Cost-effective strategies for completing the interactome

Ariel S Schwartz[1], Jingkai Yu[2], Kyle R Gardenour[2], Russell L Finley Jr[2] & Trey Ideker[1]

Comprehensive protein-interaction mapping projects are underway for many model species and humans. A key step in these projects is estimating the time, cost and personnel required for obtaining an accurate and complete map. Here we modeled the cost of interaction-map completion for various experimental designs. We showed that current efforts may require up to 20 independent tests covering each protein pair to approach completion. We explored designs for reducing this cost substantially, including prioritization of protein pairs, probability thresholding and interaction prediction. The best experimental designs lowered cost by fourfold overall and >100-fold in early stages of mapping. We demonstrate the best strategy in an ongoing project in *Drosophila melanogaster*, in which we mapped 450 high-confidence interactions using 47 microtiter plates, versus thousands of plates expected using current designs. This study provides a framework for assessing the feasibility of interaction mapping projects and for future efforts to increase their efficiency.

Analysis of molecular networks has exploded in recent years. A wide variety of technologies have been introduced for mapping networks of gene and protein interactions, including yeast two-hybrid (Y2H) assays[1–8], affinity purification coupled to mass spectrometry[9–11], chromatin immunoprecipitation measurements of transcriptional binding[12–14], synthetic-lethal and suppressor networks[15,16], expression quantitative trait loci[17–20] and many others. Using these technologies, network mapping projects are currently underway for many model species[2–4,7–13,15], microbial[21–23] and viral[24,25] pathogens, and humans[5,6].

Mapping a complete gene or protein network evokes similar challenges and considerations as mapping a complete genome sequence. In the case of the human and model genome projects, large-scale sequencing efforts have been accompanied by feasibility studies[26,27] that used mathematical formulations and pilot projects to explore strategies for genome assembly and the work required for each. In the case of interaction networks, similar feasibility studies are just beginning[28–30]. Some of the questions to be addressed are: what is the cost of completing an interactome map and what is the best strategy for minimizing that cost? Given practical cost constraints, how can the quality and coverage of interaction data be maximized? How many independent assay types are needed? How should direct pairwise tests for interaction be combined with pooled screening? What is the effect of the test sensitivity on the final cost? How should interaction predictions be incorporated, and what is their effect on the mapping cost? Which specific improvements to experimental and computational methods are likely to have the largest effect?

To approach these questions, we modeled several standard and alternative strategies for using pairwise protein-interaction experiments to determine the interactome of the fruit fly, *Drosophila melanogaster*. Our analysis showed that completing the interactome map using sequential pairwise or pooled screening is probably too costly to be practical in terms of the number of experiments required. However, this cost can be reduced substantially using a strategy that combines pooling with prioritized testing and interaction prediction. We carried out several iterations of this strategy to efficiently map 450 new high-confidence interactions in *Drosophila*.

RESULTS

Interactome mapping: problem definition

In contrast to a genome, the interactome has been more difficult to define. Some authors have argued[31] that the "true interactome" should be defined as all possible interactions encoded by a genome: that is, the set of all pairwise protein interactions that occur in at least one biological condition or cell type. The assumption is that every true interaction will be detectable by some assay and that given enough independent measurements most of the interactome could be detected. Many assay types have been described for detecting protein-protein interactions, a few of which have been adapted to large-scale screening[1,31–33]. In contrast, some interactions may be immeasurable by any available assay or will not arise in the conditions surveyed. Therefore, we use the term 'mappable interactome' for the subset of true pairwise interactions that are reproducibly detectable by any given assay method or combination of methods.

To define appropriate criteria for determining when an interactome map is 'complete', we distinguish between 'saturation' and

[1]Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, 540 East Canfield Avenue, Detroit, Michigan 48201, USA. Correspondence should be addressed to T.I. (trey@bioeng.ucsd.edu).

