced here of course contains numerous additional local networks that should prove valuable to the community at large. Our intent is for this map to serve as a public resource for interested scientists. We have deposited these interactions with FlyBase, BIND, and DIP (*56*).

References and Notes

1. E. Phizicky, P. I. Bastiaens, H. Zhu, M. Snyder, S. Fields, *Nature* **422**, 208 (2003).
2. T. Ito et al., *Proc Natl Acad Sci U S A* **98**, 4569 (2001).
3. P. Uetz et al., *Nature* **403**, 623 (2000).
4. M. D. Adams et al., *Science* **287**, 2185 (2000).
5. G. M. Rubin et al., *Science* **287**, 2222 (2000).
6. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabasi, *Science* **297**, 1551 (2002).
7. D. S. Goldberg, F. P. Roth, *Proc Natl Acad Sci U S A* **100**, 4372 (2003).
8. J. S. Bader, A. Chaudhuri, J. Chant, *Nature Biotechnology* (2003) in press.
9. A. C. Gavin et al., *Nature* **415**, 141 (2002).
10. Y. Ho et al., *Nature* **415**, 180 (2002).
11. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
12. E. Sprinzak, H. Margalit, *J Mol Biol* **311**, 681 (2001).
13. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
14. H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).
15. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
16 M. Ashburner et al., *Nat Genet* **25**, 25-9. (2000).
17. L. T. Reiter, L. Potocki, S. Chien, M. Gribskov, E. Bier, *Genome Res* **11**, 1114-25. (2001).
17a. B. W. Baron *et al*., *Proc Natl Acad Sci U S A* **90,** 5262 (1993).
17b. P. G. Hogan, L. Chen, J. Nardone, A. Rao, *Genes Dev* **17,** 2205 (2003).
18. A. J. Courey, S. Jia, *Genes Dev* **15**, 2786 (2001).
19. G. Jimenez, C. P. Verrijzer, D. Ish-Horowicz, *Mol Cell Biol* **19**, 2080 (1999).
20. G. Jimenez, A. Guichet, A. Ephrussi, J. Casanova, *Genes Dev* **14**, 224 (2000).

21. K. Jagla, M. Bellard, M. Frasch, *Bioessays* **23**, 125 (2001).
22. P. Graham, J. K. Penn, P. Schedl, *Bioessay*s **25**, 1 (2003).
23. B. R. Graveley, *Cell* **109**, 409 (2002).
24. C. Du, M. E. McGuffin, B. Dauwalder, L. Rabinow, W. Mattox, *Mol Cell* **2**, 741 (1998).
25. Y. J. Kim, P. Zuo, J. L. Manley, B. S. Baker, *Genes Dev* **6**, 2569 (1992).
26. K. W. Lynch, T. Maniatis, *Genes Dev* **10**, 2089 (1996).
27. V. Heinrichs, B. S. Baker, *Proc Natl Acad Sci U S A* **94**, 115 (1997).
28. L. Aravind, E. V. Koonin, *Trends Biochem Sci* **24**, 342 (1999).
29. B. Guglielmi, M. Werner, *J Biol Chem* **277**, 35712 (2002).
30. S. M. Feller, H. Wecklein, M. Lewitzky, E. Kibler, T. Raabe, *Mech Dev* **116**, 129 (2002).
31. J. P. Olivier et al., *Cell* **73**, 179 (1993).
32. D. Cai, S. Dhe-Paganon, P. A. Melendez, J. Lee, S. E. Shoelson, J *Biol Chem* **278**, 25323 (2003).
33. E. Yus-Najera, I. Santana-Castro, A. Villarroel, *J Biol Chem* **277**, 28545 (2002).
34. A. Sotkis et al., *Cell Commun Adhes* **8**, 277 (2001).
35. C. Peracchia, A. Sotkis, X. G. Wang, L. L. Peracchia, A. Persechini, *J Biol Chem* **275**, 26220 (2000).
36. R. J. Deshaies, *Annu Rev Cell Dev Biol* **15**, 435 (1999).
37. R. M. Feldman, C. C. Correll, K. B. Kaplan, R. J. Deshaies, *Cell* **91**, 221 (1997).
38. E. T. Kipreos, M. Pagano, *Genome Biol* **1**, REVIEWS3002 (2000).
39. D. Skowyra, K. L. Craig, M. Tyers, S. J. Elledge, J. W. Harper, *Cell* **91**, 209 (1997).
40. C. Bai et al., *Cell* **86**, 263 (1996).
41. J. P. Wing et al., *Nat Cell Biol* **4**, 451 (2002).
42. X. Dong et al., *Genes Dev* **11**, 94 (1997).
43. R. Grosskortenhaus, F. Sprenger, *Dev Cell* **2,** 29 (2002).
44. D. Ganoth et al., *Nat Cell Biol* **3**, 321 (2001).
45. D. Sitry et al., *J Biol Chem* **277**, 42233 (2002).
46. A. Christich et al., *Curr Biol* **12**, 137 (2002).
47. J. P. Wing et al., *Curr Biol* **12**, 131 (2002).
48. L. Raj et al., *Curr Biol* **10**, 1265 (2000).
49. J. A. Hoffmann, J. M. Reichhart, *Nat Immunol* **3**, 121 (2002).
50. R. S. Khush, W. D. Cornwell, J. N. Uram, B. Lemaitre, *Curr Biol* **12**, 1728 (2002).
51. S. M. Kaech, C. W. Whitfield, S. K. Kim, *Cell* **94**, 761 (1998).
52. Z. C. Lai, S. D. Harrison, F. Karim, Y. Li, G. M. Rubin, *Proc Natl Acad Sci USA* **93**, 5025 (1996).
53. S. Li, R. W. Carthew, Z. C. Lai, *Cell* **90**, 469 (1997).
54. S. Li, C. Xu, R. W. Carthew, *Mol Cell Biol* **22**, 6854 (2002)
55. C. Pazman, C. A. Mayes, M. Fanto, S. R. Haynes, M. Mlodzik *Development* **127**, 1715 (2000).
56. Accession numbers are as follows: BIND IDs 24880 to 45318.
57. We thank our colleagues at CuraGen, in particular those in the genomics facility who performed all the sequencing and prepared the cDNA libraries described in this work. We also thank S. Hossain and members of the Finley laboratory for technical assistance with Drosophila cultures and RNA preparations. J.Z., C.A.S, and R.L.F. were supported by NIH grant HG01536 (to R.L.F.).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1090289/DC1
Materials and Methods
Tables S1 to S7

11 August 2003; accepted 29 October 2003
Published online 6 November 2003;
10.1126/science.1090289
Include this information when citing this paper.

