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]

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 toward 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 sizable 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 on 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, we mounted a high-throughput effort to isolate cDNAs representing each predicted transcript of the genome (Fig. 1). These efforts used pooling and full-genome cloning in concert for maximum representation and normalization of the *Drosophila* proteome, with the concomitant drawback of possibly identifying nonbiologically 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 polymerase chain reaction (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 (*6*). 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 used 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, 6*).

After 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 (*6*). More than 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 (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 (*7–9*).

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 follows: 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

[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

interaction between a nuclear and an extra-cellular 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 (*10*, *11*). Positive examples were interacting proteins whose yeast orthologs had reported interactions (tables S3 and S4). Negative exam-

ples 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 with a stepwise procedure to eliminate statistically redundant or noninformative 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 (UTR, untranslated region; CDS, coding sequence)]. Although the 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 interactions was designed to be 0.5 (Fig. 2, A and B). The fully cross-validated false-positive and false-negative rates for the training set were 16% and 21% (*6*).

To validate the biological relevance of the statistical model, we examined Gene Ontology (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 at which 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 nonspecific (Fig. 2, A and B): Whereas 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, such that 40% of the retained interactions are likely to be biologically relevant (*6*). 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) (Fig. 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 (*12*).

Enriched and depleted protein and interaction classes. Proteins were classified according to a reduced set of GO categories (table S1) and Pfam domains (release. 8.0).
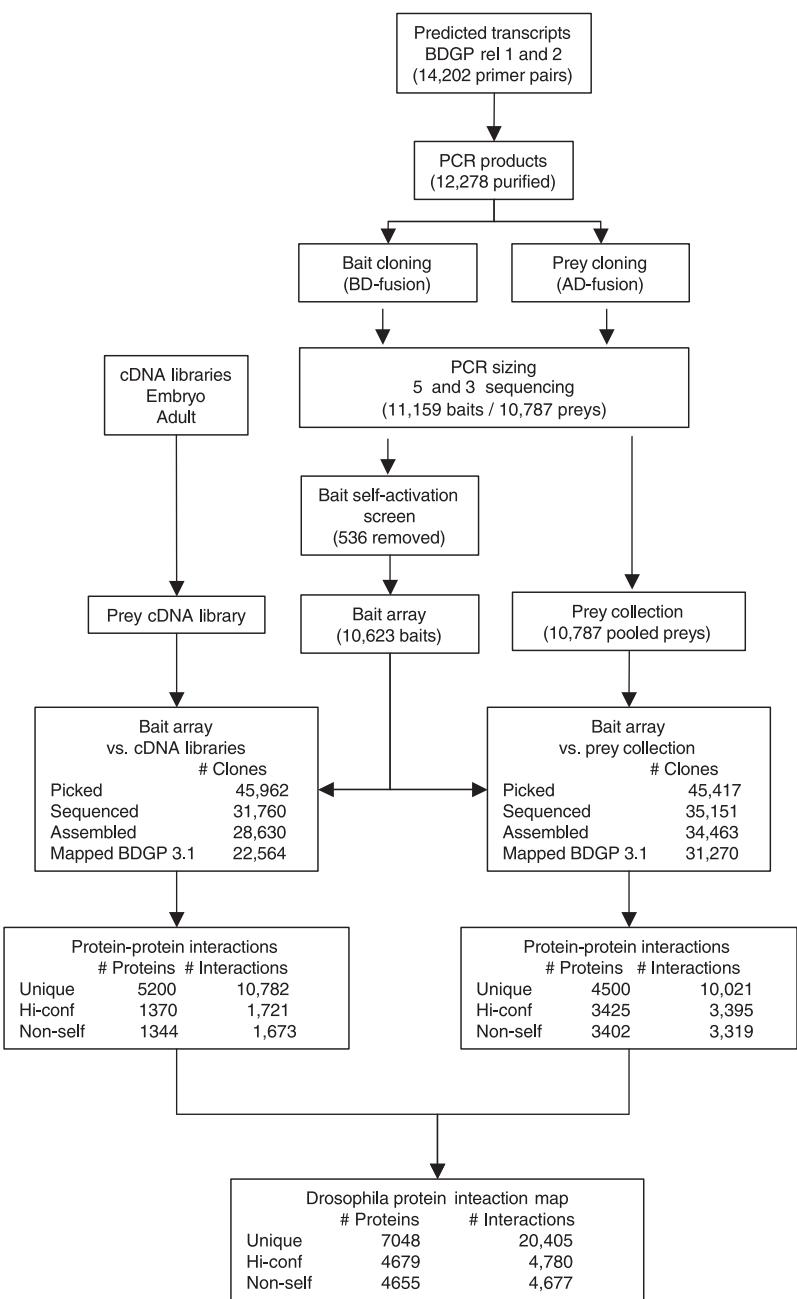

Fig. 1. Flow diagram illustrating the process used to generate the genome-scale protein-interaction map (see text for details).