RECEIVED 19 JUNE; ACCEPTED 18 NOVEMBER; PUBLISHED ONLINE 14 DECEMBER 2008; DOI:10.1038/NMETH.1283

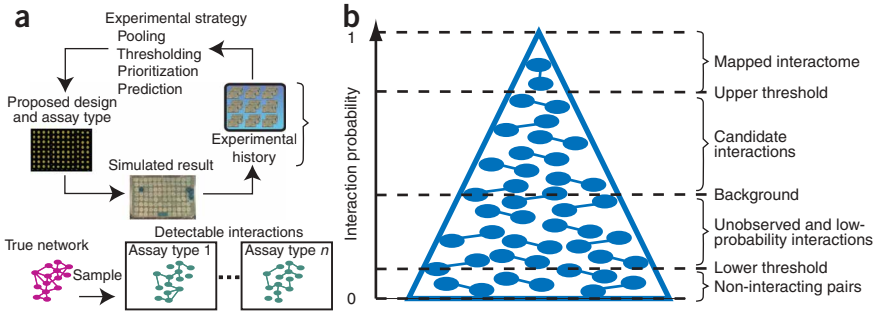

Figure 1 | Simulating an interaction mapping project. (**a**) At any given point in the project, every pair of proteins is assigned an interaction probability based on its experimental history (initially these probabilities are set to background or informed by predictions). The interaction probabilities and experimental history are used to design Y2H experiments in 96-well plate format according to one of the strategies. The result of this experiment is simulated based on the detectability of the tested interactions (given the assay type) and the pooling sensitivity. The new experimental results are recorded in the history and also used to update the interaction probabilities of the relevant protein pairs. (**b**) The pyramid represents the ordered list of protein pairs ranked by probability. Most protein pairs have a low probability of interacting and only a few pairs will move to the top of the list with high probability of interaction. Interactions with probability above an upper threshold are added to the mapped interactome, which is compared to the simulated 'true network' at intervals of 1,000 plates for reporting coverage and FDR.

'coverage'. Saturation measures the percentage of true interactions that have been experimentally observed at least once. We define coverage more strictly to mean the percentage of true interactions that have been experimentally validated with high confidence such that the percentage of false interactions (the false discovery rate; FDR) is kept below a predetermined threshold. We treat 'completion' as achieving 95% coverage of the mappable interactome at 5% FDR, which requires that the map include at least 95% of all true interactions with no more than 5% of the reported interactions being false.

A model of interactome coverage

We simulated several mapping strategies implementing various basic and sophisticated features (**Fig. 1**; see **Supplementary Fig. 1** online for flowcharts). We evaluated all strategies using a statistical model based on naive Bayes to estimate saturation and coverage of the *Drosophila* interactome as a function of the number of interaction tests. We programmed this model with the assumption that the fly interactome contains approximately 10^5 interactions overall, along with estimates for the false positive rate (FPR; the probability that a non-interacting protein pair is reported as interacting) and the false negative rate (FNR; the probability that an interacting pair is reported as non-interacting). Although the magnitudes of these errors are still under debate, previous studies[2,5,28,34,35] have reported Y2H error rates of FPR $<$ 1% with FNR of 50–80% (note that several of these studies erroneously refer to FDR as FPR). Here we used 0.2% FPR and 66% FNR.

Owing to the high FNR of a particular assay, it becomes clear that multiple assay types will likely be needed to achieve complete coverage and that these assays should be independent or at least only partially dependent. Although the precise correlations between different assay types have not been well studied, complementarity between assays has been widely assumed and occasionally demonstrated: for instance, protein interactions have been shown to be of substantially higher-confidence if they are detected in different orientations (bait-prey versus prey-bait)[2], in different Y2H screens[3,8,34], by different types of Y2H system such as LexA-based versus Gal4-based[35], or by both Y2H assay and co-affinity purification[9–11].

Basic mapping strategies in current use

We first simulated a 'basic serial' strategy (strategy 1), in which we tested all pairs of proteins for interaction sequentially. Under this formulation, achieving a saturation of 95% required eight comprehensive screens, in which we tested each protein pair by eight independent assays, equivalent to $\sim 7 \times 10^8$ pairwise tests, assuming 13,600 protein-encoding genes in fly[36] (**Fig. 2a** and **Table 1**). Moreover, 93% of all observed interactions in this network were false positives (including 99% of interactions observed exactly once and 21% of interactions observed twice; **Fig. 2b**). The

