SUPPLEMENTARY FIGURES
Supplementary Figure 1:
Flow charts detailing the different interactome mapping strategies

1) Basic serial

2) Pooling
3) Thresholding

Start

Set interaction prior probabilities to background probability.

Set first binary and pooled assay types. Set upper & lower thresholds. Distribute all strains into bait and prey arrays.

Reached Target Coverage?

Yes

End

No

Are there enough candidates (interactions) with probabilities between the prior and upper threshold?

Yes

Test next plate of binary interactions with highest probabilities between the lower & upper thresholds.

Switch binary assay type.

Update interaction probabilities & Interactome map.

No

Completed comprehensive screen with current pooled assay type?

Yes

Switch to next pooled assay type.

Switch to next bait array.

Screened current bait array against all prey arrays?

Yes

No

Phase I pooling: mate pool of prey strains with each strain in the bait array.

Phase II assays: mate positive baits with individual prey in positive pools

Switch to next prey array.

4) Predictions

Start

Set interaction prior probabilities using classifier predictions.

Set first binary and pooled assay types. Set upper & lower thresholds. Distribute all strains into bait and prey arrays.

Reached Target Coverage?

Yes

End

No

Are there enough candidates (interactions) with probabilities between the prior and upper threshold?

Yes

Test next plate of binary interactions with highest probabilities between the lower & upper thresholds.

Switch binary assay type.

Update interaction probabilities & Interactome map.

No

Completed comprehensive screen with current pooled assay type?

Yes

Switch to next pooled assay type.

Switch to next bait array.

Screened current bait array against all prey arrays?

Yes

No

Phase I pooling: mate pool of prey strains with each strain in the bait array.

Phase II assays: mate positive baits with individual prey in positive pools

Switch to next prey array.
SUPPLEMENTARY TABLES

Supplementary Table 1: Interactome completion times for a range of error parameters. Number of complete screens required for 95% saturation and 95% coverage for different values of false-discovery and false-negative rates.

| FNR | FDR   |
|-----|-------|
|     | 10%   | 50%   | 82%   |
| 50% | 5, 8  | 5, 11 | 5, 11 |
| 66% | 8, 12 | 8, 17 | 8, 17 |
| 90% | 29, 46| 29, 61| 29, 61|

Supplementary Table 2: Results of yeast two-hybrid test of predicted interactions.

| Predictions due to conservation (Sharan) | Pairs tested | Pairs positive | Percent positive |
|-----------------------------------------|--------------|---------------|-----------------|
| Predictions combining multiple evidence types | 606          | 171           | 28.2            |
| All predictions                         | 2047         | 279           | 13.6            |
| All predictions                         | 2516         | 378           | 15.0            |
| Random pairs                            | 2354         | 72            | 3.1             |
SUPPLEMENTARY METHODS

**Predicting Drosophila protein-protein interactions by combination of multiple evidence types.** New protein-protein interactions were predicted through an integrative probabilistic approach combining multiple lines of evidence. Our method integrated many features, trained prediction models using different classifiers, and then combined their results to obtain the final prediction. Cross-validation and independent evaluation showed that our predictions were significantly enriched for true protein-protein interactions.

The prediction model consisted of two stages. In the first stage, both a decision tree (C4.5) and Naive Bayes models were trained separately using the attributes described below. Each of the two methods assigned a numeric value between 0 and 1 to each pair of proteins, representing the probability that the pair has an interaction. The two probabilities were then combined in a logistic regression model. Probabilities reported by the logistic regression model were taken to be the final probability representing the possibility that a protein pair is a true interaction. We predicted protein interactions for pairs with probabilities greater than 0.95 and removed proteins having more than 80 predictions to obtain a final set of reasonable size. The final set of predictions contained 24,798 interactions covering 4,702 fly proteins with an average distance of 4.15 links between proteins.

**Attributes used in training and prediction.** The following attributes were used in training and interaction prediction:

1) **Gene expression**
   a) Gene expression data throughout embryogenesis have been previously obtained using Affymetrix microarray analysis by Tomancak et al.\(^1\) These data were used to compute Pearson correlation coefficients between -1 and 1 for all pairs of genes.
   b) Gene expression data covering approximately 5,000 genes throughout the developmental life cycle of Drosophila have been previously obtained by cDNA microarrays in Dr. Kevin White’s group\(^2\). These data were used to compute Pearson correlation coefficients between -1 and 1 for all pairs of genes.
   c) We also used pairs of genes that are co-expressed across evolution, i.e. multiple genomes\(^3\). A pair was encoded as 1 if the two genes are co-expressed in at least two organisms and as 0 otherwise.