We first identified protein classes significantly enriched or depleted in the high-confidence network (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 (*13*), we used a randomization method to calculate statistical significance (*6*). Enriched class-pairs involving structural domains (Pfam annotations) may represent binding modules and could provide the biological rules for building multiprotein complexes. We identified 67 pairs of Pfam domains enriched with a *P* value of 0.05 or better after correcting for multiple testing (table S6). These include known domain pairs (F-box/Skp1, $P = 9 \times 10^{-20}$; LIM/LIM binding, $P = 5 \times 10^{-8}$; actin/cofilin, $P = 2 \times 10^{-7}$) as well as domain pairs involving domains of unknown function (DUF227/DUF227, $P = 9 \times 10^{-5}$; cullin/DUF298, $P = 0.0003$). 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 (*14–16*). 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 in 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 peaks at 9 to 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 $K(\ell; N, J) = N/[1+ (N-1)J^{-\ell}]$ is the number of proteins within ℓ links of a central protein and the symbol $'$ indicates differentiation with respect to ℓ, $K'(d; N, J) = N(N-1)(\ln J)J^{-\ell}/[1 + (N-1)J^{-\ell}]^2$. Although 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 (*7*). 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 $\Pr(\ell)$ that the
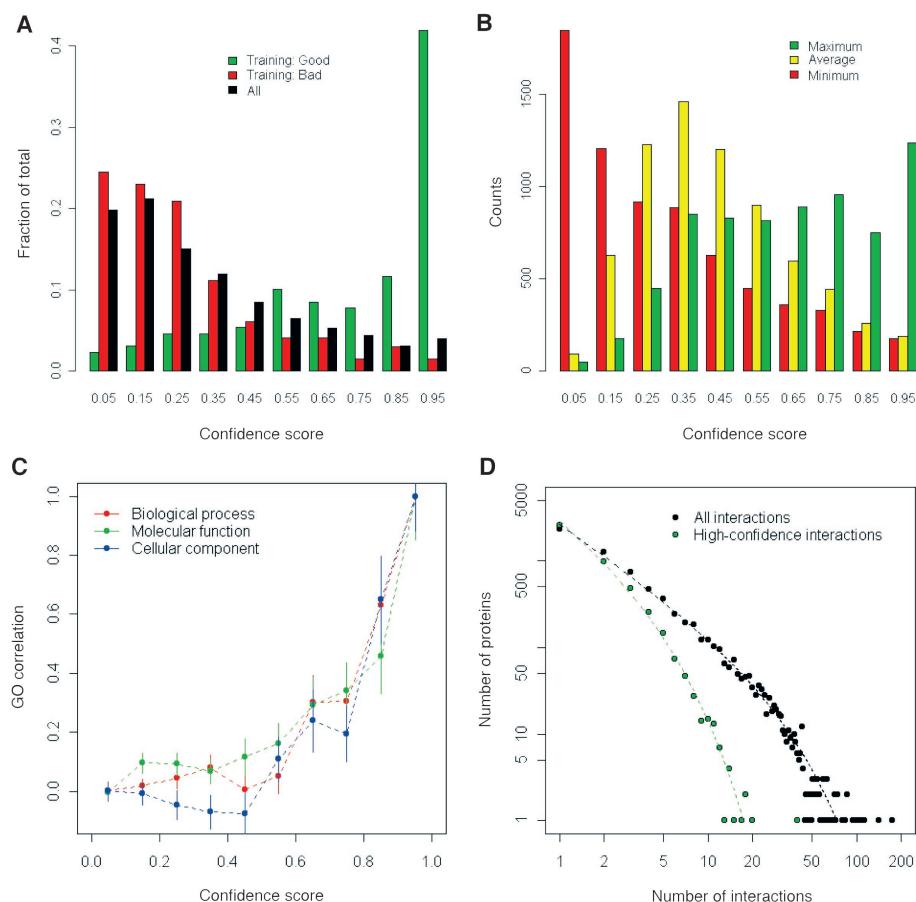

Fig. 2. Confidence scores for protein-protein interactions (**A**) *Drosophila* protein-protein interactions have been binned according to confidence score for the entire set of 20,405 interactions (black), the 129 positive training set examples (green), and the 196 negative training set examples (red). (**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 annotation (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, $Prob(n) \sim n^{-\alpha}\exp^{-\beta n}$, indicated by the dashed lines, with r^2 for the fit greater than 0.98 in either case (all interactions: $\alpha = 1.20 \pm 0.08$, $\beta = 0.038 \pm 0.006$; high-confidence interactions: $\alpha = 1.26 \pm 0.25$, $\beta = 0.27 \pm 0.05$). Note that the power-law exponents are within 1σ for the two interaction sets.

shortest path has ℓ links is

$$\Pr(\ell) = (N_1 N_2 - 1)^{-1}[K'(\ell; N_1, J_2) +$$

$$K'(\ell; N_1, N_2 J_1) + \int_0^\ell dx K'(x; N_1, N_2 J_1)K'$$

$$(\ell - x; N_2, J_2)]$$

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 two-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, $P = 0.16$; see (6) for fitting parameters].

Within multiprotein 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 two-level model in which proteins are organized into N_1 complexes with N_2 proteins per complex, with J_1 be-

tween-complex links and J_2 within-complex links per protein, the number of loops is

$$(\text{\# loops of perimeter } L) = (J_1 + J_2)^L/2L +$$

$$N_1(J_2^L/2L)\exp(-L^2/2N_2)$$

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 two-level model provides a significantly better fit than the one-level model ($P = 4.5 \times 10^{-5}$); for the random network, the fits are indistinguishable ($P = 0.996$).