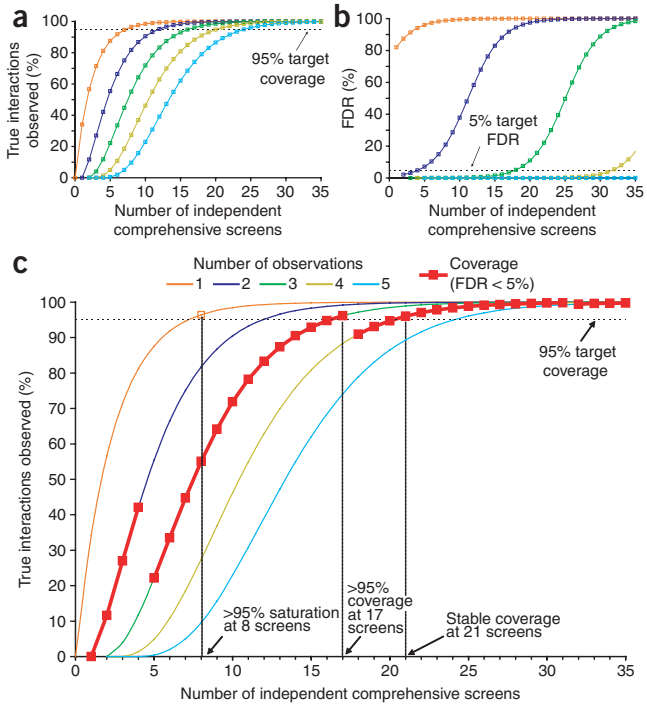

Figure 2 | Analysis of the coverage and saturation of the fly interactome as a function of the number of independent screens. (**a**) The percentage of true interactions that are observed the indicated number of times as a function of the number of times they are tested with independent assays. (**b**) The FDR of interactions that are observed the indicated number of times as a function of the number of times they are tested with independent assays. To achieve FDR $<$ 5% interactions should be observed at least twice when tested with $<$5 independent assays, and at least three times when tested with 5–17 assays. (**c**) The effective coverage at FDR $<$ 5% is shown by embedding the observation threshold from the data presented in **b** into the curves shown in **a**. Although saturation is achieved after 8 screens, 21 screens are required for 95% coverage at FDR $<$ 5%.

Table 1 | Summary of the features and performance of the different strategies

Strategy	Data presented	Pooling	Repeated screens	Pooling sensitivity (%)	Thresholding and prioritization	Prediction	Fruit fly interactome: Intermediate (50%) coverage cost	Fruit fly interactome: Complete (95%) coverage cost	Human interactome: Intermediate (50%) coverage cost	Human interactome: Complete (95%) coverage cost
Basic serial										
1	**Fig. 2c**	No	NA	NA	No	No	7.5M	19.9M	18.9M	51.8M
Pooling										
2.1	**Fig. 3a**	Yes	4	40	No	No	1.6M	4.4M	4.2M	11.3M
2.2	**Fig. 3a**	Yes	1	20	No	No	1.6M	4.9M	3.9M	12.3M
2.3	**Fig. 3a**	Yes	1	40	No	No	1.4M	4.1M	3.5M	10.4M
2.4	**Fig. 3a**	Yes	1	100	No	No	1.4M	3.7M	3.5M	9.5M
Thresholding	**Fig. 3b**									
3.1	**Fig. 3a**	Yes	1	20	Yes	No	443K	1.9M	1.1M	4.8M
3.2	**Fig. 3**	Yes	1	40	Yes	No	391K	1.7M	969K	4.3M
3.3	**Fig. 3b**	Yes	1	100	Yes	No	354K	1.5M	916K	3.9M
Prediction	FPR, FNR, FDR (%)									
4.1	10, 20, 99.2	**Fig. 3b**	Yes	1	40	Yes	Yes	249K	1.6M	611K	4.1M
4.2	1, 50, 95.0	**Fig. 3b**	Yes	1	40	Yes	Yes	111K	1.4M	293K	3.6M
4.3	5, 20, 98.3	**Fig. 3b**	Yes	1	40	Yes	Yes	126K	1.3M	313K	3.3M
4.4	1, 20, 92.2	**Fig. 3b**	Yes	1	40	Yes	Yes	28K	925K	69K	2.3M

Interaction costs are given in units of total number of plates (K, thousands; M, millions) required for 50% or 95% coverage. When 95% coverage is achieved more than once, the greatest cost is presented.
NA, not applicable.

false positives predominate because, although the 0.2% FPR seems low, the number of non-interacting protein pairs is far in excess of the number of true interactions.