Fig. 1. Flow diagram illustrating the process utilized to generate the genome-scale protein interaction map (See text for details).

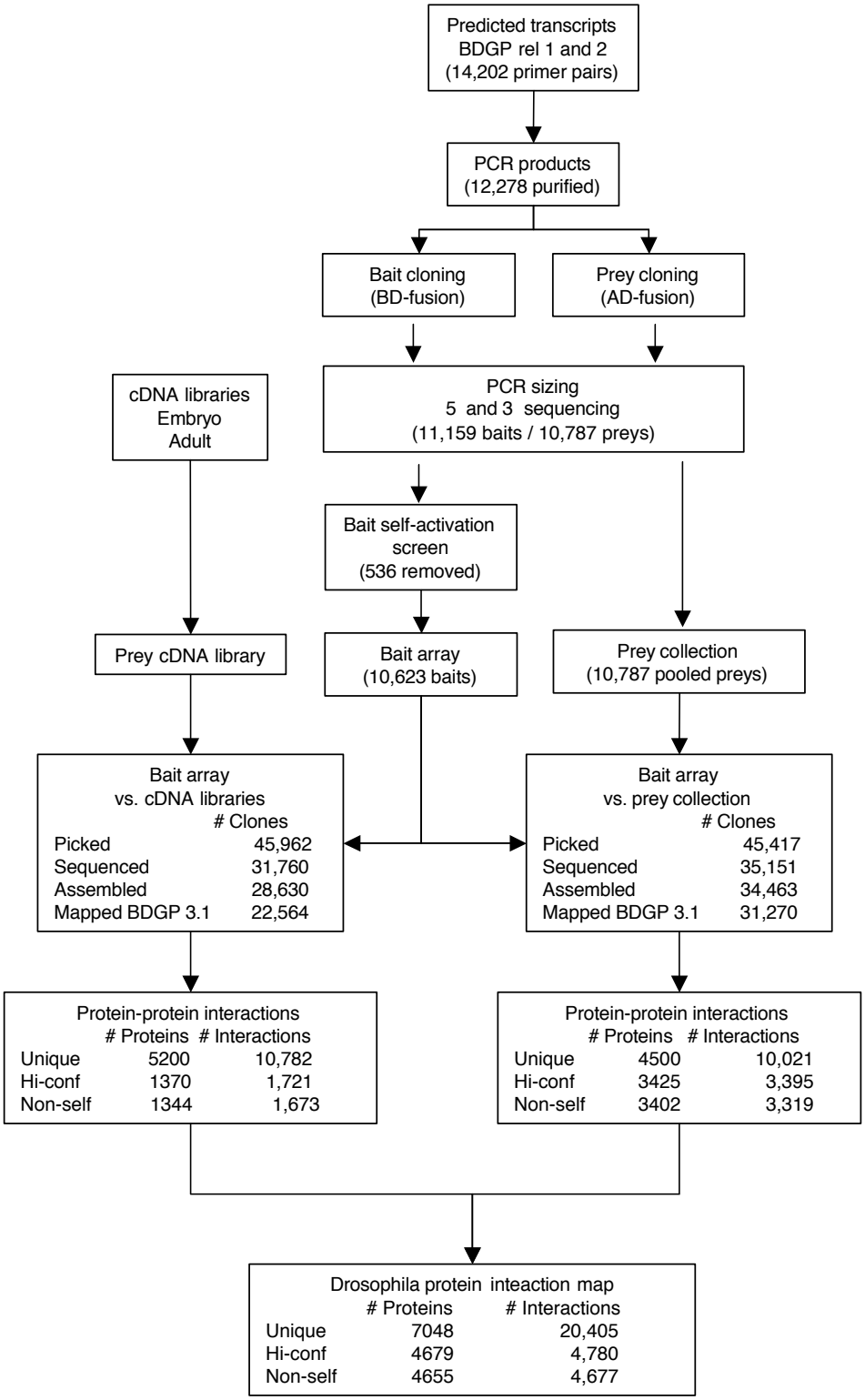
Fig. 2. Confidence scores for protein-protein interactions (**A**) *Drosophila* protein-protein interactions have been binned according to confidence score for the entire set of 20,405 interactions (black), the 129 positive training set examples (green), and the 196 negative training set examples (red). (**B**) The 7048 proteins identified as participating in protein-protein interactions have been binned according to the minimum, average, and maximum confidence score of their interactions. Proteins with high-confidence interactions total 4679 (66% of the proteins in the network, and 34% of the protein-coding genes in the *Drosophila* genome). (**C**) The correlation between GO annotations for interacting protein pairs decays sharply as confidence falls from 1 to 0.5, then exhibits a weaker decay. Correlations were obtained by first calculating the deepest level in the GO hierarchy at which a pair of interacting proteins shared an annotated (interactions involving unannotated proteins were discarded). The average depth was calculated for interactions binned according to confidence score, with bin centers at 0.05, 0.1, ..., 0.95. Finally, the correlation for the bin centered at x was defined as $[Depth(x)–Depth(0)] / [Depth(0.95)–Depth(0)]$. This procedure effectively controls for the depth of each hierarchy, and for the probability that a pair of random proteins shares an annotation (**D**) The number of interactions per protein is shown for all interactions (black circles) and for the high-confidence interactions (green circles). Linear behavior in this log-log plot would indicate a power-law distribution.

