

**Bridging Structural Biology and Genomics:  
Assessing genome-wide protein-protein interaction datasets using the structures of  
known protein complexes**

Aled M. Edwards<sup>1,4</sup>, Bart Kus<sup>1</sup>, Ronald Jansen<sup>2</sup>, Dov Greenbaum<sup>2</sup>, Jack Greenblatt<sup>1</sup> and  
Mark Gerstein<sup>2,3,4</sup>

<sup>1</sup>Banting and Best Department of Medical Research  
University of Toronto  
C.H. Best Institute  
112 College St.  
Toronto, Ontario  
Canada M5G 1L6

<sup>2</sup>Department of Molecular Biophysics and Biochemistry

<sup>3</sup>Dept. of Computer Science

Yale University  
266 Whitney Avenue  
New Haven, CT 06520

<sup>4</sup>Corresponding authors

## ***Abstract***

There is a large effort to map protein-protein interactions on a genome-wide scale. The utility of the resulting interaction networks depends on the reliability of the experimental methods and coverage of the approaches. Known macromolecular complexes provide a defined and objective set of protein interactions to which biochemical and genetic data can be compared and validated. We show a significant fraction of the protein-protein interactions in genome-wide datasets, as well as many of the individual interactions reported in the literature, were inconsistent with the known 3D structures of three recent complexes (RNA polII, Arp2/3 and the proteasome). Furthermore, comparison amongst genome-wide datasets and between them and a larger (but less well resolved) group of 174 complexes also shows marked inconsistencies. Finally, individual interaction datasets, being inherently noisy, are best used when integrated together, and we show how simple Bayesian approaches can combine them, significantly decreasing error rate.

## **Background**

Protein-protein interactions constitute one of the most important regulatory mechanisms in cells; they underlie intercellular communication, signal transduction, and the regulation of gene expression. Indeed, most cellular processes are coordinated by specific protein interactions. The completion of the human genome project has led to the realization that the genome comprises fewer proteins than had been previously believed<sup>1,2</sup>. It is now thought that the complexity underlying the biology of higher organisms may arise not from an increase in the number of proteins, but rather as a result of combinatorial interactions among them<sup>3</sup>.

Many large-scale protein interaction datasets have been published, each attempting to completely characterize the interactome - the set of all protein interactions in a cell<sup>22,23,27,28</sup> (See supplementary material website for a summary of the different techniques, <http://genecensus.org/integrate/interactions>.) These data sets, while extensive, miss many of the interactions, and report spurious interactions as well (e.g. see ref. 28). In order to use these data sets to build robust and statistically significant protein interaction networks, it will be essential to quantify the intrinsic error rates in these experiments. This is challenging yet critical: challenging, in that it is difficult to define a basis protein interaction dataset with which to validate the experimentally-identified interactions; yet critical, because in yeast, for example, there are a possible 18 million possible protein interactions (given 6,000 proteins), of which only a small fraction is biologically relevant. With some many potential interactions, even a very small false positive rate can create a large amount of spurious interactions that swamp out the real ones.

Recently, some papers have attempted to evaluate the quality of these interaction datasets, by either integrating additional annotation, e.g. synexpression among interacting proteins, or through comparison with reference interaction sets<sup>30, 36-37,41</sup>. The conclusions

from these studies were that the interaction datasets contain false positive information and are missing many true protein interactions.

Our paper provides two complementary analyses of the reliability of protein interaction datasets. First, we use information provided through X-ray crystallography of several macromolecular complexes to provide validated interactions that serve as a basis set for comparison with the interaction databases. Second, on a somewhat larger scale, we compared the interactions in the known complexes in the curated MIPS database, with those that had been determined using genome-wide approaches. (See website for more extensive discussion of MIPS complexes).

### **Assessing interactions with the 3D structures of three complexes**

Interactions discovered using structural biology of large complexes provide an excellent test set to assess the quality and coverage of protein interaction data sets. Over the past two years, the structures of several large, conserved protein complexes have been determined, and from these structures a list of verified protein-protein interactions can be extracted. The structures that we selected (RNA polymerase II, the proteasome, and the Arp2/3 complex) were solved independently of the experimental interaction data sets, and the 3D structures were determined after most of the biochemical and genetic studies were published. Thus, the structures provided objective and verified interactions to assess the reliability of biochemical and genetic protein interaction studies.

The 3D structures of RNA polymerase II, the proteasome, and the Arp2/3 complex, which comprise 10, 14 and 7 subunits respectively, were examined to identify stable protein-protein interaction interfaces, defined as more than  $800\text{\AA}^2$  of contacting surface area. This  $800\text{\AA}^2$  threshold was chosen based on the lower limits of the interaction interfaces of binary protein-protein complexes in the structural database (PDB)<sup>4</sup>; the “average” protein interaction interface is  $1600\text{\AA}^2$ <sup>5</sup>.

For RNA polymerase II, of the 45 possible interactions that could exist among its ten subunits ( $10 \times 9 / 2$ ), 13 clear protein-protein interactions were found in the crystal structure<sup>6,7</sup>. Of the 21 possible interactions that could occur among the Arp2/3 subunits, 6 exist in the crystal structure<sup>8</sup>. Of the 91 possible interactions that could occur among the subunits in the proteasome, 14 were observed in the crystal structure<sup>9</sup>. The ribosome was eliminated as a candidate for our analysis because it is largely held together via protein-RNA interactions, confounding attempts to verify all the protein-protein interactions.