To ensure an overall FDR < 5%, we found that every interaction must be reported by at least three independent assays. After eight screens, 55% of the interactome was covered under these conditions. We achieved the coverage goal of 95% only after 21 comprehensive pair-wise screens (**Fig. 2c**). We observed this overall outcome, that the number of experiments required to reach full coverage is two to three times that required to reach saturation, over a range of error parameters (**Supplementary Table 1** online). Clearly, completing the interactome map under these conditions is impractical, as it would require testing 92 million protein pairs 21 separate times.

To reduce the number of tests, assays such as Y2H typically use pooled screens in which a single protein 'bait' is tested for interaction against pools of protein 'preys' (phase I)[37]. For pools that test positive, pairwise tests are immediately conducted between the bait and each prey in the pool (phase II; this second phase can also be conducted by sequencing[3,5]). The benefit of pooling is that large numbers of potential interactions can be sampled at relatively low cost. This comes at the expense of FNR, as the chance a true interaction is missed in the pool is higher than the chance it would be missed by direct pairwise tests[37]. Through simulation, we found that basic two-phase pooling (pooling strategies 2.1–2.4, in which we assumed the number of screens and pooling sensitivity as indicated in **Table 1**) does indeed achieve a four- to fivefold reduction in coverage cost over pairwise testing (~4 million plates for pooling compared to ~20 million plates for basic serial). However, assuming the rate of interaction screening of a typical laboratory (for example, ~2,400 plate-matings per person per year), pooled screens would still require approximately 1,700 person-years to complete the *Drosophila* protein-interaction network.

Advanced mapping strategies

We next considered an interaction mapping strategy that, rather than treating all protein pairs equally, maintains a rank-ordered list of pairs according to their probabilities of interaction (thresholding strategies 3.1–3.3, in which we assumed the pooling sensitivity as indicated in **Table 1**). All probabilities started at the background frequency of interaction for random protein pairs (as for basic-serial and pooling strategies). We initially tested protein interactions using pooled screening and, after each two-phase pooled experiment, increased the probabilities for interactions that were observed and decreased the probabilities for interactions that we tested but did not observe. Unlike previous strategies, however, we declared protein pairs with probability greater than an upper threshold (95%) to be definite 'interactors' and removed them from subsequent testing (**Fig. 1b**). Likewise, we declared interactions with probability beneath a lower threshold to be 'non-interactors' and also removed them from further consideration. The motivation for thresholding was to more quickly exclude the overwhelming number of non-interacting protein pairs. Finally, we defined 'candidate interactions' as those with probabilities between the upper threshold and background. When candidates were available they were always tested immediately using pairwise assays, before resorting to pooling, until their probabilities were pushed above the upper threshold or below background. The motivation for prioritizing candidate interactions was to more quickly cover the interactions likely to be positive. Overall, thresholding resulted in more than a twofold cost reduction compared to pooling (**Table 1** and **Fig. 3a**).

Lastly, we considered whether computational prediction of interactions, based on prior knowledge and data, could hasten the time to interactome completion (prediction strategies 4.1–4.4, in which we assumed the FPR, FNR and FDR as indicated in **Table 1**). Various prediction methods have been proposed based on evolutionary conservation[38–40], that is, transfer of interaction measurements from one species to another or based on integrating