2) **Domain-domain interaction**
   It is believed that most protein-protein interactions are mediated by domain-domain interactions. Interdom is a database of putative interacting protein domains derived from multiple data sources\(^4\). This attribute was encoded as 1 if the gene pair contains an interacting domain pair, and as 0 otherwise.

3) **Interologs (conserved interaction)**
   Similar to Matthews et al.\(^5\), we defined Interologs as fly protein pairs for which corresponding homologs had been shown to interact in human, yeast, or worm. Interologs were encoded as 1 or 0.
a) Human interologs were derived from human protein interaction data downloaded from HPRD9.
b) Yeast interologs were derived from yeast protein interaction data downloaded from DIP7.
c) Worm interologs were derived from C. elegans protein interaction data downloaded from DIP7.

4) Genetic interactions
Genetic interactions were downloaded from the FlyBase8. These data were encoded as 1 if a gene pair is a genetic interaction, and 0 otherwise.

5) Gene Ontology annotations9
a) Molecular functions: for each pair of genes A-B, we counted how many gene pairs (out of all possible binary gene combinations) have the exact same GO Molecular Function annotations as A-B.
b) Cellular components: for each pair of genes A-B, we counted how many gene pairs (out of all possible binary gene combinations) have the exact same GO Cellular Component annotations as A-B.

The above two data items have integer values from 0 to 98 million, the number of all possible gene pairs of Drosophila melanogaster. Biological process annotations were reserved for evaluation purpose and were not used in the training process.

6) Flybase anatomy annotations
FlyBase provides a controlled vocabulary annotating where a gene is expressed10. The assumption is that when two proteins are expressed in the same location, they are more likely to interact than proteins not found to be expressed at the same location. Each gene was assigned a list of anatomy terms. This attribute was encoded as 1 or 0, where 1 means two genes share at least one anatomy term and 0 means otherwise.

**Obtaining training positives and negatives.** We used a set of reported high-confidence yeast two hybrid results (2,689 interactions) as our positive training examples. Specifically, we took the union of the following as our positive training set:

1) Interactions from Giot et al.11 with confidence score greater than 0.8.
2) Interactions from Stanyon et al.12 with reporter score sums greater than 3.
3) Interactions from Formstecher et al.13 with scores of ‘A’, ‘B’, or ‘C’. These authors assigned confidence scores to their data such that $A > B > C > D > E$.

Negative training examples were synthesized by selecting random protein pairs. Any interactions that had been detected by the yeast two-hybrid screens were deleted from this set (to enrich for true negatives). The final negative training set contained 131,761 random pairs of Drosophila genes.

**Computational evaluation of prediction results.** The accuracy of the predictions was assessed using the technique of five-fold cross-validation. By this approach, 20% of the positive and negative data are withheld during training and, once training is completed, the model predictions are tested on the withheld data. This process is repeated five times,
with a different fraction of data withheld each time. The average precision over these
five trials was 89%; the recall was 27%.

As another means of evaluating prediction results, we withheld GO Biological
Process annotations from training and prediction and instead used them to score the
biological relevance of the predicted interactions. We computed the number of pairs
sharing at least one annotation as well as the average depth of shared annotations. We
then generated 2000 sets of random gene pairs, each set with as many pairs as we
predicted, and computed the above two indexes for all random sets. The actual predicted
interactions had a much larger number of pairs sharing GO biological process annotations
(5832) than the distribution of shared annotations among random pairs, which had a
maximum value of 740 (p<2.2 x 10^{-16}). Similarly, the average depth of shared
annotations for the predictions (7.5) was significantly larger than the depths of the sets of
random pairs, which reached a maximum of 6.5 (<2.2 x 10^{-16}).