An assumption in our analysis is that the protein interactions defined on the basis of the structural biology of stable macromolecular complexes provide an objective set of interactions. This assumption is supported by two observations. First, many macromolecular assemblies -- particularly, RNA polymerase II, Arp2/3 and the proteasome -- are very stable and can be purified to homogeneity without loss of subunits. In fact, the RNA polymerase II complex is even stable in high concentration of

chaotropic agents, such as urea<sup>10</sup>. Therefore, the subunit-subunit interactions found in the crystal structure are unlikely to have formed randomly. Second, many crystallized proteins are enzymatically active. For example, it is known that the crystallized form of the active RNA polymerase II elongation complex has the same repertoire of specific protein interactions as does the native form<sup>11</sup>

### ***False positives and negatives***

For each of the three protein complexes, a list of biochemically and genetically-defined interactions were extracted from genome datasets, as well as from the biochemical and genetic literature. The structural, biochemical and genetic lists were compared to assess overlap, consistency, and the rates of false positive and false negative information. For clarification:

A *false positive* (FP) is defined as an interaction that was documented in one of the datasets but did not exist in the crystal structure.

A *true positive* (TP) is defined as an interaction that was documented in one of the datasets and did, in fact, exist in the crystal structure.

A *false negative* (FN) is defined as an interaction that had been tested experimentally and failed to score as an interaction, but is known to exist in the crystal structure.

A *true negative* (TN) is defined as an interaction that had been tested experimentally and failed to score as an interaction, and is known *not* to occur in the crystal structure.

For each method, we define the false positive rate as the number of false positive interactions reported per total reported interactions (FP/(FP + TP)), and the false negative rate as the fraction of known interactions that are not identified (FN/(FN + TP)).

### ***Comparing structural and biochemical interactions in the literature***

Prior to the elucidation of the 3D structures for the three above-mentioned complexes, many of the subunits had been subjected to *in vitro* and *in vivo* binding experiments, chemical cross-linking and far-western analysis (see Table 1 for an overview of data relating to RNA polymerase II). The aim of these experiments, in the absence of crystallographic information, had been to deduce the protein interaction map for a given complex. With 3D structures now in hand, we compared the subunit-subunit interactions defined by these experiments with the interactions revealed in the crystal structures.

In one series of experiments with RNA polymerase II subunits, every recombinant subunit was cloned into baculoviruses as either a GST-tagged or an untagged protein. Every combination of individual GST-tagged subunits and non-tagged subunits were co-expressed and precipitated, generating a comprehensive set of inter-subunit pairwise interactions<sup>12-16</sup>. This is a common and accepted strategy ('pull-down') to define binary protein-protein interactions. Remarkably, 61% of the interactions in these experiments were false positives. Some of the false positive interactions were even "validated" by several different biochemical approaches<sup>12,16</sup> and, in a few cases,

spurious contact sites were also mapped in detail<sup>16-17</sup>. The false negative rate was 38%.

Biochemical interactions were also extracted from a series of far-western experiments in which the RNA polymerase II subunits were resolved by denaturing gel electrophoresis, transferred to a solid support, renatured and over-laid with various purified subunits. 44% of the interactions in the RNA polymerase II crystal structure were not observed with this method whereas 53% of the reported interactions were false positives<sup>16, 17</sup>.

Chemical cross-linking<sup>18</sup> proved a more effective method for identifying biochemical interactions between RNA polymerase II subunits. The cross-linking data contained no false negatives and 10 true positives; there were 7 false positives (41%) although for 3 of them we did find subunit-subunit contacts in the crystal structure with a contact area below our 800 Angstrom<sup>2</sup> cutoff.

Cross-linking approaches were also used to study the Arp2/3 and proteasome complexes. For the Arp2/3 complex, 86% of the interactions found by cross-linking were present in the crystal structure<sup>19-20</sup>, while the false negative rate was 50%. For the proteasome, only 5 of the 14 interactions (36%) reported in the cross-linking study were not present in the crystal structure<sup>21</sup> whereas the false negative rate was 79%. The relative success of cross-linking approaches (25-41% false positive rate) suggests that this method may be preferable to others, particularly for relatively stable macromolecular complexes. Unfortunately, chemical cross-linking may be challenging to perform on a genomic scale, and the efficacy of cross-linking studies is dependent on the particular protein complexes and the juxtaposition of the appropriate amino acid residues that react with the cross-linking reagent.

### ***Comparing structural and genetic interactions***

The yeast two-hybrid (Y2H) method identifies interactions between two recombinant proteins expressed in yeast cells. Current estimates suggest that the false positive rate could range from 47 to 91% in the genome-wide Y2H screens<sup>24</sup>. Our analysis of the structural information for RNA pol II, Arp2/3 and the proteasome was unable to contribute to the estimation of the false positive rate in Y2H screens for two reasons. First, the structural information was derived from yeast complexes, and therefore we could not rule out the possibility that the interaction between two subunits could be bridged by other, endogenous yeast components. Second, the Y2H method may have also been capable of detecting transient interactions not revealed in the structure, but that might exist during complex assembly.