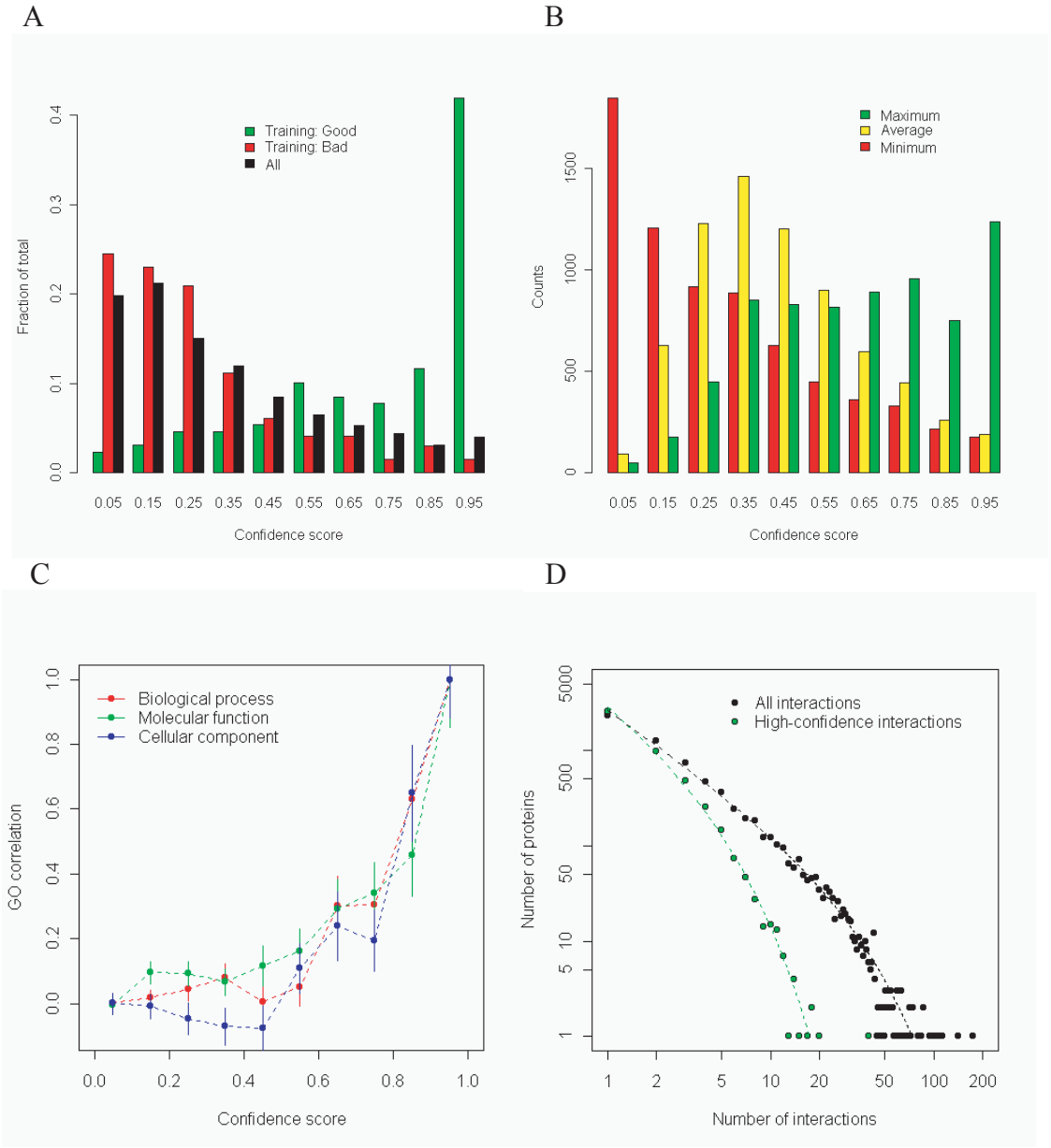
Although regions of each distribution appear linear, neither distribution may be adequately fit by a single power-law. Both may be fit, however, by a combination of power-law and exponential decay, Prob(n) ~ $n^{-\alpha}\exp^{-\beta n}$, indicated by the dashed lines, with r^2 for the fit greater than 0.98 in either case (all interactions: $\alpha = 1.20\pm0.08$, $\beta = 0.038\pm0.006$; high-confidence interactions: $\alpha = 1.26\pm0.25$, $\beta = 0.27\pm0.05$). Note that the power-law exponents are within 1-σ for the two interaction sets.

Fig. 3. Statistical properties of the refined *Drosophila* interaction map. The high-confidence *Drosophila* protein-protein interactions form a small-world network with evidence for a hierarchy of organization. Network properties are presented for the giant connected component, in which 3659 pairwise interactions connect 3039 proteins into a single cluster (see text for details). (**A**) The probability distribution for the shortest path between a pair of proteins the actual network (green points) is peaked at 9–11 links, with a mean of 9.4 links. In contrast, an ensemble of randomly rewired networks shows a mean separation of 7.7 links between proteins. Biological organization may be responsible for flattening the actual network by enhancing links between proteins that are already close. (**B**) Clustering, or enhancement of connections between proteins that are already close, is analyzed quantitatively by counting the number of closed loops (triangles, squares, pentagons, etc) in which the perimeter is formed by a series of proteins connected head-to-tail, with no protein repeated. The actual network (green points) shows an enhancement of loops with perimeter up to 10–11 relative to the random network (red points). In both (A) and (B), the 1-level and 2-level models produce nearly indistinguishable fits for the random networks, indicating the absence of structured clustering.

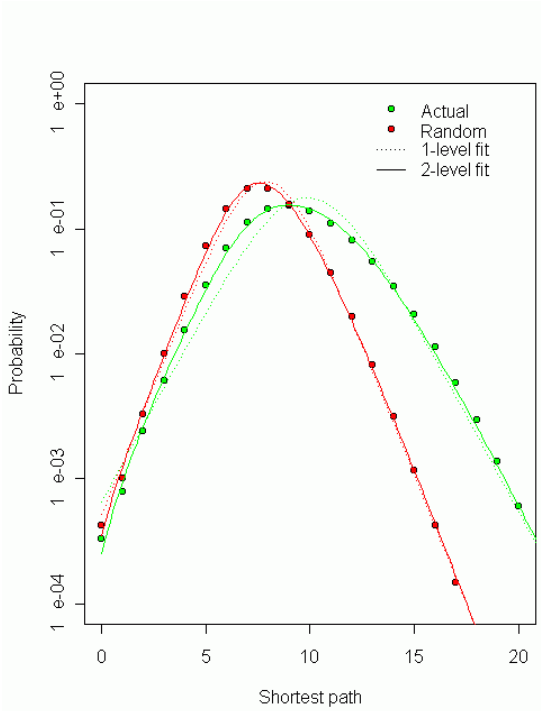
Fig. 4. Global views of the protein interaction map. (**A**) Protein family/human disease ortholog view. Proteins are color-coded according to protein family as annotated by the Gene Ontology hierarchy. Proteins orthologous to human disease proteins have a jagged starry border. Interactions were sorted according to interaction confidence score and the top 3000 interactions are shown with their corresponding 3522 proteins. This corresponds roughly to a confidence score of 0.62 and higher. (**B**) Subcellular localization view. This view shows the fly interaction map with each protein colored by its Gene Ontology Cellular Component annotation. This map has been filtered by only showing proteins with less than or equal to 20 interactions and with at least one Gene Ontology annotation (not necessarily a cellular component annotation). We show proteins for all interactions with a confidence score of 0.5 or higher. This results in a map with 2346 proteins and 2268 interactions.

Fig. 5. Local pathway views. (**A**) Regulation of transcription repression by Groucho and CtBP proteins. (**B**) Splicing complex associated with sex determination. (**C**) Signaling complex linking Src kinases with downstream effectors via adaptor proteins. (**D**) Regulation of surface transporters and channels by Calcium signaling. (**E**) Drosophila Skp pathway. (**F**) Examples of local pathway views identified in the interaction network.