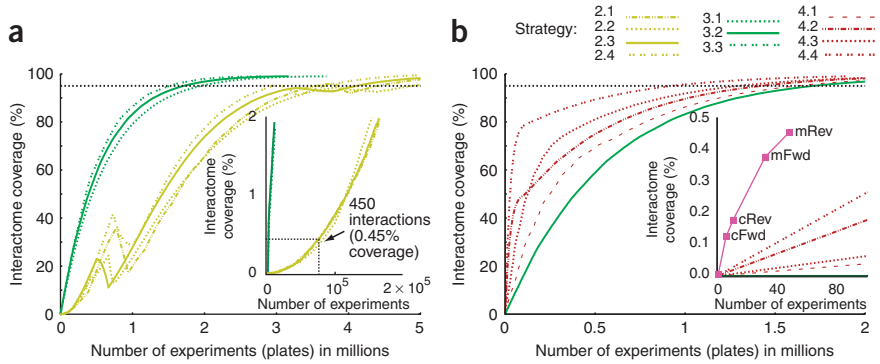

Figure 3 | Fly and human interactome coverage costs for different experimental strategies. (**a**) Comparison of the pooling strategies 2.1–2.4 with thresholding strategies 3.1–3.3, which combine pooling with direct experiments based on thresholding and prioritization. Inset, close-up view showing the number of plates required to add the first 450 interactions to the map using pooling. (**b**) Performance of the prediction strategies 4.1–4.4 over a range of FPR, FNR and FDR of the predictions. Inset, close-up view including an experimental proof-of-principle based on predictions from network conservation (cFwd, cRev) or multiple types of evidence (mFwd, mRev). Fwd and Rev denote the experiments performed using forward and reverse Y2H assays, respectively.

conservation with additional features such as co-expression and co-annotation[41–46]. Such predictions impact the experimental design by setting the initial probabilities of interaction for each protein pair in lieu of background probabilities. In the prediction strategy, we explored the effect of setting these probabilities using theoretical prediction methods simulated over a range of predetermined prediction accuracies (a range of different values for FPR, FNR and corresponding FDR of the predictions). We found that even predictors with very high FDRs could have a major impact on the mapping cost (**Fig. 3b** and **Table 1**). For example, a predictor with 92.2% FDR gave a fourfold reduction in cost over pooling, with a >50-fold reduction in cost to achieve 50% coverage and hundreds-fold cost savings in the early stages of mapping. Moreover, the 92.2% FDR means that even a predictor that makes 12 false predictions for every true one can lead to a major reduction in the cost of interactome completion. The best prediction method required approximately 385 person-years to achieve 95% coverage of the *Drosophila* protein network and 12 person-years to achieve 50% coverage. Thus, while obtaining full coverage of an interactome map may still be some years away, a draft scaffold providing half coverage might be feasibly achieved by a team of about 12 technicians working for one year.

From theory to practice: an experimental proof of concept

Given the good performance of the prediction strategy in simulations, we explored an experimental implementation in which *Drosophila* protein interactions were predicted using a method based on cross-species analysis[38] (**Fig. 4a**). According to this method, existing protein-interaction networks in baker's yeast, nematode and fruit fly are aligned based on sequence similarity to identify conserved interaction clusters, and these alignments are used to transfer interactions that have been observed in some species but not yet in others (**Fig. 4b**). A total of 1,294 interactions had been previously predicted using this method[38], each of which we prioritized as a candidate with high initial probability (92.4%) based on an estimated FDR of 7.6% (**Supplementary Methods** online).

As this prior probability was much greater than the background probability of other protein pairs (0.1%), we began by using the pairwise LexA Y2H assay[47] to test all 606 predictions for which sequence-verified clones were available. Of these, 136 tested positive for interaction and 470 tested negative. After each 96-well plate test (seven plates total), we updated the interaction probabilities, resulting in an increase to >99.9% for pairs testing positive and a decrease to 90.5% for pairs testing negative. As the 136 'positives' now had probability greater than the upper threshold (95%), all of these could be added to the interactome map and removed from subsequent testing.

Although the remaining 470 predictions had tested negative once, their high probability (90.5%) still prioritized them as candidate interactions. Therefore, as dictated by the prediction strategy, we tested these pairs immediately using a second assay type. For this second assay, we ran the LexA Y2H assay in a 'reverse' orientation in which the two proteins cloned as bait and prey, respectively, were exchanged as prey and bait. We tested 251 of the 470 predictions for which sequence-verified clones were available in the 'opposite orientation'. This resulted in 35 pairs testing positive, elevating these interactions to probability >99.9% and adding them to the map. The pairs that tested negative in the reverse orientation were downgraded to 88.1% probability. Overall, after performing the Y2H assay in both forward and reverse orientations, we identified 171 new interactions out of 606 protein pairs, a success rate of 28% (**Supplementary Table 2** online). Although we ended our experiment at this point, the prediction strategy could be continued by next testing the 'double negatives' (pairs testing negative in both 'orientations' of LexA Y2H) using a third type of assay such as Gal4-based Y2H.