We were able to use the structural information to estimate the false negative rates in Y2H screens. Our analysis was restricted to the Arp2/3 and proteasome complexes because the RNA polymerase II subunits were omitted in the Y2H (they are known to score positively in the absence of a protein partner). The individual proteasomal subunits were screened against the rest of the yeast proteome by Cagney and colleagues<sup>25</sup>, and 12

interactions between the subunits were revealed. Of these 12 interactions, eight were also found in the crystal structure and have contact surfaces that exceeded  $800\text{\AA}^2$ . Since the crystal structure revealed 14 interactions between subunits, this directed Y2H approach missed six out of fourteen interactions (43%). When interactions among Arp2/3 subunits were tested directly using Y2H screens, 71% of subunit-subunit interactions present in the Arp2/3 crystal structure were not detected<sup>26</sup>.

### ***Comparison to genome-wide two-hybrid***

The proteasome and Arp2/3 subunits were also analyzed as part of several genome-wide Y2H screens. In the first, carried out by Uetz and colleagues<sup>27, 31</sup>, five interactions involving proteasome subunits and other proteins were uncovered, but not one interaction between two known proteasomal subunits was found. This data set also did not contain any interactions between Arp2/3 subunits. A subsequent two-hybrid screen carried out by Ito et al.<sup>28</sup> identified 30 interactions between proteasome subunits and non-proteasome proteins, but only one intra-complex interaction. This interaction was, in fact, present in the crystal structure. In order to reduce the extent of false positive information in their dataset, Ito and colleagues created a “core data set” that was filtered (taking only data validated by more than three interaction sequence tags<sup>28</sup>) to reduce false positive information. This core data set eliminated protein-protein interaction data that did not appear at least three times in their screen. This filtering process eliminated the single correct proteasome interaction from the dataset. The Ito et al. screen correctly identified one interaction between Arp2/3 subunits, and this interaction was maintained in the “core data set”.

In summary, we estimate that the false negative rate in directed two-hybrid screens ranges from about 43-71%, while the rate was even higher in genome-scale applications of the Y2H method.

### ***Large-scale pull-down experiments***

Biochemical interactions can also be discovered using ‘pull-down’ experiments *in vivo*. In these experiments, a protein tag is engineered onto a specific subunit and the gene for the tagged subunit added back into the cell. The tagged protein is then purified from cell lysates using affinity chromatography, and the co-purifying proteins identified. For multi-protein complexes, this method does not provide information about binary protein interactions, but rather describes the collection of proteins that are stably associated with the tagged protein. Gavin et al. and Ho et al. reported the results from large-scale pull-down experiments in two recent publications<sup>22-23</sup>. For the three complexes we studied, the *in vivo* “pull down” method was quite successful in identifying subunits that interact within the complex. In the case of RNA polymerase, two of the ten polymerase subunits were tagged and the co-purifying proteins identified. Half of the subunits known to interact directly with these subunits were detected, for a false negative rate of 50%. When a subset of the Arp2/3 and the proteasome subunits were tagged, there were no false negatives; all interactions present in the crystal structure whose contacting surface

are  $>800\text{\AA}$  were found in these data sets. The *in vivo* ‘pull-down’ experiments seemed to have fewer false negatives than other methods, though the rate for RNA polymerase II was 50%.

In conclusion, we have used a small test set of structure-based interactions to assess the quality of several protein interaction datasets, and have quantified significant sources of error. Can the results of our structural analysis of complexes be generalized? We cannot be sure. With regard to the analysis of the 3D structures of large complexes, it is possible that inter-subunit interactions distinct from those seen in the crystal might occur, but these complexes were selected to be quite stable in nature. It is also likely that the discrepancies between the information in the crystal structure and the biochemical literature results from protocols used in high-throughput studies, which are acknowledged to generate unvalidated data in the interest of speed and efficiency. Even if these 3D complexes represent special cases, our analysis deserves some attention, because it is probably the first objective assessment of the protein interaction datasets.

## **Broadening the comparisons to 174 complexes**

### ***Complexes in the MIPS database***

Our structure-based analysis of the 31 proteins within the three complexes points to the potential for significant errors in large-scale protein interaction datasets. In order to broaden our study, we looked at a larger set of complexes listed in the MIPS database. This comprises complexes that are manually annotated from the yeast literature. From the MIPS complexes catalogue, we identified 174 protein complexes that contain from 2 to 81 proteins each. The structures of the complexes in the MIPS complexes catalog are generally not known; consequently, it is impossible to perform 3D-structure-based analyses. However, we can still use them as a quality control for the genome-wide data sets. Their use broadens our analysis by orders of magnitude.

A theoretical minimum and maximum number of interactions within these complexes can be calculated. For example, the maximum number of interactions would arise if each protein in these complexes interacted with every other in the complex (an unreasonable assumption for the very large complexes). Among the 174 protein complexes that we selected, there would be a theoretical maximum of 8250 interactions among all the proteins. The theoretical minimum number of interactions among the proteins in all the MIPS complexes would be 834, since the lowest possible number of interactions in a complex is the number of proteins itself minus one (when the proteins are arrayed like beads on a string, thus all proteins in the complex are connected to at least one and at most two other protein). Our estimations also presume that the complexes are maintained by protein interactions and not by other molecules, such as RNA in the case of the ribosome. Of course, the number of true interactions among these proteins lies in the middle of this range between minimum and maximum, probably closer to the minimum.