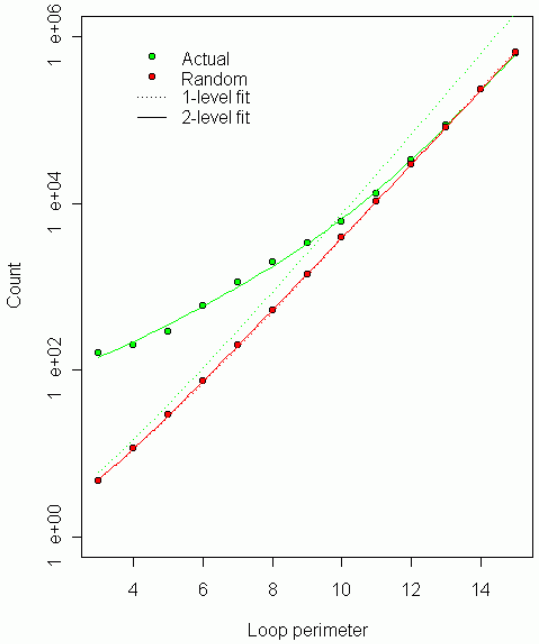

Predicted transcripts
BDGP rel 1 and 2
(14,202 primer pairs)
PCR products
(12,278 purified)
Bait cloning
(BD-fusion)
Prey cloning
(AD-fusion)
cDNA libraries
Embryo
Adult
PCR sizing
5 and 3 sequencing
(11,159 baits / 10,787 preys)
Bait self-activation
screen
(536 removed)
Prey cDNA library
Bait array
(10,623 baits)
Prey collection
(10,787 pooled preys)
Bait array
vs. cDNA libraries
Clones
Picked 45,962
Sequenced 31,760
Assembled 28,630
Mapped BDGP 3.1 22,564
Bait array
vs. prey collection
Clones
Picked 45,417
Sequenced 35,151
Assembled 34,463
Mapped BDGP 3.1 31,270
Protein-protein interactions
Proteins # Interactions
Unique 5200 10,782
Hi-conf 1370 1,721
Non-self 1344 1,673
Protein-protein interactions
Proteins # Interactions
Unique 4500 10,021
Hi-conf 3425 3,395
Non-self 3402 3,319
Drosophila protein inteaction map
Proteins # Interactions
Unique 7048 20,405
Hi-conf 4679 4,780
Non-self 4655 4,677

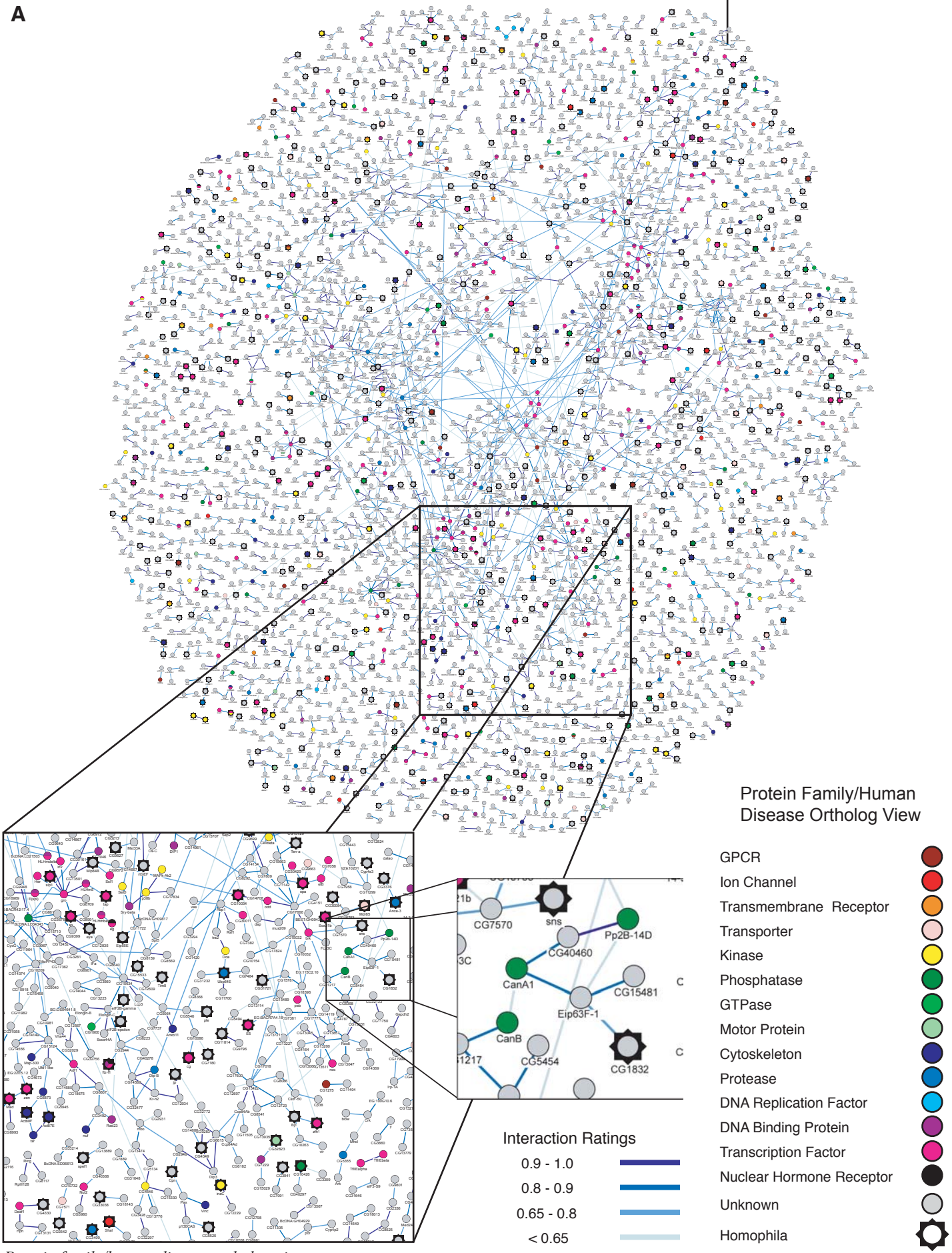

A
B
C
D
Training: Good
Training: Bad
All
Fraction of total
Confidence score
Maximum
Average
Minimum
Counts
Confidence score
Biological process
Molecular function
Cellular component
GO correlation
Confidence score
All interactions
High-confidence interactions
Number of proteins
Number of interactions