**Yeast two-hybrid test of predicted interactions.** We used the LexA-based yeast two-
hybrid mating assay\textsuperscript{14} to test two sets of predicted interactions: 469 protein pairs
predicted to interact by Sharan at al.\textsuperscript{15} based on cross-species conservation, and 2047
interactions predicted using a classifier built on multiple lines of evidence (see main text).
At the same time, we tested a background set of 2354 randomly selected pairs of proteins.
Assays were conducted using sequence-verified yeast two-hybrid clones as previously
described\textsuperscript{16}. Briefly, BD and AD yeast clones were re-arrayed into 96-well plates (using
a Beckman Biomek Span 8 robot) and grown to saturation in 100 µl of the appropriate
media for each clone type. A Beckman Biomek FX 96-channel robot was used to plate
6µl of each clone type onto a single YPD plate, which was then incubated for 36-48
hours to allow mating and formation of diploids. Yeast cells were replicated to two
different diploid selection plates to assay for two-hybrid reporter activity: one plate
lacking leucine to assay the LEU2 reporter, and the other containing X-Gal to assay the
lacZ reporter. Growth on the Leu– plate and “blueness” on the X-Gal plate were scored
using scales of 0 - 3 and 0 - 5, respectively, as previously described\textsuperscript{17}. These two values
were corrected by subtracting the LEU2 or lacZ scores that the BD clone produced by
itself (in a separate auto-activation assay). Pairs for which the sum of corrected LEU2
and lacZ scores exceeded 0.9 were recorded as positives. The quantitative results for all
assayed pairs are shown in Supplementary Table 2.

**Interaction probability model.** We use a naïve-Bayesian probabilistic model to compute
the posterior probability \( p_{ij}^t \) that proteins \( i \) and \( j \) interact \textit{in vivo} given the experimental
evidence at time \( t \). The results of two experiments on a given interaction are assumed to
be conditionally independent if they employ different assay types or different orientations
(\( i \) as bait and \( j \) as prey vs. \( j \) as bait and \( i \) as prey). The prior probability \( p_{ij}^0 \) is set as the
background probability for an interaction or using the probability assigned by a classifier
via network conservation or other predictive features.

When only direct experiments are used with no pooling, the posterior probability of
each interaction is updated after every direct experiment at time \( t \) given the false-positive
and false-negative rates of assay \( A^t \) and the interaction probability at time \( t-1 \). For
simulations reported in the main text, we used \( fpr = 0.2\% \) and \( fnr = 66\% \) for all assays
and \( p^0 = 0.1\% \). The following is a derivation of the update equation given in the main text:

\[
p_y^t = p(\text{ij interact} | \text{the prior probability } p_{ij}^0 \text{ and the results of experiments } x^1 \text{ through } x^t).
\]

For brevity we write this expression as \( p_y^t = p(y_{ij} | X_{ij}^{(t)}) \).

Using Bayes rule we get

\[
p_y^t = \frac{p(y_{ij})p(X_{ij}^{(t)} | y_{ij})}{p(X_{ij}^{(t)})} ,
\]

where:

- \( p(y_{ij}) = p_{ij}^0 \) is the prior probability that \( i \) and \( j \) interact,
- \( p(X_{ij}^{(t)} | y_{ij}) \) is the conditional probability of observing experimental results \( x^1 \) through \( x^t \) given that proteins \( i \) and \( j \) interact \textit{in vivo}, and
- \( p(X_{ij}^{(t)}) \) is the unconditional probability of observing experimental results \( x^1 \) through \( x^t \) and can be expressed as: \( p_{ij}^0 p(X_{ij}^{(t)} | y_{ij}) + (1 - p_{ij}^0) p(X_{ij}^{(t)} | \neg y_{ij}) \).

Dividing by the numerator gives:

\[
p_y^t = \frac{1}{1 + \frac{(1 - p_{ij}^0) p(X_{ij}^{(t)} | \neg y_{ij})}{p_{ij}^0 p(X_{ij}^{(t)} | y_{ij})}}.
\]

Using the conditional independence assumption \( p(X_{ij}^{(t)} | y_{ij}) \) and \( p(X_{ij}^{(t)} | \neg y_{ij}) \) can be expressed as the product of \( t \) independent probabilities:

\[
p(X_{ij}^{(t)} | y_{ij}) = \prod_{l=1}^{t} p(x_{ij}^l | y_{ij}) \quad \text{and} \quad p(X_{ij}^{(t)} | \neg y_{ij}) = \prod_{l=1}^{t} p(x_{ij}^l | \neg y_{ij}),
\]