The two-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 versus 2.2), between-complex neighbors per protein (0.1 versus 0.8), and proteins in each complex (7 versus 40). This difference arises in part because we used 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 ef-

fective dimensionality, are known to depend on the distance scale measured. Thus, although 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 multiprotein 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 disks are color-coded according to broad classes of molecular functions as taken from the Gene Ontology annotations (see legend) (17). 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) (18). The linkage of proteins altered in human disease to enzyme classes, some of which are potential drug targets, 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 (19), 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 (20). 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 subcellular 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
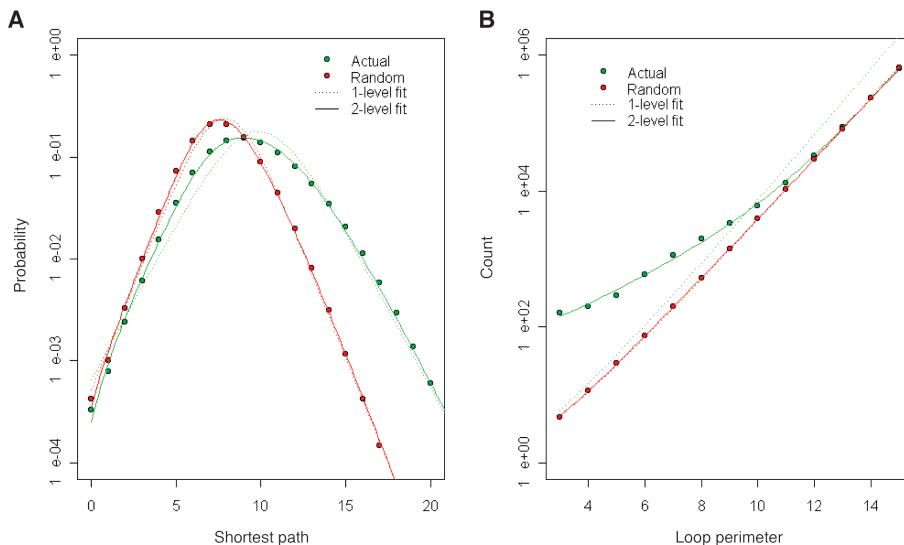
Fig. 3. Statistical properties of the refined *Drosophila* interaction map. The high-confidence *Drosophila* protein-protein interactions form a small-world network with evidence for a hierarchy of organization. Network properties are presented for the giant connected component, in which 3659 pairwise 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 in the actual network (green points) peaks at 9 to 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 to 11 relative to the random network (red points). In both (A) and (B), the one-level and two-level models produce nearly indistinguishable fits for the random networks, indicating the absence of structured clustering.