A means of predicting additional protein interactions is to probabilistically integrate many different lines of evidence into a single classifier[41–46]. Thus, we applied a machine learning–based classifier for predicting interactions that combined many relevant features including gene expression, domain-domain interactions, conserved protein-protein interactions, genetic interactions and shared gene annotations (**Supplementary Methods**). We used this approach to generate 24,798 high-confidence predictions. We randomly selected 2,047 of these for testing using the forward-orientation Y2H assay and, as above, retested the negative pairs using the reverse-orientation Y2H assay (for which clones were available). In total, using this procedure we added 279 new high-confidence interactions to the map, a 13.6% success rate (**Supplementary Table 2**). Combining both conservation-based and multiple evidence–based predictions, we added 450 new protein-protein interactions to the *Drosophila* interactomes map based on data from experiments using 47 96-well plates (**Fig. 3**). To establish the background rate of interaction, we also tested 2,354 randomly chosen pairs, 72 of which were positive in the Y2H assay, yielding a 3% background rate. These results show that both types of prediction are highly enriched for true interactions. Note that even if all predicted interactions were true, in our model the expected confirmation rate would be limited by the false negative rate of the Y2H assay, equal to 1 − FNR = 33%.

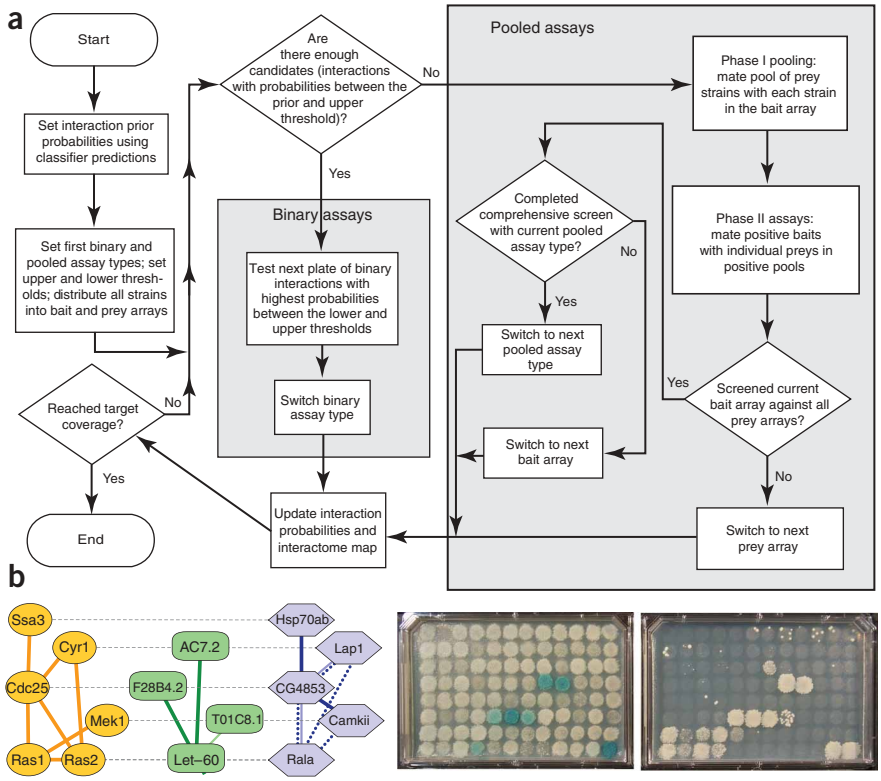

Figure 4 | Design and implementation of the prediction strategy for mapping the interactome. (**a**) State diagram for the prediction strategy, which combines interaction predictions with direct and pooled experiments to reduce the intermediate and total costs of interactome mapping. (**b**) Making conservation-based interaction predictions as reported in reference 38. Colored nodes and links represent proteins and protein-protein interactions, respectively, measured for yeast (orange), worm (green) or fly (blue). Gray horizontal dashed lines connect sequence-similar protein families across the three species. Representative plates are shown for tests of conservation-based predictions using the *lacZ* reporter (5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal) cleavage assay; left) or the *LEU2* reporter (LexA-based system; right).