### ***Overlap between MIPS complexes and genome-wide Y2H interaction datasets***

In Table 2, we present a comprehensive overview of the overlap between the MIPS complexes and the other data sets. Each dataset is defined on a different subset of genes, which makes comparison between the sets non-trivial. Moreover, when we look at the intersections between the datasets we find that they are fairly small. (This is illustrated in Figure 2.) This arises for two reasons: (i) the different subsets of genes and, additionally, (ii) the different interactions for the same sets of proteins in each dataset. Consequently, in the analysis we consider both how many interactions and how many genes are shared between the datasets and the complexes. This allows us to separate the two effects from one another.

We can use the Uetz data as an example to see how this works. For example, within the 174 MIPS complexes, there are 871 distinct proteins, and the Uetz data set overlaps the complexes on a subset of 179 proteins. Of these there are 78 proteins involved in 50 known intra-complex interactions -- i.e. where the observed Uetz interactions correspond to a protein-pair in a MIPS complex, generating a true positive. However, these 78 proteins are also involved in 14 interactions that connect different MIPS complexes, thus potentially representing newly discovered or false positive interactions. The remaining 101 of the 179 proteins are only involved in interactions not found in any MIPS complex, again representing new interactions or false-positives.

How many of the MIPS complex interactions do the Uetz data miss? This, of course, depends on how many interactions we count in the MIPS complexes. We can estimate that the 78 proteins that are both in the Uetz data and the MIPS complexes must be connected by between 102 and 333 existing interactions (using minimum and maximum numbers of interactions for complexes; see above and the website for a detailed derivation of these values). Thus, the Uetz data missed between 52 and 283 (102 - 50 and 333 - 50) existing interactions, implying a false negative rate between 51 and 85% (52/102 and 283/333).

One can analyze the other two-hybrid datasets in a similar fashion. The Ito data has a false negative rate between 76 – 96%; the corresponding values for the Ito core data are 45 – 74%.

### ***Overlap between MIPS complexes and genome-wide “pull-down” interaction datasets***

In contrast to the yeast two-hybrid experiments, which represent a method where proteins are studied outside their normal cellular milieu, the comprehensive *in vivo* pull-down experiments capture the protein in its natural state. Consequently, one might expect the pull-down experiments to be somewhat more sensitive and miss fewer existing interactions.

The right panel in Table 2 shows the comparison between the two pull-down experiments (Gavin et al.<sup>22</sup> and Ho et al.<sup>23</sup>) with the MIPS complexes catalog, using a similar approach as for the two-hybrid. Overall, the false negative rates of the Ho and Gavin datasets, are 50% and 85%, thus confirming our expectation. However, these datasets



also contain 1578 (77% of 2042) and 2975 (63% of 4718) interactions between *different* MIPS complexes, thus representing either newly discovered or false positive interactions.

Thus, it appears that the pull-down methods are very sensitive; they discover real interactions at a high rate, but the datasets are likely populated by false positives.

The relatively low false negative rates in the *in vivo* pull-down experiments are remarkable, even with the caveat that the experimenters had the benefit of prior knowledge of the contents of the MIPS database. What might be the reasons that this approach is apparently superior? First, MIPS is largely populated by stable complexes, which are amenable to pull-down methods because they can remain intact for hours during affinity chromatography. Second, the pull-down experiments probe the complexes in an active state; presumably the proteins retain their structural integrity during the purification process. Why does the TAP-tagging approach used in the Gavin et al. paper appear to be superior to that used in the Ho et al. paper in reducing the false negative rate? We speculate that the differences arise from differences in the experimental strategies. In the TAP-tagging approach, the engineered proteins are recombined using homologous recombination into the yeast genome, and are expressed at normal levels from their natural promoter. The TAP-tagged proteins can therefore incorporate into the normal cellular environment. These features all contribute to the relative success of this method. By contrast, the method used by Ho and colleagues, which involved over-expression of the engineered protein, may have increased the level of non-specifically associated proteins, and made the interpretation of the results more difficult.

## Data integration

We have seen that both genome-wide and smaller conventional interaction datasets can be noisy and inaccurate. If the noise is not systematically biased, it should be possible to improve the accuracy of the datasets by integrating the information within them. The recent trend is to combine a variety of experimental information -- including transcription co-expression, co-localization, essentiality data, and functional annotations -- to help predict and validate protein-protein interactions.<sup>30,34-40</sup> In general, data integration, combining multiple, independent sources, should increase the degree to which the known complex interactions are found (i.e. increasing the coverage) while decreasing the number of errors in the data.<sup>34,35,37</sup>

We attempted to quantify the improvements in merging datasets by using the structural information for RNA polymerase II. As described in Table 1, the false positive rates in each of the RNA polymerase II datasets ranged from 41 to 67%. We combined the various interaction datasets for RNA polymerase II using a simple Bayesian network (see table 1 and the supplementary website), where we essentially multiplied the probability of interaction across the various datasets. The combined data has a false positive rate of 30%, which is lower than that for any of the individual datasets. Moreover, the combined data gives a complete coverage of the 45 possible subunit-subunit pairs in the RNA polymerase II structure, whereas the individual interaction studies only cover between 6 and 35 of them. In other words, in contrast to the merged dataset, none of the interaction

studies in the literature provide information on all of these possible interactions. (In particular, the sum of true positives and true negatives is 36 in the combined data versus only 23 or less for the individual datasets.)