localization, and potential function, of proteins that have not been studied or annotated previously. In Fig. 4B, a local protein-interaction network is enlarged that 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 by PSORT I and 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; interactions between compartments, which represent intercompartment communication or protein shuttling, would be more rare. As summarized in table S6, we observe strong enrichment of nuclear-nuclear, cytoplasm-cytoplasm, cytoskeleton-cytoskeleton, and endoplasmic reticulum–endoplasmic reticulum interactions. Intercompartment interactions (e.g., nucleus-plasma membrane, extracellular-nucleus) tend to be depleted from the data set, consistent with the view that intercompartment communication is a relatively rare regulatory event. Although 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 Fig. 5A. The binding partners of the two corepressors are largely nonoverlapping, which concurs with existing evidence that CtBP and Gro repressors independently mediate short- and long-range transcriptional repression (*21*). CtBP interacts with a range of transcription factors including homeodomain, nuclear hormone receptor, and C2-H2 Zn-finger proteins, along with the NC2 α 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 (HLH) domain proteins. Gro interactors are known to interact through C-terminal WRPW motifs or the engrailed homology 1 (eh1) domains (*22, 23*). Each HLH protein shown interacting with Gro [Her, dpn, E(spl), HLHm3, HLHm5, HLHmdelta] has 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 (*24*).
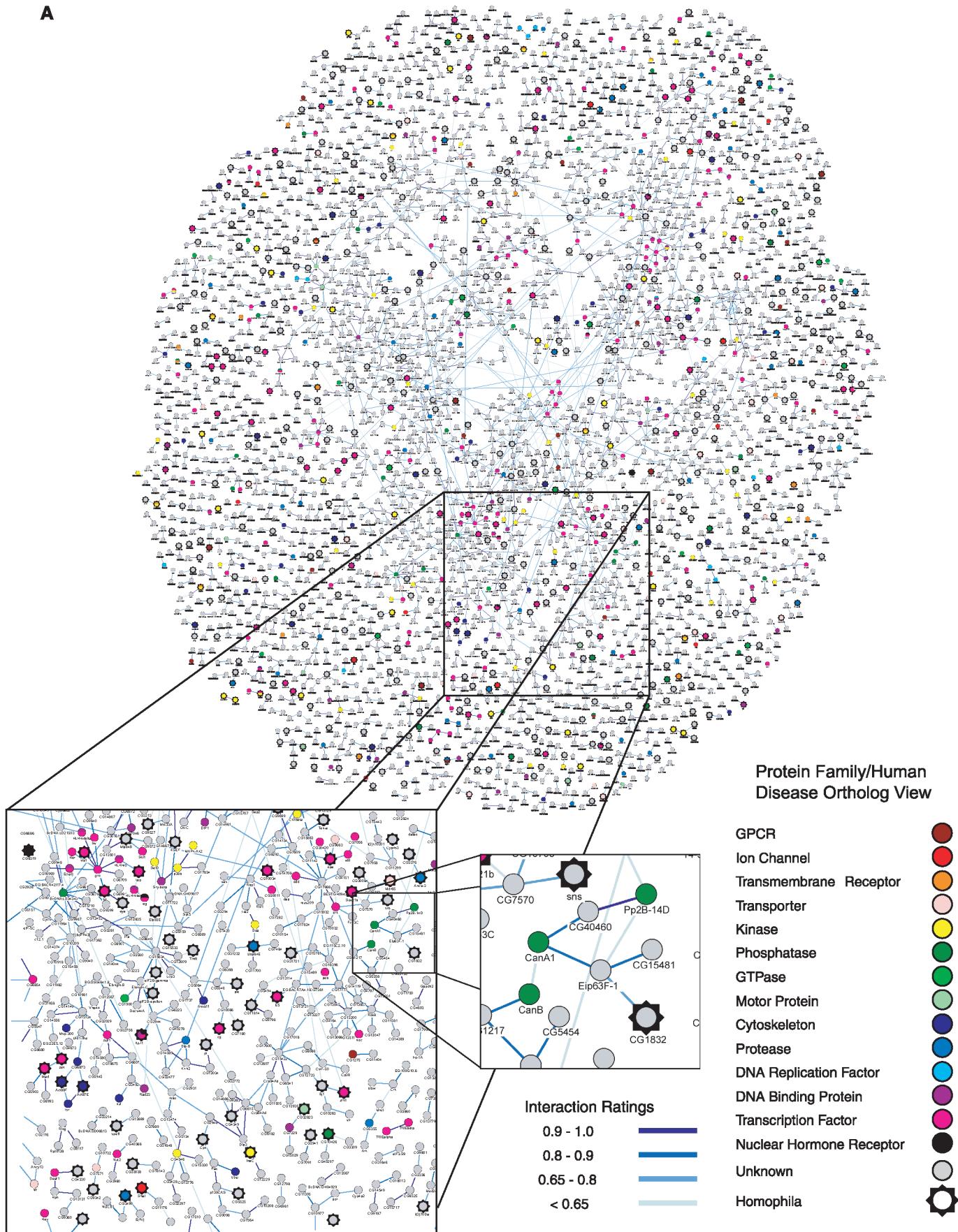
Splicing. Figure 5B shows an extensive network of proteins involved in RNA metabolism. The network captures the regulation of sex determination from X:A ratio to the machinery responsible for the splicing of doublesex and fruitless mRNAs (*25, 26*). Existing evidence indicates that both Tra-2 and Rbp1 are substrates of Doa kinase (*27*). 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 (*28–30*). Three previously unknown proteins (CG14323, CG6843, CG31211) are linked to splicing components through an extensive set of interactions. Although 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 (*31, 32*).

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 (Src homology 2/phosphotyrosine binding) domains and recruit other proteins via their SH3 domains. Within the Sevenless tyrosine kinase pathway, drk is known to recruit dos. (*33, 34*), whereas here drk potentially recruits CG13358 and Nek2, a serine/threonine kinase. A previously unknown adaptor protein, CG2079, with PTB and PH (pleckstrin homology) 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 (*35*). Two previously unknown proteins that interact with the bifunctional adaptor proteins in the pathway are CG15022 and CG13358. Inspection of their sequences indicated that they both have polyproline 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), previously unknown calmodulin-like proteins [CG11165, CG31958, EG:BACR7A4.12 (CG11638)], calmodulin-binding proteins, and the calcineurin family of calmodulin-dependent serine/threonine 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 (*36–38*). A potentially important link between the tyrosine transporter hoe1 and two calcineurin phosphatases is shown. A mutation in the human homolog of hoe1 causes ocular albinism.

Cell cycle regulation: Figure 5E shows the network surrounding the Skp protein complex (SCF complex) that targets proteins to ubiquitin-mediated proteasomal degradation (*39*). Target proteins are recruited to the Skp complex by F-box proteins (*40–42*). Among the Skp proteins, only SkpA reportedly binds F-box proteins (*43*). 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 (*44*). Other notable 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 (*45*) and is reported to be an inhibitor of the anaphase-promoting complex (APC) (*46*). 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 (*47, 48*). The Sickle (skl) protein is a recently described DIAP-binding protein that induces apoptosis (*49, 50*). The presence of skl in the Skp complex suggests that, like Morgue, it may target DIAP1 for degradation by the SCF complex. As shown in Fig. 5D, skl protein also interacts with several calmodulin-binding proteins. It is tempting to speculate that skl may

A

Protein Family/Human
Disease Ortholog View

GPCR
Ion Channel
Transmembrane Receptor
Transporter
Kinase
Phosphatase
GTPase
Motor Protein
Cytoskeleton
Protease
DNA Replication Factor
DNA Binding Protein
Transcription Factor
Nuclear Hormone Receptor
Unknown
Homophila

Interaction Ratings

0.9 - 1.0
0.8 - 0.9
0.65 - 0.8
< 0.65

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

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.