By definition,

\[
p(x_{ij}^l | y_{ij}) = \begin{cases} 1 - fnr(A^l) & \text{if a true interaction } ij \text{ is tested and observed in experiment } l \\ fnr(A^l) & \text{if a true interaction } ij \text{ is tested but not observed in experiment } l \end{cases}
\]

and

\[
p(x_{ij}^l | \neg y_{ij}) = \begin{cases} fpr(A^l) & \text{if a false interaction } ij \text{ is tested and observed in experiment } l \\ 1 - fpr(A^l) & \text{if a false interaction } ij \text{ is tested but not observed in experiment } l \end{cases}.
\]

Substituting back and using a simple induction we get the update equation:

\[
p_y^t = \begin{cases} 1 & \text{if interaction } ij \text{ is tested and observed in experiment } t \\ \frac{1 - fpr(A^t)}{1 + \frac{fpr(A^t)}{fnr(A^t)} \frac{1 - p_{ij}^{t-1}}{p_{ij}^{t-1}}} & \text{if interaction } ij \text{ is tested but not observed in experiment } t \end{cases}.
\]
Two-phase pooling. For strategies that utilize two-phase pooling experiments, a pooling-sensitivity parameter $\rho$ is defined. Intuitively, this parameter represents the probability that a pool with a single detectable interaction will show a positive result. Given a pool of 96 AD clones mated with a single BD clone, $k$ is defined to be the number of assay-specific detectable interactions in the sample space ($k \leq 96$). Assuming that all detectable interactions in the sample space are independently likely to contribute to a positive pool, the probability that the pool does not show a positive result is $(1 - \rho)^k$, and the probability that the pool shows a positive result is given by $1 - (1 - \rho)^k$ (at least one success in $k$ independent Bernoulli trials).

According to the two-phase pooling protocol, every positive BD clone from phase I is individually mated with the 96-well AD array in phase II. Since direct mating is highly reproducible using the same assay type $^{17,18}$, we assume that the results of direct mating are deterministic given $\delta_{ij}(A')$, the detectability of an interaction in a given assay type.

As in the direct experiment case, $p'_{ij} = \frac{1}{1 + \frac{(1 - p_{ij}^0)p(X_{ij}^{(t)}|y_{ij})}{p_{ij}^0p(X_{ij}^{(t)}|\overline{y}_{ij})}}.$

However, when the $X_{ij}^{(t)}$ include multiple pooled experiments using the same assay type, the interaction probability update equation is more involved. This is due to the fact that repeated pooled experiments are conditionally independent only given the detectability $\delta_{ij}(A')$, but not given $y_{ij}$ alone. Therefore,

$$p(X_{ij}^{(t)}|y_{ij}) = p(X_{ij}^{(t-1)}|y_{ij},\delta_{ij}(A')) \cdot p(x_{ij}^t|y_{ij},\delta_{ij}(A')) \cdot p(\delta_{ij}(A')|y_{ij}) + p(X_{ij}^{(t-1)}|y_{ij},\overline{\delta}_{ij}(A')) \cdot p(x_{ij}^t|y_{ij},\overline{\delta}_{ij}(A')) \cdot p(-\delta_{ij}(A')|y_{ij}).$$

Assuming that $k$ is small (which is generally true for most pools) we can approximate $p(\delta_{ij}(A')|y_{ij}) \approx \rho$ and $p(-\delta_{ij}(A')|y_{ij}) \approx 1 - \rho$. Substituting back, and using simple induction we get the following update rules.