Similarly, we can show the benefits of data integration for the genome-wide data sets (Figure 2). First, for each genome-wide dataset (e.g. the Gavin TAP-tag set), we identified the fraction of interactions that contain information about the MIPS complexes. Then we looked at all possible pairwise intersections among the genome-wide sets (e.g. Gavin and Ho) and again looked at how well they found the MIPS complex interactions. Finally, we looked at triplet and quadruplet intersections amongst the datasets (e.g. Gavin and Ho and Uetz). One can see that as we progress from individual datasets to pairwise intersections and then to even higher order overlaps, the accuracy increased (as measured by the fraction of interactions that represent protein pairs within the same MIPS complex). That is, in any individual dataset, only a small fraction of the interactions involved a partner within the same MIPS complex. However, in the final, merged dataset, for instance, the majority of the interactions were consistent with the MIPS data.

The intersections of the datasets tend to be rather small, however, taken together, the datasets are complementary and increase the amount of interactions covered. These observations are consistent with the results from the focused analysis of RNA polymerase II.

## **A summary and a strategy**

Our analysis has two main conclusions. First, biochemical and genetic methods, when applied on a small or large scale, have and will introduce significant numbers of spurious interactions into the datasets of binary protein interactions. One must also conclude that genome-wide screens for individual protein-protein interactions have not yet uncovered the bulk of protein-protein interactions. The validity of the interaction datasets can be improved with the use of structural information about protein complexes. Taken together, it would be difficult to consider mining the current binary interaction data to model intracellular networks.

Second, the *in vivo* pull-down methods, particularly the TAP-tagging approach in yeast, appears to have a sufficiently low false negative rate to warrant a comprehensive analysis of the yeast genome. In such an approach, each protein in every complex would be tagged and associated proteins identified. The false positive results would be minimized by considering only the interactions common to all proteins in the complex.

However, it is not clear if the relative success of the TAP-tagging approach will extend to higher eukaryotes. The TAP-tagging method is particularly powerful in yeast because the tagged gene is inserted into the proper location in the genome and under the normal transcriptional controls. Higher eukaryotes do not have efficient machinery for homologous recombination and therefore making TAP-tagged proteins on a genome-scale is impractical. Expression of the tagged protein in higher eukaryotes will likely have to

be driven from a heterologous promoter and the protein expressed in unnatural amounts. If this is the case, we can perhaps expect to get error rates equivalent to those obtained by Ho and colleagues (15 - 50% false negative and 63 - 77% false positive), who used this over-expression approach in yeast.

Our analysis validates current efforts to create databases that take into account the inherently statistical and error-prone nature of the current protein interaction literature and genome-wide experiments. If this is done correctly, one can achieve tremendous benefit from systematically integrating different datasets. Moreover, the networks being described by the current interaction databases need to be described in some sort of statistical or probabilistic terms in order to be physically realistic. Second, it will be important to derive structures for large protein complexes in order to provide precise and accurate information on protein interactions. Indeed, one of the most enduring values of large-scale structural biology may be the identification of accurate protein-protein interactions via complex crystal structures.

## **Acknowledgments**

The authors wish to acknowledge funding from the NIH Structural Genomics Initiative (P50 GM62413-01) (AME, MG), the Ontario Research and Development Challenge Fund (AME) and the Keck foundation (MG). AME is a CIHR Scientist. JG is a CIHR Distinguished Scientist and an International Scholar of the HHMI.

## **FIGURE 1**

We systematically compared the interaction datasets with the complexes. In general, we do not know the structure of the complexes in the MIPS catalog, so we simply counted which interactions from the genome-wide datasets are between proteins that are in the same MIPS complex. Figure 1 illustrates schematically the situation: the proteins in the dashed circle all belong to the same protein complex, and the dashed lines between the proteins indicate all theoretically possible protein-protein interactions within this complex. The red lines indicate schematically the interactions of a genome-wide dataset.

## **FIGURE 2**

Figure 2 shows the size of the different genome-wide datasets and their possible intersections and their consistency with the MIPS complexes catalog.

The bars (relating to the left y-axis) indicate the number of interactions in each individual dataset and each possible intersection of the datasets (for instance, "Ito + Uetz" contains only interactions that are both within Ito and Uetz). The dotted line (relating to the right y-axis) shows what fraction of these interactions overlap with protein pairs within the same MIPS complex. We show this for both the individual datasets (left), pair-wise intersections of datasets (middle) and higher order intersections (3 or more datasets). As the degree of intersection among the datasets increases, the fraction of interactions within

the same MIPS complex increases. The different datasets are complementary and cover more interactions than each dataset individually.