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 for degradation has been reported (*51*). 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 data set. Three of these pathways are described here; the others are described in the supporting online material.

Innate immunity. The Imd pathway is a well-characterized *Drosophila* signaling complex involved in the innate immune response against Gram-negative bacteria (*52*). 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 complexes 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 the ubiquitin pathway represses IMD signaling (*53*) 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 *Caenorhabditis elegans* protein complex LET-23-LIN-2-LIN-7, which is involved in the polarized localization of the LET-23 receptor [epidermal growth factor (EGF) receptor] during vulval development (*54*). 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 the *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 (*55*, *56*). 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
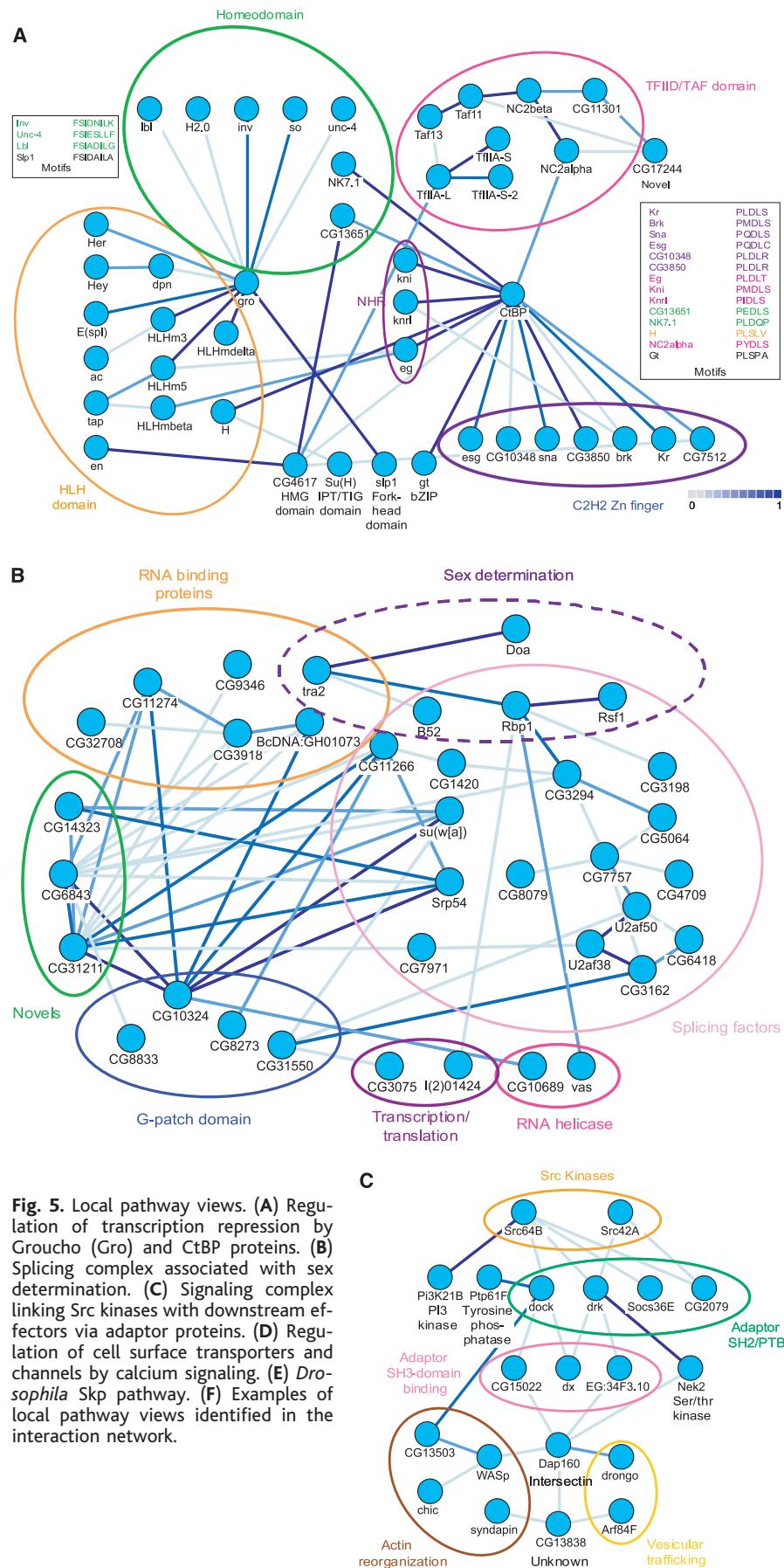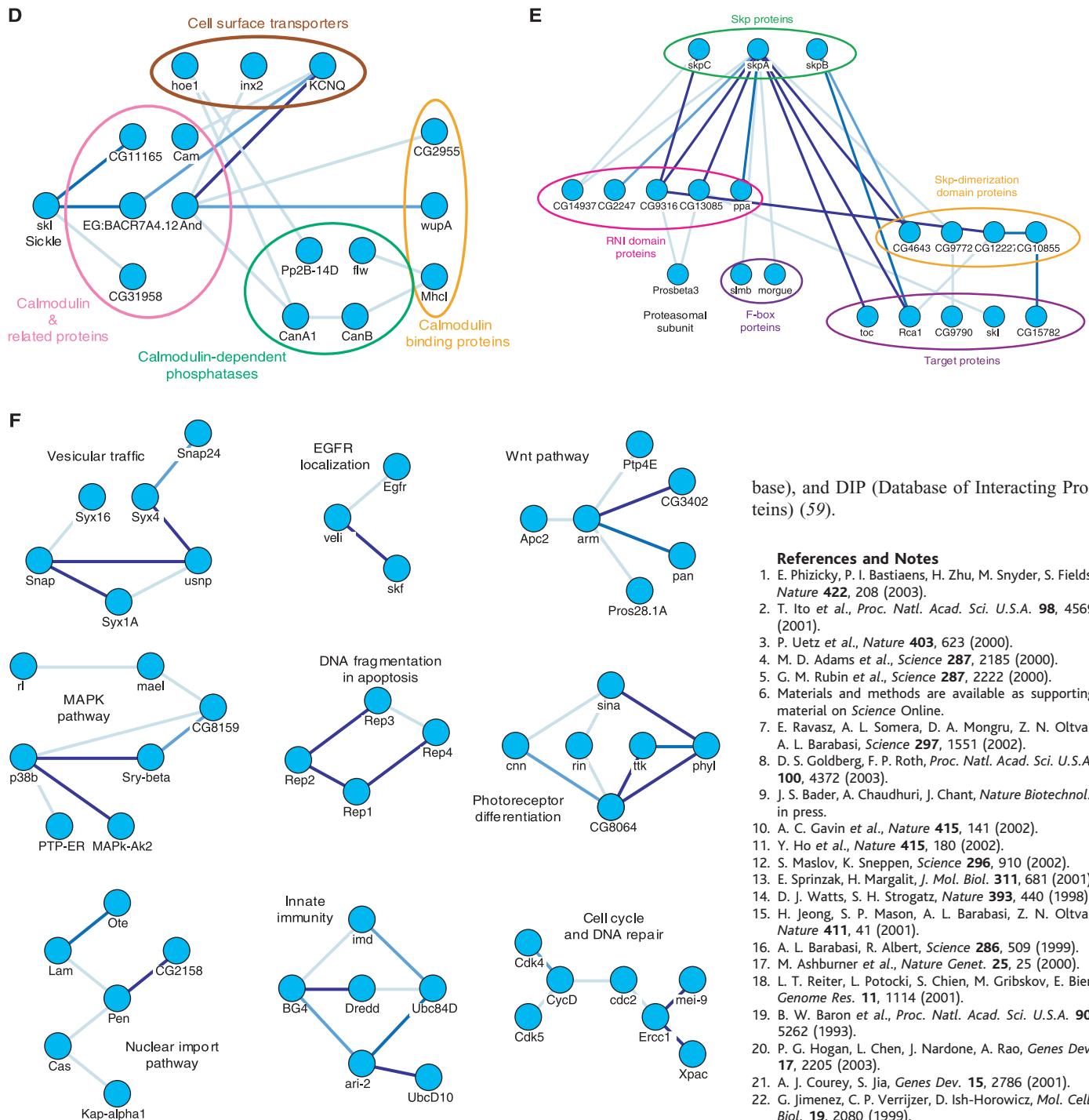
Fig. 5. Local pathway views. (**A**) Regulation of transcription repression by Groucho (Gro) and CtBP proteins. (**B**) Splicing complex associated with sex determination. (**C**) Signaling complex linking Src kinases with downstream effectors via adaptor proteins. (**D**) Regulation of cell surface transporters and channels by calcium signaling. (**E**) *Drosophila* Skp pathway. (**F**) Examples of local pathway views identified in the interaction network.