A


Actual
Random
1-level fit
2-level fit
Probability
1 e+00
1 e-01
1 e-02
1 e-03
1 e-04
0
5
10
15
20
Shortest path

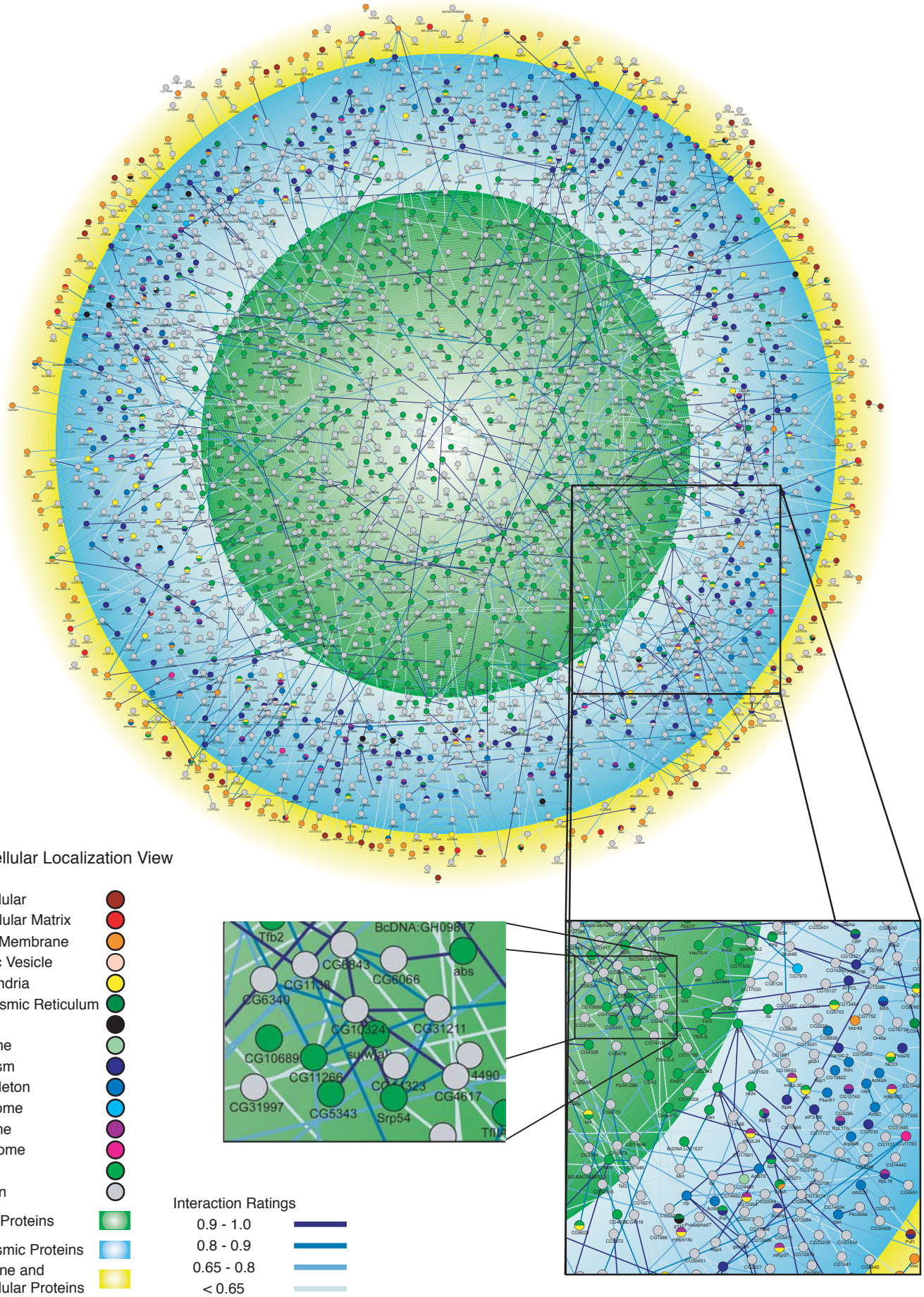
B

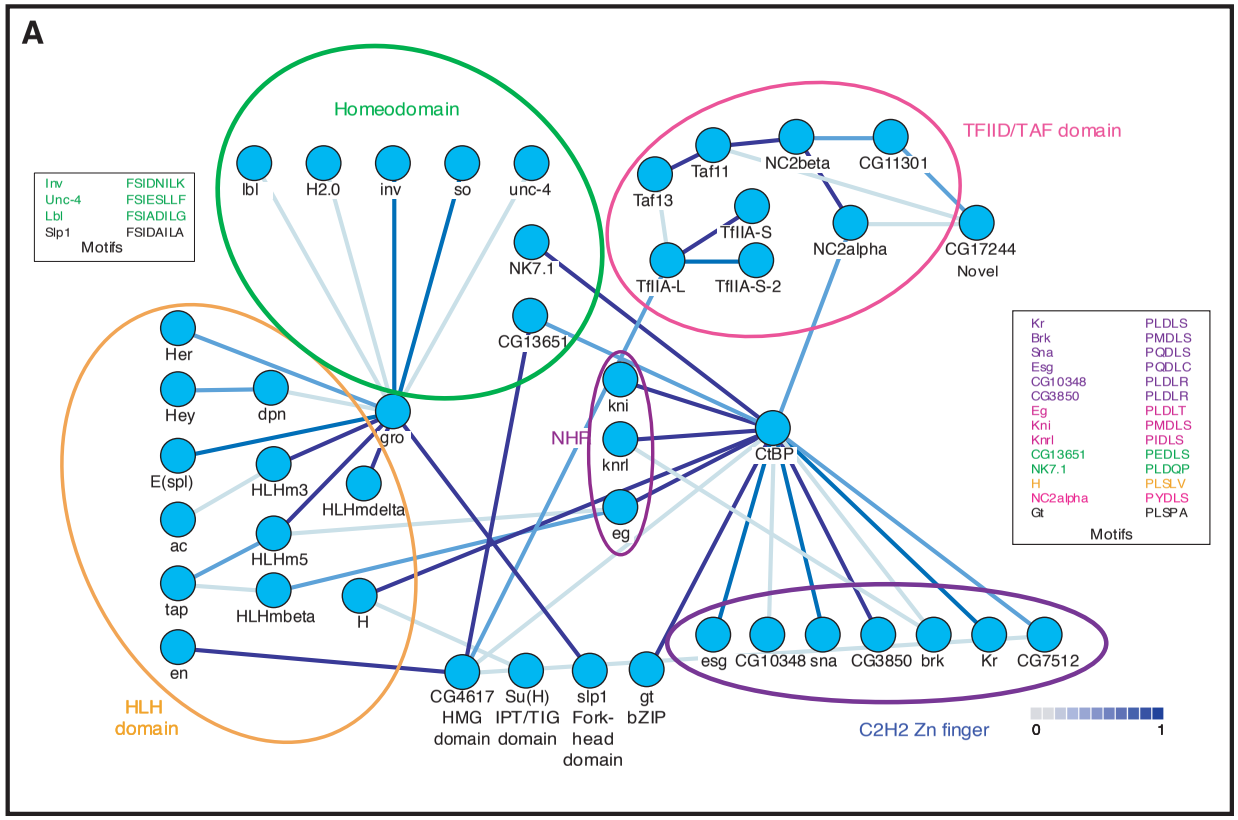

Actual
Random
1-level fit
2-level fit
Count
1 e+06
1 e+04
1 e+02
1 e+00
4
6
8
10
12
14
Loop perimeter