Protein pairs in the sample space of the first phase that do not show a positive BD clone have their posterior probabilities updated using the following equation:

$$p'_{ij} = \frac{1}{1 + \frac{1 - \text{fpr}(A')\rho}{\text{fnr}(A') + (1 - \text{fnr}(A'))(1 - \rho)} \cdot \frac{\text{fpr}(A') + (1 - \text{fnr}(A'))(1 - \rho)}{1 - \text{fpr}(A') + \text{fpr}(A')(1 - \rho)} \cdot \frac{1 - p_{ij}^{t-1}}{p_{ij}^{t-1}}}$$

where $\tau_{ij}(A')$ captures the number of times pair $ij$ has been in the sample space of the first phase of a pooling experiment using assay type $A'$ without showing a positive result. When an interaction is tested directly in the second phase of a pooling experiment its posterior probability is updated as follows:
Lower and upper thresholds. The Thresholding and Prediction strategies stop testing a protein pair once its posterior probability of interaction is greater than the upper bound ($P_u$) or less than the lower bound ($P_l$). $P_u = 1 - FDR$, where $FDR$ is the target false-discovery-rate (95%). This ensures that the interactome map at any time $t$ always includes less than $1 - FDR$ false positives. The lower bound is defined in order to stop testing protein pairs that are very unlikely to interact in vivo. This reduces the cost spent on testing non-interacting pairs, which compose the majority of the candidate interactions. The lower bound is a function of the target coverage level ($C$), and should leave out less than $100(1 - C)$ percent of the true interactions. We estimate the lower bound by defining $L$ to be the smallest integer, such that the probability that a true interaction be tested directly $L$ times with zero positive results is less than $1 - C$. This leads to the following inequality: $fnr^L \leq 1 - C$.

Solving for $L$ we get, $L \geq \frac{\ln(1 - C)}{\ln(fnr)}$.

Using the above equation $p_{ij}^t = \left\{ \begin{array}{ll} 1 & \text{if interaction } ij \\
\frac{1}{1 + \frac{fpr(A^t)}{fpr(A^t)} + (1 - fnr(A^t))(1 - \rho)^{\tau_y(A^t)} \cdot \frac{1 - p_{ij}^{t-1}}{p_{ij}^{t-1}}} & \text{if interaction } ij \text{ tested but not observed} \\
\frac{1}{1 - fnr(A^t) + fpr(A^t)} & \text{in experiment } t
\end{array} \right.$

and setting $p_{ij}^0 = \left( 1 - \frac{1 - p_{ij}^{0}}{p_{ij}^{0}} \right)^{\ln(1 - C)}$, we get the lower bound probability:

$$P_l(C) = \frac{1}{1 + \left( \frac{1 - p_{ij}^{0}}{p_{ij}^{0}} \right)^{\frac{\ln(1 - C)}{\ln(fnr)}}}.$$
SUPPLEMENTARY REFERENCES

1. Tomancak, P., et al. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* 3, RESEARCH0088 (2002).

2. Arbeitman, M.N., et al. Gene Expression During the Life Cycle of Drosophila melanogaster. *Science* 297, 2270-2275 (2002).

3. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* 302, 249-255 (2003).

4. Ng, S.K., Zhang, Z. & Tan, S.H. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics (Oxford, England)* 19, 923-929 (2003).

5. Matthews, L.R., et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* 11, 2120-2126 (2001).

6. Peri, S., et al. Human protein reference database as a discovery resource for proteomics. *Nucleic acids research* 32, D497-501 (2004).

7. Xenarios, I., et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research* 30, 303-305 (2002).

8. Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P. & Gelbart, W.M. FlyBase: genomes by the dozen. *Nucleic acids research* 35, D486-491 (2007).

9. The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nat Genet* 25(2000).

10. Grumbling, G. & Strelets, V. FlyBase: anatomical data, images and queries. *Nucleic acids research* 34, D484-488 (2006).

11. Giot, L., et al. A protein interaction map of Drosophila melanogaster. *Science* 302, 1727-1736 (2003).

12. Stanyon, C.A., et al. A *Drosophila* Protein-Interaction Map Centered on Cell-Cycle Regulators. *Genome Biology* 5, R96 (2004).

13. Formstecher, E., et al. Protein interaction mapping: a Drosophila case study. *Genome Res* 15, 376-384 (2005).

14. Finley, R.L., Jr. & Brent, R. Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. *Proc Natl Acad Sci U S A* 91, 12980-12984 (1994).
15. Sharan, R., et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102**, 1974-1979 (2005).

16. Stanyon, C.A., et al. A Drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol* **5**, R96 (2004).

17. Zhong, J., Zhang, H., Stanyon, C.A., Tromp, G. & Finley, R.L., Jr. A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. *Genome Res* **13**, 2691-2699 (2003).

18. Rual, J.F., et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).