D

Cell surface transporters

hoe1 inx2 KCNQ

CG11165 Cam

skl
Sickle

EG:BACR7A4.12 And

CG31958

Calmodulin
&
related proteins

CG2955

wupA

Pp2B-14D flw

CanA1 CanB

Mhcl

Calmodulin
binding proteins

Calmodulin-dependent
phosphatases

E

Skp proteins

skpC skpA skpB

CG14937 CG2247 CG9316 CG13085 ppa

RNI domain
proteins

Prosbeta3

Proteasomal
subunit

slmb morgue

F-box
porteins

Skp-dimerization
domain proteins

CG4643 CG9772 CG12227 CG10855

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

CG8064

Photoreceptor
differentiation

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

two separate domains in phyl that bind Sina and ttk (57). A previously unknown interactor in our pathway is rin (rasputin), a RasGAP protein that functions in eye development as a regulator of RAS signaling (58). 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). Importantly, 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 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, GRID (General Repository of Interaction Datasets), BIND (Biomolecular Interaction Network Database), and DIP (Database of Interacting Proteins) (59).

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. Materials and methods are available as supporting material on *Science* Online.
7. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabasi, *Science* **297**, 1551 (2002).
8. D. S. Goldberg, F. P. Roth, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4372 (2003).
9. J. S. Bader, A. Chaudhuri, J. Chant, *Nature Biotechnol.*, in press.
10. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
11. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
12. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
13. E. Sprinzak, H. Margalit, *J. Mol. Biol.* **311**, 681 (2001).
14. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
15. H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).
16. A. L. Barabasi, R. Albert, *Science* **286**, 509 (1999).
17. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
18. L. T. Reiter, L. Potocki, S. Chien, M. Gribskov, E. Bier, *Genome Res.* **11**, 1114 (2001).
19. B. W. Baron *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 5262 (1993).
20. P. G. Hogan, L. Chen, J. Nardone, A. Rao, *Genes Dev.* **17**, 2205 (2003).
21. A. J. Courey, S. Jia, *Genes Dev.* **15**, 2786 (2001).
22. G. Jimenez, C. P. Verrijzer, D. Ish-Horowicz, *Mol. Cell. Biol.* **19**, 2080 (1999).
23. G. Jimenez, A. Guichet, A. Ephrussi, J. Casanova, *Genes Dev.* **14**, 224 (2000).
24. K. Jagla, M. Bellard, M. Frasch, *Bioessays* **23**, 125 (2001).
25. P. Graham, J. K. Penn, P. Schedl, *Bioessays* **25**, 1 (2003).
26. B. R. Graveley, *Cell* **109**, 409 (2002).
27. C. Du, M. E. McGuffin, B. Dauwalder, L. Rabinow, W. Mattox, *Mol. Cell* **2**, 741 (1998).
28. Y. J. Kim, P. Zuo, J. L. Manley, B. S. Baker, *Genes Dev.* **6**, 2569 (1992).
29. K. W. Lynch, T. Maniatis, *Genes Dev.* **10**, 2089 (1996).
30. V. Heinrichs, B. S. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 115 (1997).
31. L. Aravind, E. V. Koonin, *Trends Biochem. Sci.* **24**, 342 (1999).
32. B. Guglielmi, M. Werner, *J. Biol. Chem.* **277**, 35712 (2002).

33. S. M. Feller, H. Wecklein, M. Lewitzky, E. Kibler, T. Raabe, *Mech. Dev.* **116**, 129 (2002).
34. J. P. Olivier *et al.*, *Cell* **73**, 179 (1993).
35. D. Cai, S. Dhe-Paganon, P. A. Melendez, J. Lee, S. E. Shoelson, *J. Biol. Chem.* **278**, 25323 (2003).
36. E. Yus-Najera, I. Santana-Castro, A. Villarroel, *J. Biol. Chem.* **277**, 28545 (2002).
37. A. Sotkis *et al.*, *Cell Commun. Adhes.* **8**, 277 (2001).
38. C. Peracchia, A. Sotkis, X. G. Wang, L. L. Peracchia, A. Persechini, *J. Biol. Chem.* **275**, 26220 (2000).
39. R. J. Deshaies, *Annu. Rev. Cell Dev. Biol.* **15**, 435 (1999).
40. R. M. Feldman, C. C. Correll, K. B. Kaplan, R. J. Deshaies, *Cell* **91**, 221 (1997).
41. E. T. Kipreos, M. Pagano, *Genome Biol.* **1**, REVIEWS3002 (2000).
42. D. Skowyra, K. L. Craig, M. Tyers, S. J. Elledge, J. W. Harper, *Cell* **91**, 209 (1997).
43. C. Bai *et al.*, *Cell* **86**, 263 (1996).
44. J. P. Wing *et al.*, *Nature Cell Biol.* **4**, 451 (2002).
45. X. Dong *et al.*, *Genes Dev.* **11**, 94 (1997).
46. R. Grosskortenhaus, F. Sprenger, *Dev. Cell* **2**, 29 (2002).
47. D. Ganoth *et al.*, *Nature Cell Biol.* **3**, 321 (2001).
48. D. Sitry *et al.*, *J. Biol. Chem.* **277**, 42233 (2002).
49. A. Christich *et al.*, *Curr. Biol.* **12**, 137 (2002).
50. J. P. Wing *et al.*, *Curr. Biol.* **12**, 131 (2002).
51. L. Raj *et al.*, *Curr. Biol.* **10**, 1265 (2000).
52. J. A. Hoffmann, J. M. Reichhart, *Nature Immunol.* **3**, 121 (2002).
53. R. S. Khush, W. D. Cornwell, J. N. Uram, B. Lemaitre, *Curr. Biol.* **12**, 1728 (2002).
54. S. M. Kaech, C. W. Whitfield, S. K. Kim, *Cell* **94**, 761 (1998).
55. Z. C. Lai, S. D. Harrison, F. Karim, Y. Li, G. M. Rubin, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5025 (1996).
56. S. Li, R. W. Carthew, Z. C. Lai, *Cell* **90**, 469 (1997).
57. S. Li, C. Xu, R. W. Carthew, *Mol. Cell. Biol.* **22**, 6854 (2002).
58. C. Pazman, C. A. Mayes, M. Fanto, S. R. Haynes, M. Mlodzik, *Development* **127**, 1715 (2000).
59. Accession numbers are as follows: BIND IDs 24880 to 45318.
60. 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.). The interactions are also available at our Web sites www.curagen.com and www.jhubiomed.org.