Testing the conditional independence between assay types

An underlying assumption of our simulations is that different assay types are conditionally independent. To examine the extent to which this assumption holds, we compared Y2H data for protein pairs tested in both forward and reverse orientations. Overall, we obtained Y2H test results in both orientations for 309 conservation-based predictions (including data reported above as well as that from additional tests; **Supplementary Data** online). Of these, we observed 58 positives in the Y2H assay in the forward orientation and 50 positives in the reverse orientation, for an average positive rate of $r = 17\%$ ($(58 + 50) / (309 \times 2)$). We identified 15 positives in both orientations, representing 4.9% of the tests. Assuming all predictions are true interactions, this percentage is very close to that predicted by conditional independence, for which 3.1% of tests are expected to be positive in both orientations (r^2). If some predictions are not true as expected, the percentages come into even better agreement; for example, a prediction FDR of 20% predicts that 4.8% positives would arise in both orientations. We performed a similar analysis on 1,572 combined-evidence predictions that we tested in both orientations, leading to similar agreement with the conditional independence assumption.

DISCUSSION

The interactions predicted by cross-species conservation were at least as accurate as we had assumed in our simulations. In contrast, their power to prioritize interactions is dependent on the network coverage in other species, and the long-term viability of this approach will depend on obtaining greater numbers of predictions than the 1,294 that are currently available. As interactome maps progress across an ever-widening array of species, these maps might be dynamically cross-compared to continually generate sufficient numbers of candidate interactions for testing. The second set of predictions, made by integrating various lines of evidence, had a lower success rate than the predictions based solely on conservation. Their potential utility is higher, however, as the number of available predictions is nearly 20 times that of the conservation-based predictions and could be increased further by including lower-confidence predictions. Even with a lower success rate, the performance of the integrated classifier was superior to the best theoretical predictor we simulated.

Predictions lead to a lower interactome mapping cost for two reasons. First, predicted protein pairs are much more likely than arbitrary pairs to be true. Second, protein pairs with high prior probabilities do not require repeated positive measurements to confirm them as true interactions. Both effects underlie the finding that 450 new predicted interactions could be added to the interaction map using just 47 microtiter plates. In contrast, the pooling strategy would require nearly 10^5 plates to add this number of interactions to the map.

One might intuitively object that, rather than test predicted interactions, a better strategy would focus on the 'novel' areas of the interactome that have never before been suggested by any species or dataset. The problem with such an approach is that it would very quickly produce an interactome map with a very high error rate. Conversely, the rationale behind the thresholding and prediction strategies is that one should first clean up the map by validating predicted interactions using real experiments, and only then resort to testing random protein pairs in pools.

A second objection might be that prioritizing candidate interactions requires the corresponding Y2H baits and preys to be rearrayed in microtiter plates in different orders over the course of an interaction mapping project. Although we did not include the cost of rearraying in our analysis, in our laboratory (Finley) these costs are greatly alleviated through robotic transfer systems. Certainly, failure to rearray leads to an about fourfold increase in cost and an about tenfold increase in the early stages of mapping (**Table 1**).

Regardless, mapping the interactome remains a daunting task. Our study makes it clear that achieving 95% coverage of an interactome requires many more screens than one pass through all pools or over all protein pairs. If complete coverage is to be

obtained in the near future, it will be necessary to invoke better strategies for experimental design, technologies reporting fewer false negatives or both. In terms of experimental design, we showed that the cost of completion is reduced substantially by carefully ordering pooled screens. In terms of technology, our study underscores the importance of decreasing the FNR or of different assays that provide independent samples of a protein pair. Even if the error rates are lower than assumed here, advanced mapping strategies are still likely to be worthwhile (**Supplementary Table 1**). Here we used two types of Y2H assay, forward and reverse orientations, to obtain multiple samples which appear largely independent. If the assays were partially dependent, multiple tests might still be worth the cost as long as they were not perfectly correlated (and the dependence could be handled quantitatively using a statistical model). In our study, the conditional independence assumption leads to a best-case scenario or lower bound on the number of interaction tests that will likely be required to achieve full coverage of an interactome. Additional work will be needed to better characterize the relative dependencies among the wide range of other interaction assays that are now available: if the presently available assays are highly dependent, then the required number of tests will be greater than we estimated here.