## Table 1

Table 1 shows the (dis)agreement between the various interaction datasets culled from the literature with the crystal structure of RNA polymerase II. The left side gives statistics for the datasets individually, whereas we group the datasets by method on the right side.

We show the count of TP, FN, TN, and FP, the coverage (that is, the number of unique subunit-subunit pairs the data covers), the false negative rate (FN/(FN + TP)) as well as the false positive rate defined as FP/(TP + FP) for each dataset.

(The table does not contain the Ulmasov and Larkin data referred to in the text since they both contain only one TP, which is already covered by the Yasui data. Note that for 3 FPs in the cross linking data, the subunits are actually in contact in the structure, but with less than our 800 Angstrom<sup>2</sup> cutoff.)

In addition, we show the corresponding statistics for a dataset that combines the individual datasets using a Bayesian procedure. The integration of the various data sources both increases the coverage (45 subunit-subunit pairs for the combined data versus 6 – 35 pairs for the individual data) and lowers the false positive rate (30% for the combined data versus 41 – 67% for the individual data).

If we think of each data source  $k$  as evidence  $e_k$  for or against a protein-protein interaction, we can compute a combined belief in an interaction  $I$  from:

$$O(I | e_1, e_2 \dots e_N) = L(e_1, e_2 \dots e_k | I)O(I)$$

where

$$O(I) = \frac{P(I)}{P(-I)}$$

are the prior odds of interaction  $I$ ,

$$O(I | e_1, e_2 \dots e_N) = \frac{P(I | e_1, e_2 \dots e_N)}{P(-I | e_1, e_2 \dots e_N)}$$

are the posterior odds of interaction  $I$ ,

and

$$L(e_1, e_2 \dots e_k | I) = \frac{P(e_1, e_2 \dots e_k | I)}{P(e_1, e_2 \dots e_k | -I)}$$

is the likelihood ratio.

The prior odds are  $O(I) = 13/(45 - 13)$  since there are 13 interactions among the 45 possible subunit-subunit pairs in the crystal structure.

Assuming conditional independence of the evidence (meaning that each experiment is only dependent on whether there is an interaction and thereafter they are independent of the outcome of the other experiments), the likelihood ratio can be written as:

$$L(e_1, e_2 \dots e_k | I) = \prod_{k=1}^N L(e_k | I)$$

The likelihood ratio  $L(e_k | I)$  for experiment  $k$  can be computed from the experimental data.

## TABLE 2

Table 2 shows the size of the genome-wide datasets and their overlap with the MIPS complexes catalog. The overlap is first shown in terms of the number of proteins, then in terms of the number of interactions (or protein pairs). Note that we consider the maximum number of possible protein pairs between all subunits for the complexes in the MIPS catalog and the Ho and Gavin datasets.

The number of interactions that overlap with the MIPS catalog is very small for each of the individual genome-wide datasets. This is partly due to the fact that the genome-wide datasets cover different proteins than those that the MIPS catalog contains. When we normalize for the protein set shared between the MIPS catalog and the genome-wide dataset in each case, the relative overlap becomes more substantial. For instance, of the 31304 interactions (protein pairs) in the Ho dataset, 4718 are between proteins that the MIPS catalog contains. Of these 4718 interactions, 1743 correspond to protein pairs that are within the same MIPS complex (true positives, *TP*), whereas 2975 are between proteins in different MIPS complexes (these either represent newly discovered interactions or false positives, *FP*). The Gavin dataset missed 300 protein pairs within the same MIPS complex (false negatives, *FN*).

These statistics allow us to measure the sensitivity  $TP/(TP + FN)$  of each genome-wide dataset, that is, the fraction of protein pairs within the same MIPS complex that each genome-wide dataset recovers (normalized for the set of proteins shared with the MIPS catalog). The Gavin dataset recovers the largest amount of interactions (86%), followed by the Ho dataset (50%) and the Y2H datasets (Uetz 15% - 49%, Ito 4% - 24%). The sensitivity of the Ito core dataset (26% - 55%) is higher than for the total Ito dataset.

### Supplementary Information

<http://genecensus.org/integrate/interactions>

### References

1. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
3. Claverie, J.M. Gene number. What if there are only 30,000 human genes? *Science* **291**, 1255-7 (2001).
4. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).
5. Janin, J., Miller, S. & Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol* **204**, 155-64 (1988).
6. Cramer, P. et al. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* **288**, 640-9 (2000).
7. Cramer, P., Bushnell, D.A. & Kornberg, R.D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863-76 (2001).