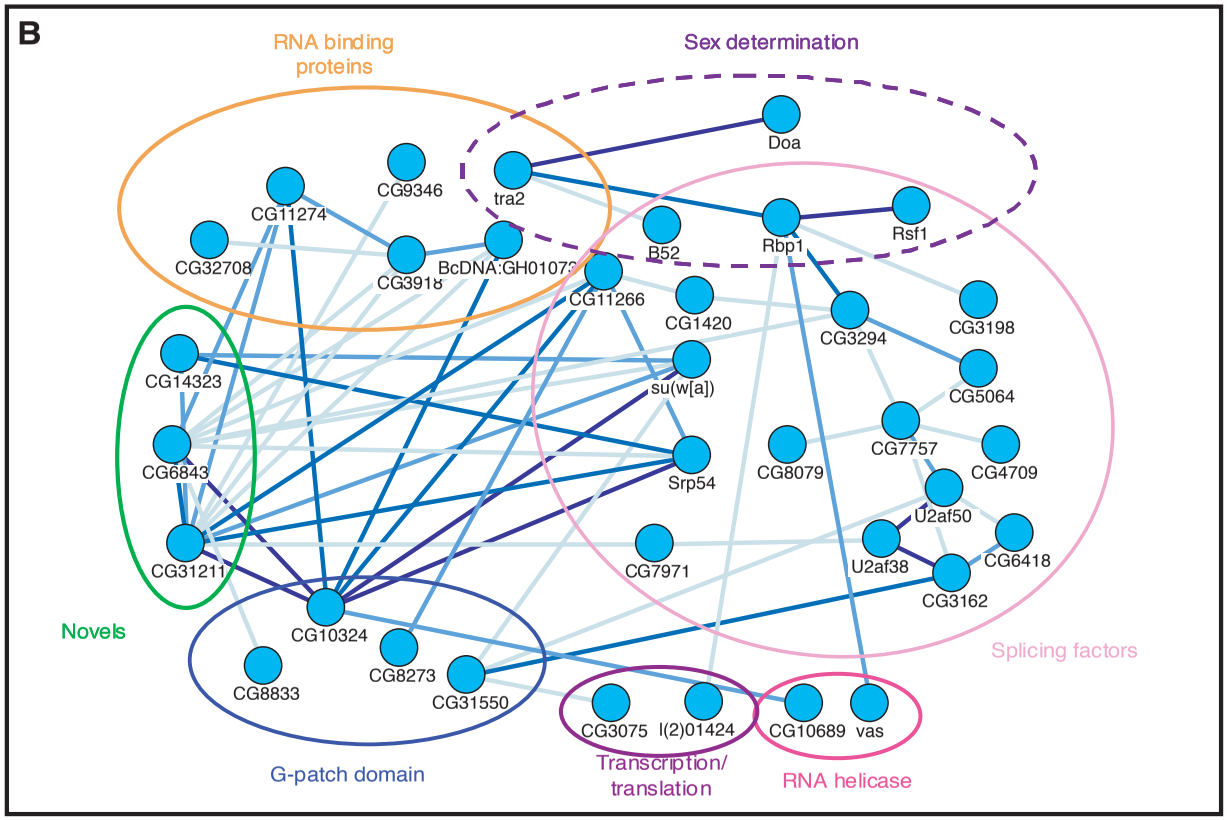


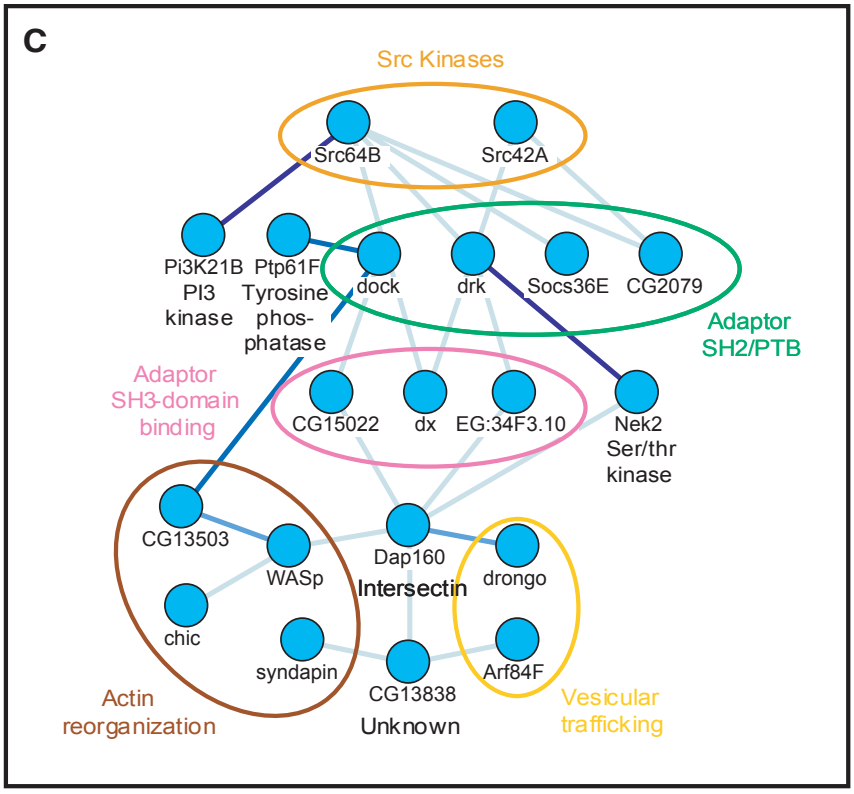
Protein family/human disease ortholog view.