REPORTS

Probing the Threshold to H Atom Transfer Along a Hydrogen-Bonded Ammonia Wire

Christian Tanner, Carine Manca, Samuel Leutwyler*

We characterized the entrance channel, reaction threshold, and mechanism of an excited-state H atom transfer reaction along a unidirectionally hydrogen-bonded "wire" $-O-H\cdots NH_3\cdots NH_3\cdots NH_3\cdots N$. Excitation of supersonically cooled 7-hydroxyquinoline$\cdot(NH_3)_3$ to its vibrationless S_1 state produces no reaction, whereas excitation of ammonia-wire vibrations induces H atom transfer with a reaction threshold ≈ 200 wave numbers. Further translocation steps along the wire produce the S_1 state 7-ketoquinoline$\cdot(NH_3)_3$ tautomer. Ab initio calculations show that proton and electron movement along the wire are closely coupled. The rate-controlling S_1 state barriers arise from crossings of a $\pi\pi^*$ with a Rydberg-type $\pi\sigma^*$ state.

Proton and H atom transfer mechanisms and dynamics are of fundamental importance for acid-base chemistry (*1–13*). The transport of protons displays a number of unusual properties, in part because the conduction mechanism of protons is different from that of other ions (*2–4*). Direct experimental probing of proton-transfer reactions at the molecular level is difficult because of the short times, microscopic length scales, and the solvent fluctuations involved. Recently, the autoionization and excess proton transfer processes in water have been studied with the use of ab initio molecular dynamics methods (*3, 4*). Also, proton relay systems in enzymes, ion channels, and membrane-spanning proteins ("proton wires") have become topics of intense theoretical study (*5–8*).

In our experiments (*14*), a hydrogen-bonded ammonia "wire" of $NH_3\cdots NH_3\cdots NH_3$ is attached to the aromatic "scaffold" molecule 7-hydroxyquinoline (7HQ), which offers an O–H and a N hydrogen-bonding site that are spaced far enough apart to stretch the ammonia wire (Fig. 1A). Furthermore, the O–H group is acidic and can serve to inject a proton or H atom, whereas the quinolinic N site is an acceptor group (*15–18*). Hence, H^+ or H-atom translocation along this wire is directionally biased (diode-like). We use $S_1 \leftarrow S_0$ optical excitation of 7HQ to render the O–H group more acidic and the N atom more basic (*15–18*).

High-level ab initio calculations predict the gas-phase 7HQ$\cdot(NH_3)_3$ cluster in Fig. 1A to be the most stable structure, whereas the tautomeric 7-ketoquinoline (7KQ) form is 36 kJ mol^{-1} less stable. This energetic ordering is also indicated by the S_0 state potential in Fig. 1B. Hence, neither proton nor H atom conduction occurs in the ground-state 7HQ$\cdot(NH_3)_3$ cluster (*17, 18*). As we show below, after the $S_1 \leftarrow S_0$ excitation, the enol 7HQ injects an H atom into the ammonia wire. Three successive H atom transfers occur in the O–H $\rightarrow$ N direction by a single-file mechanism to form the 7KQ tautomer. Each step involves a different H atom, analogous to the Grotthus proton conduction mechanism in water (*1–7*). At this point, the hydrogen bond polarity of the wire is fully reversed (Fig. 1A). The 7HQ$\cdot(NH_3)_3$ cluster is large enough to allow the multistep H atom transfer reaction to proceed without dissociation of the system. Minimal system size, which is crucial for theory, is achieved by using each NH_3 and the 7HQ molecule alternately as reactant, solvent, or product. Here, we spectroscopically characterize the entrance channel region of the first H atom transfer, report the observation of a low-energy reaction threshold, and analyze the subsequent steps by fluorescence measurements and ab initio theory.

The $S_1 \leftarrow S_0$ two-color resonant two-photon ionization (2C-R2PI) spectrum of 7HQ$\cdot(NH_3)_3$ (Fig. 2A) displays an intense origin at 28,798.4 cm^{-1} [in the ultraviolet (UV) at a wavelength of 347.1 nm] as well as ammonia-wire vibrations 125 to 200 cm^{-1} above the origin. About 200 cm^{-1} above the origin, the 2C-R2PI spectrum falls off abruptly, and only traces of band structure are observed at +256 and +332 cm^{-1}. The $S_1 \leftarrow S_0$ UV–UV-hole-burning (HB) spectrum of 7HQ$\cdot(NH_3)_3$ (Fig. 2B) also exhibits narrow vibronic bands. From the origin up to 200 cm^{-1}, the band positions are identical to those in Fig. 2A, but the HB spectrum exhibits sharp bands far above the 200 cm^{-1} falloff of the R2PI signal. Complementing these measurements, excitation of the $S_1 \rightarrow S_0$ emission of 7HQ$\cdot(NH_3)_3$ in the 0 to 200

Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, CH-3012 Bern, Switzerland.

*To whom correspondence should be addressed. E-mail: samuel.leutwyler@iac.unibe.ch