8. Robinson, R.C. et al. Crystal structure of Arp2/3 complex. *Science* **294**, 1679-84 (2001).
9. Groll, M. et al. Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* **386**, 463-71 (1997).
10. Edwards, A.M., Kane, C.M., Young, R.A. & Kornberg, R.D. Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. *J Biol Chem* **266**, 71-5 (1991).
11. Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A. & Kornberg, R.D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* **292**, 1876-82 (2001).
12. Acker, J. et al. Interactions between the human RNA polymerase II subunits. *J Biol Chem* **272**, 16815-21 (1997).
13. Kimura, M. & Ishihama, A. Involvement of multiple subunit-subunit contacts in the assembly of RNA polymerase II. *Nucleic Acids Res* **28**, 952-9 (2000).
14. Larkin, R.M. & Guilfoyle, T.J. Reconstitution of yeast and Arabidopsis RNA polymerase alpha-like subunit heterodimers. *J Biol Chem* **272**, 12824-30 (1997).
15. Ulmasov, T., Larkin, R.M. & Guilfoyle, T.J. Association between 36- and 13.6-kDa alpha-like subunits of Arabidopsis thaliana RNA polymerase II. *J Biol Chem* **271**, 5085-94 (1996).
16. Yasui, K., Ishiguro, A. & Ishihama, A. Location of subunit-subunit contact sites on RNA polymerase II subunit 3 from the fission yeast *Schizosaccharomyces pombe*. *Biochemistry* **37**, 5542-8 (1998).
17. Miyao, T., Yasui, K., Sakurai, H., Yamagishi, M. & Ishihama, A. Molecular assembly of RNA polymerase II from the fission yeast *Schizosaccharomyces pombe*: subunit-subunit contact network involving Rpb5. *Genes Cells* **1**, 843-54 (1996).
18. Ishiguro, A. et al. Two large subunits of the fission yeast RNA polymerase II provide platforms for the assembly of small subunits. *J Mol Biol* **279**, 703-12 (1998).
19. Mullins, R.D., Stafford, W.F. & Pollard, T.D. Structure, subunit topology, and actin-binding activity of the Arp2/3 complex from *Acanthamoeba*. *J Cell Biol* **136**, 331-43 (1997).
20. Mullins, R.D. & Pollard, T.D. Structure and function of the Arp2/3 complex. *Curr Opin Struct Biol* **9**, 244-9 (1999).
21. Kopp, F. et al. Subunit arrangement in the human 20S proteasome. *Proc Natl Acad Sci U S A* **94**, 2939-44 (1997).
22. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
23. Ho, Y. et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-3 (2002).
24. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research? *Genome Res* **11**, 1971-3 (2001).
25. Cagney, G., Uetz, P. & Fields, S. Two-hybrid analysis of the *Saccharomyces cerevisiae* 26S proteasome. *Physiol Genomics* **7**, 27-34 (2001).
26. Zhao, X., Yang, Z., Qian, M. & Zhu, X. Interactions among subunits of human Arp2/3 complex: p20-Arc as the hub. *Biochem Biophys Res Commun* **280**, 513-7 (2001).
27. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-7 (2000).

28. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
29. Mewes, H.W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **28**, 37-40 (2000).
30. Jansen, R. et al. Relating whole-genome expression data with protein-protein interactions. *Genome Research* **12**, 37-46 (2002).
31. Cho, R. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
32. Schwikowski, B. et al. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1242-3 (2000).
33. Walhout, A. et al. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **17**, 88-94 (2000).
34. Gerstein, M. et al Proteomics. Integrating interactomes. *Science* 295 284-7 (2002)
35. Jansen, R. et al Integration of Genomic Datasets to Predict Protein Complexes in Yeast *Journal of Structural and Functional Genomics* (In Press )
36. Kemmeren, P. et al Protein interaction verification and functional annotation by integrated analysis of genome-scale data *Mol Cell* 9 1133-43 (2002)
37. von Mering, C. et al Comparative assessment of large-scale data sets of protein protein interactions *Nature* 417 399-403 (2002)
38. Deane CM et al Protein interactions: Two methods for assessment of the reliability of high-throughput observations *Molecular and Cellular Proteomics* (In Press)
39. Ge H, et al Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae* *Nat Genet* 29 482-486 (2001)
40. Tong AH, et al A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules *Science* 295 321-4 (2002)
41. Aloy, P. & Russell, R.B Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A* 2002 99 5896-901 (2002)
42. Pearl, J Probabilistic reasoning in intelligent systems, Morgan Kaufmann Publishers, San Francisco (1988)