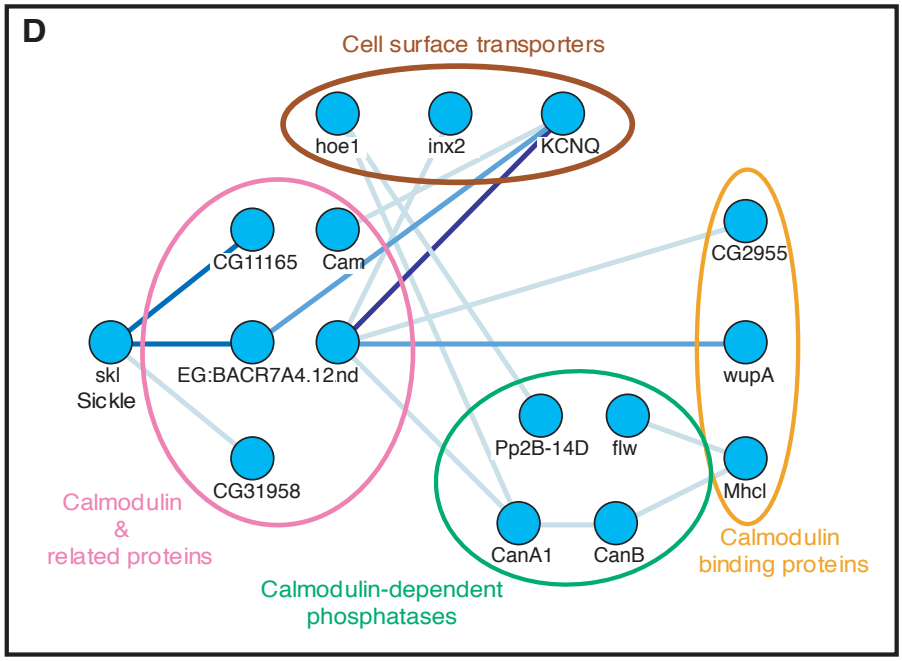
B

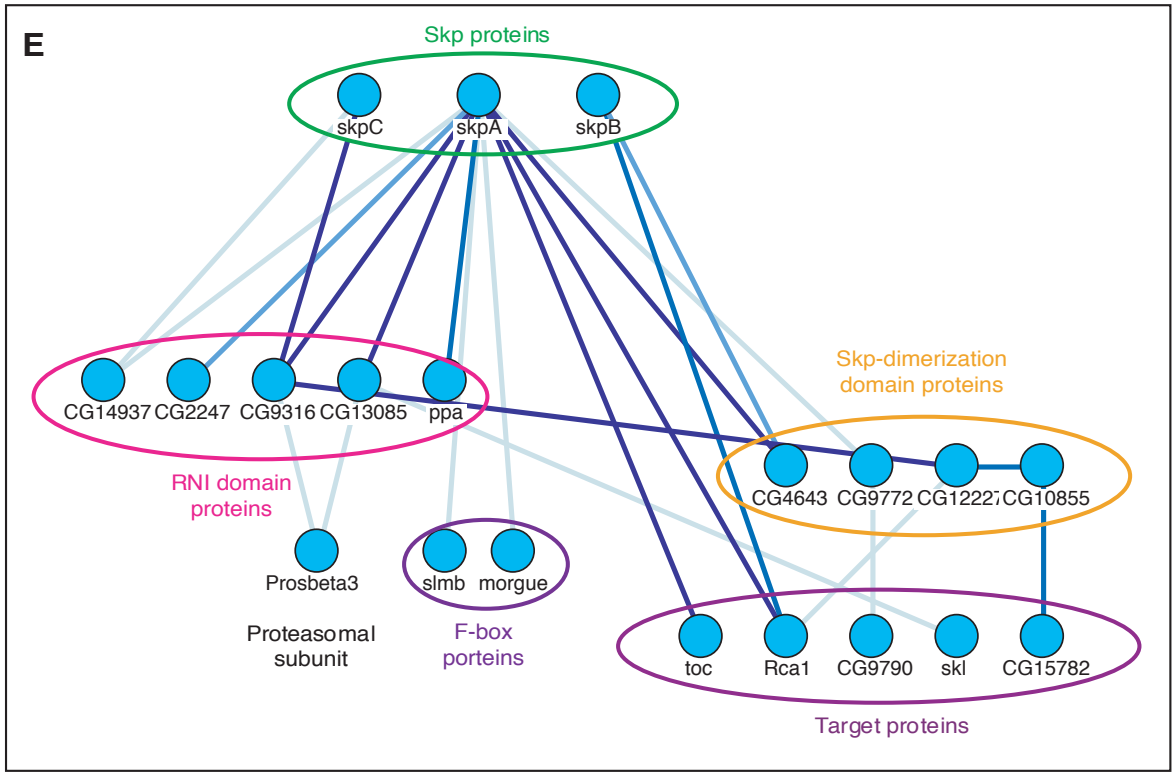

Sub-Cellular Localization View
Extracellular
Extracellular Matrix
Plasma Membrane
Synaptic Vesicle
Mitochondria
Endoplasmic Reticulum
Golgi
Lysosome
Cytoplasm
Cytoskeleton
Peroxisome
Ribosome
Centrosome
Nucleus
Unknown
Nuclear Proteins
Cytoplasmic Proteins
Membrane and Extracellular Proteins
Interaction Ratings
0.9 - 1.0
0.8 - 0.9
0.65 - 0.8
< 0.65
Tfb2
BcDNA:GH09817
CG6843
CG6066
abs
CG1138
CG6340
CG10324
CG31211
CG10689
CG11266
CG31997
CG5343
Srp54
CG4617
4490