METHODS

Simulation procedure. 'True' reference interactomes for fly and human were generated by random sampling of interactions from the set of all possible pairs of proteins using the interaction probabilities in the String database[45]. Protein pairs not included in the String database were sampled using a low background probability such that the total number of interactions in the sampled interactomes agreed with current estimates of interactome sizes[29] ($\sim$100,000 fly interactions and $\sim$260,000 human interactions). The detectability of each protein pair was independently sampled for each new assay type (representing a new type of measurement technology or new bait/prey orientation) using a 66% FNR for true interactions and 0.2% FPR for false interactions (corresponding to 82% FDR). Once an interaction was defined as detectable or undetectable, direct pairwise experiments were assumed to be 100% reproducible for a given protein pair and assay. For pooled assays, each detectable interaction in the sample space of a pool was assumed to be observed in the pool with probability equal to the 'pooling sensitivity' (**Table 1**). Pools with at least one observed interaction were declared positive. For each strategy, after every 1,000 experiments the mapped interactomes were compared to the 'true' interactomes and the coverage and FDRs were recorded.

Yeast two-hybrid test of predicted interactions. We used the LexA-based Y2H mating assay[47] using sequence-verified clones as previously described[35] (**Supplementary Methods**).

Data availability. The International Molecular Exchange Consortium through IntAct[48]: IM-9552 (new protein interactions identified). The data are also available at *Drosophila* Interactions Database: Finley YTH v3.0.

Additional methods. Descriptions of the interaction probability model, the combined-evidence method for interaction prediction, the computation of thresholds and the Y2H test protocol are available in **Supplementary Methods**.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank S. Bandyopadhyay for critical reading of the manuscript, I. Bronner, K. Gulyas, B. Mangiola and H. Zhang for expert technical assistance with the two-hybrid assays, and R. Karp and R. Sharan for discussions of earlier versions of this work. This work was supported by US National Institutes of Health grants RR018627, GM070743 and HG001536.

AUTHOR CONTRIBUTIONS

A.S.S. and T.I. formulated the probabilistic model and performed the simulations. J.Y., K.R.G. and R.L.F. generated all new reported Y2H data. A.S.S., R.L.F. and T.I. wrote the paper.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Fields, S. High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.* **272**, 5391–5399 (2005).
2. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
3. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
4. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
5. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
7. Suzuki, H. *et al.* Protein-protein interaction panel using mouse full-length cDNAs. *Genome Res.* **11**, 1758–1765 (2001).
8. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
9. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
10. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
11. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
12. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
13. Pokholok, D.K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
14. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
15. Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
16. Collins, S.R., Schuldiner, M., Krogan, N.J. & Weissman, J.S. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.* **7**, R63 (2006).
17. Bao, L. *et al.* Combining gene expression QTL mapping and phenotypic spectrum analysis to uncover gene regulatory relationships. *Mamm. Genome* **17**, 575–583 (2006).
18. Chesler, E.J., Lu, L., Wang, J., Williams, R.W. & Manly, K.F. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat. Neurosci.* **7**, 485–486 (2004).
19. Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.* **2**, e172 (2006).
20. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
21. Rain, J.C. *et al.* The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
22. Parrish, J.R. *et al.* A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol.* **8**, R130 (2007).
23. LaCount, D.J. *et al.* A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438**, 103–107 (2005).
24. Uetz, P. *et al.* Herpesviral protein networks and their interaction with the human proteome. *Science* **311**, 239–242 (2006).
25. von Brunn, A. *et al.* Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS ONE* **2**, e459 (2007).

26. Lander, E.S. & Waterman, M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
27. Weber, J.L. & Myers, E.W. Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–409 (1997).
28. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
29. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006).
30. Lappe, M. & Holm, L. Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.* **22**, 98–103 (2004).
31. Cusick, M.E., Klitgord, N., Vidal, M. & Hill, D.E. Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14** (special issue 2), R171–R181 (2005).
32. Kocher, T. & Superti-Furga, G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat. Methods* **4**, 807–815 (2007).
33. Parrish, J.R., Gulyas, K.D. & Finley, R.L. Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17**, 387–393 (2006).
34. Deane, C.M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349–356 (2002).
35. Stanyon, C.A. *et al.* A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol.* **5**, R96 (2004).
36. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
37. 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).
38. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
39. 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).
40. Boulton, S.J. *et al.* Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127–131 (2002).
41. Ben-Hur, A. & Noble, W.S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21** Suppl 1, i38–i46 (2005).
42. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
43. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
44. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15**, 945–953 (2005).
45. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
46. Yu, H., Paccanaro, A., Trifonov, V. & Gerstein, M. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**, 823–829 (2006).
47. Finley, R.L. Jr & Brent, R. Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc. Natl. Acad. Sci. USA* **91**, 12980–12984 (1994).
48. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).