**Figure 1**

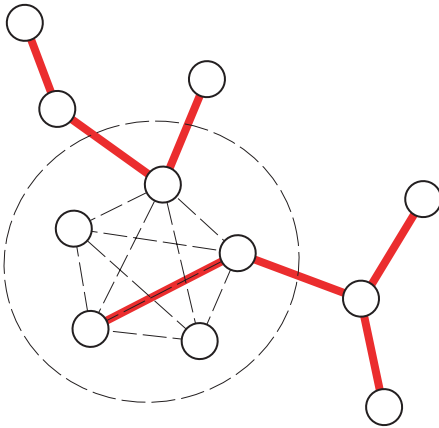
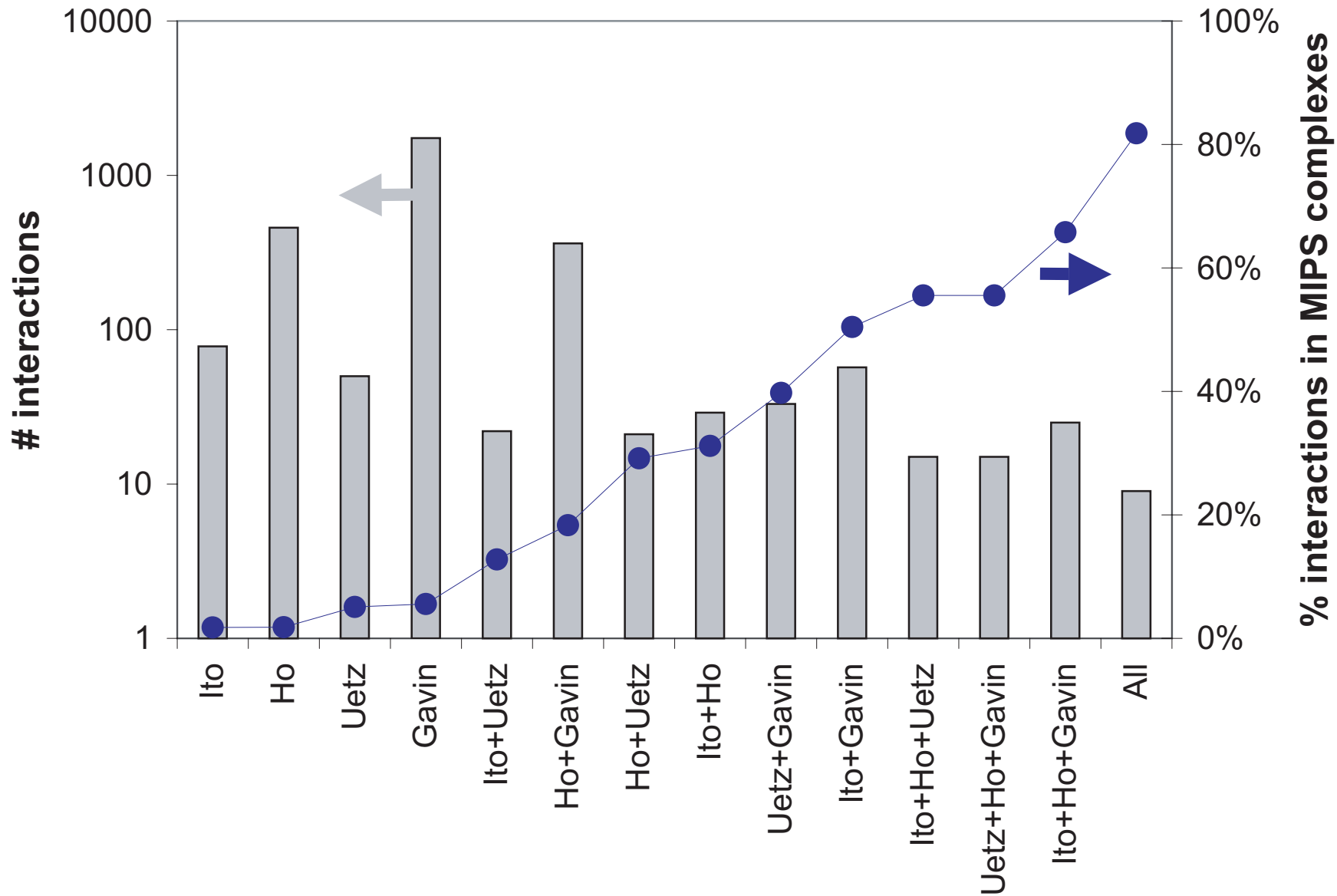




Figure 2



**Table 1**

	Dataset								Method		
	Miyao et al. far-west'n (Rpb5 as probe)	Ishiguro et al. cross linking	Ishiguro et al. far-west'n (Rpb6/8/10/11/12 as probe)	Acker et al. GST-pulldown (X and GST-Y)	Acker et al. GST-pulldown (GST-X and Y)	Kimura et al. pulldown	Yasui et al. far west'n	Combined	Pulldown	Far western	Cross linking
TP	2	10	6	6	7	2	1	7	15	9	10
FN	3	0	2	4	3	2	2	6	9	7	0
TN	6	3	17	14	16	2	2	29	32	25	3
FP	4	7	5	11	9	3	1	3	23	10	7
Coverage	15	20	30	35	35	9	6	45	35	20	36
False negative rate FN/(FN + TP) [%]	60	0	25	40	30	50	67	46	38	44	0
False positive rate FP/(TP + FP) [%]	67	41	45	65	56	60	50	30	61	53	41

## Table 2

Dataset	yeast two-hybrid			<i>in vivo</i> pull-down	
	Uetz	Ito	Ito core	Ho	Gavin
<b># proteins</b>	1044	3278	786	1578	1361
# proteins overlapping with MIPS	179	453	142	331	472
# proteins linked to partner is in same MIPS complex	78	121	73	214	436
# not linked to MIPS partner	101	332	69	117	36
<b># interactions</b>	981	4393	754	25333	31304
# interactions between proteins, both in MIPS catalog	64	177	51	2042	4718
# interactions between proteins in same MIPS complex (TP)	50	78	46	464	1743
# interactions between proteins in different MIPS complexes (FP or new)	14	99	5	1578	2975
<b># MIPS protein pairs missed (FN)</b>	52 - 283	250 - 1682	38 - 131	464	300
<b>FN/(TP + FN) (false negative rate)</b>	51% - 85%	76% - 96%	45% - 74%	50%	15%