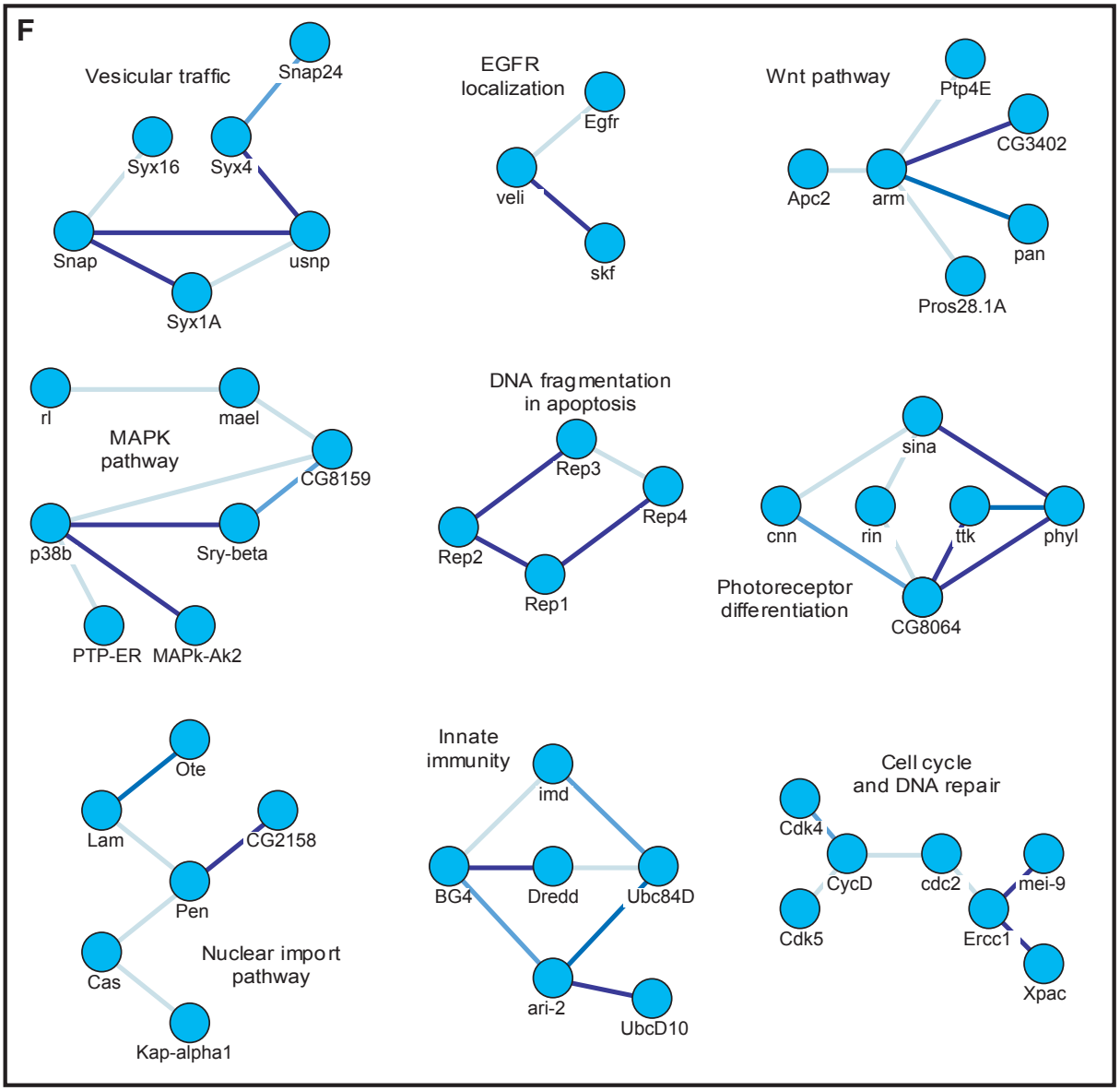

A
Homeodomain
TFIID/TAF domain
Inv FSIDNILK
Unc-4 FSIESLLF
Lbl FSIADILG
Slp1 FSIDAILA
Motifs
lbl
H2.0
inv
so
unc-4
NK7.1
CG13651
Taf11
NC2beta
CG11301
Taf13
TfIIA-S
NC2alpha
CG17244
Novel
TfIIA-L
TfIIA-S-2
Her
Hey
dpn
gro
E(spl)
HLHm3
HLHmdelta
ac
HLHm5
tap
HLHmbeta
H
en
NHF
kni
knrl
eg
CtBP
Kr PLDLS
Brk PMDLS
Sna PQDLS
Esg PQDLC
CG10348 PLDLR
CG3850 PLDLR
Eg PLDLT
Kni PMDLS
Knrl PIDLS
CG13651 PEDLS
NK7.1 PLDQP
H PLSLV
NC2alpha PYDLS
Gt PLSPA
Motifs
esg
CG10348
sna
CG3850
brk
Kr
CG7512
CG4617 HMG domain
Su(H) IPT/TIG domain
slp1 Fork-head domain
gt bZIP
HLH domain
C2H2 Zn finger
0
1


B
RNA binding proteins
Sex determination
Doa
tra2
CG9346
CG11274
CG32708
CG3918
BcDNA:GH01073
B52
Rbp1
Rsf1
CG11266
CG1420
CG3294
CG3198
CG14323
su(w[a])
CG5064
CG6843
Srp54
CG8079
CG7757
CG4709
U2af50
CG31211
CG7971
U2af38
CG6418
CG3162
Novels
CG10324
CG8273
CG8833
CG31550
Splicing factors
CG3075
l(2)01424
CG10689
vas
G-patch domain
Transcription/ translation
RNA helicase


C
Src Kinases
Src64B
Src42A
Pi3K21B
PI3
kinase
Ptp61F
Tyrosine
phos-
phatase
dock
drk
Socs36E
CG2079
Adaptor
SH2/PTB
Adaptor
SH3-domain
binding
CG15022
dx
EG:34F3.10
Nek2
Ser/thr
kinase
CG13503
WASp
chic
syndapin
Dap160
Intersectin
drongo
Arf84F
CG13838
Unknown
Actin
reorganization
Vesicular
trafficking


D
Cell surface transporters
hoe1
inx2
KCNQ
CG11165
Cam
CG2955
skl
Sickle
EG:BACR7A4.12nd
wupA
Pp2B-14D
flw
CG31958
Mhcl
CanA1
CanB
Calmodulin
&
related proteins
Calmodulin-dependent
phosphatases
Calmodulin
binding proteins


E
Skp proteins
skpC
skpA
skpB
CG14937
CG2247
CG9316
CG13085
ppa
RNI domain proteins
Skp-dimerization domain proteins
CG4643
CG9772
CG12227
CG10855
Prosbeta3
Proteasomal subunit
slmb
morgue
F-box porteins
toc
Rca1
CG9790
skl
CG15782
Target proteins


F
Vesicular traffic
Snap24
Syx16
Syx4
Snap
usnp
Syx1A
EGFR
localization
Egfr
veli
skf
Wnt pathway
Ptp4E
CG3402
Apc2
arm
pan
Pros28.1A
rl
mael
MAPK
pathway
CG8159
p38b
Sry-beta
PTP-ER
MAPk-Ak2
DNA fragmentation
in apoptosis
Rep3
Rep4
Rep2
Rep1
sina
cnn
rin
ttk
phyl
Photoreceptor
differentiation
CG8064
Ote
Lam
CG2158
Pen
Cas
Nuclear import
pathway
Kap-alpha1
Innate
immunity
imd
BG4
Dredd
Ubc84D
ari-2
UbcD10
Cell cycle
and DNA repair
Cdk4
CycD
cdc2
mei-9
Cdk5
Ercc1
Xpac