# Literature-curated protein interaction datasets

Michael E Cusick[1,2,9], Haiyuan Yu[1,2,9], Alex Smolyar[1,2], Kavitha Venkatesan[1,2,8], Anne-Ruxandra Carvunis[1–3], Nicolas Simonis[1,2], Jean-François Rual[1,2,8], Heather Borick[1,2,8], Pascal Braun[1,2], Matija Dreze[1,2], Jean Vandenhaute[4], Mary Galli[5], Junshi Yazaki[5,6], David E Hill[1,2], Joseph R Ecker[5,6], Frederick P Roth[1,7] & Marc Vidal[1,2]

**High-quality datasets are needed to understand how global and local properties of protein-protein interaction, or 'interactome', networks relate to biological mechanisms, and to guide research on individual proteins. In an evaluation of existing curation of protein interaction experiments reported in the literature, we found that curation can be error-prone and possibly of lower quality than commonly assumed.**



An essential component of systems biology is discovery of the network of all possible physical protein-protein interactions (PPIs), the 'interactome' network[1–3]. There are two complementary ways to obtain comprehensive PPI information. One is to systematically test all pairwise combinations of proteins for physical interactions at proteome scale with a high-throughput assay[3]. The alternative is to curate all publications in the literature, each describing one (or a few) PPI(s) assayed at low throughput[4], and then make the curation accessible in interaction databases. As neither strategy can come close to allowing us to discover the full interactomes yet[5–7], the matter arises as to which strategy can best fill in the missing pieces.

## High-throughput protein interaction assays

Two approaches are in frequent use for high-throughput mapping of protein interactions at proteome scale. Yeast two-hybrid assays attempt to identify binary interactions[8,9], whereas co-affinity purification followed by mass spectrometry identifies presence in a protein complex[10] but may not accurately determine the binary interactions between proteins within a complex[7]. Other technologies exist for mapping both binary interactions and presence in the same complex[11], but none can yet be routinely scaled up for high-throughput assays, although recently, a protein complementation assay allowed a large-scale mapping of the yeast interactome[12].

## Curating protein interactions

Manual curation of protein interactions from literature began with pioneering curation for the yeast *Saccharomyces cerevisiae* by the Yeast Proteome Database (YPD)[13]. Those early efforts demonstrated that effective curation was possible and also broadly aimed to capture all types of functional and genomic information, not only PPIs. Genomic databases dedicated to a single model organism arose in parallel with genome sequencing projects, for example, the Saccharomyces Genome Database (SGD)[14] and The *Arabidopsis* Information Resource (TAIR)[15]. Although initially devoted to sequence information, many of these databases eventually added many types of literature-curated information, including PPI data. In time, the publications reporting PPIs exceeded the capacity of specialized genome databases and led to

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG), Unité Mixte de Recherche 5525 Centre National de la Recherche Scientifique (CNRS), Faculté de Médecine, Université Joseph Fourier, 38706 La Tronche Cedex, France. [4]Unité de Recherche en Biologie Moléculaire, Facultés Universitaires Notre-Dame de la Paix, 61 Rue de Bruxelles, 5000 Namur, Wallonia, Belgium. [5]Genomic Analysis Laboratory and [6]Plant Biology Laboratory, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. [7]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [8]Present addresses: Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (K.V.), Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (J.-F.R.) and Department of Biological Sciences, 132 Long Hall, Clemson University, Clemson, South Carolina 29634, USA (H.B.). [9]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu) or M.E.C. (michael_cusick@dfci.harvard.edu).

PUBLISHED ONLINE 30 DECEMBER 2008; DOI:10.1038/NMETH.1284

## BOX 1 INTEROLOGS

Interologs are *in silico* predictions of protein interactions in one species between a pair of proteins whose orthologs are known to interact in another species[62–64]. The assumption that interologs are more likely true than not is widely held[65,66]. Recent evaluations have now revisited this assumption in several ways[24,67].

The most important question is where to draw the line for interspecies transfer. For instance, is mouse-human transfer close enough but more evolutionarily distant mammals not close enough? Actually, it is not the species relatedness but the sequence relatedness that really matters. Interolog transfers are only accurate for especially high sequence similarity[24,64]. Hence, interolog predictions with low sequence conservation should not be accepted, even between closely related species[24].

Investigations of intrinsic disorder in proteins have also unsettled the certainty that protein interactions are highly conserved. There are two types of interacting surfaces in proteins. Domain-domain interactions are more prevalent in stable protein complexes, whereas domain-disorder interactions are more transient[2,68,69]. Domain-disorder interactions evolve much faster than domain-domain interactions[70]. The proportion of protein interactions that are of the domain-disorder type versus the domain-domain type is not known, even approximately, for any species. Still, the likely considerable proportion of poorly conserved domain-disorder interactions means that the proportion of nonconserved interactions is substantial[24].

In the one experimental test of interologs so far, only one-third of the sample set of yeast interactions found by yeast two-hybrid were reproduced by yeast two-hybrid between the *C. elegans* orthologs[63]. Perhaps the large evolutionary distance between yeast and worm precluded a higher success rate, and mouse-human interologs might have a better success rate, but that supposition has not been experimentally tested.

In light of all these reappraisals, curation policies are changing. For instance, one interaction database has stopped transferring nonhuman interactions to human[19], a change from earlier practice[48]. Other interaction databases may follow suit. Alternatively, those interactions predicted by interolog extrapolation could be explicitly delineated in databases from those experimentally demonstrated, so the user could chose the appropriate data to examine. Either policy becomes complicated because species of the interactors are not often provided in publications[30,33]. Overall, it would seem best practice to only curate the species for which there is direct experimental evidence; in reality, doing so is difficult.



the creation of databases dedicated to PPIs, for example, the Munich Information Center for Protein Sequence (MIPS) protein interaction database[16], the Biomolecular Interaction Network Database (BIND)[17], the Database of Interacting Proteins (DIP)[18], the Molecular Interaction database (MINT)[19] and the protein Interaction database (IntAct)[20]. More recent PPI curation efforts, the Biological General Repository for Interaction Datasets (BioGRID)[21] and the Human Protein Reference Database (HPRD)[22], have attempted larger-scale curation of data from more manuscripts and more interactions.

### High-throughput efforts versus literature curation

High-throughput approaches contrast in several attributes with literature-curation strategies (**Table 1**). Literature-curated collections represent the accumulation of thousands of small-scale, 'hypothesis-driven' investigations, whereas high-throughput experiments are 'discovery-based', designed to discover new biology without a priori expectations of what could be learned. Because literature-curated datasets are hypothesis-driven, biological functions of interacting proteins often, though not always, can be inferred from the actual study design. Discovery-based high-throughput datasets do not present this advantage, though function can sometimes be inferred through additional analyses[23]. Hypothesis-driven studies set up an inevitable study bias[7], in that what has been successfully investigated before tends to be investigated again, whereas high-throughput screens avoid study bias[24]. The completeness, or the portion of the proteome that has been tested for interactions[5], can be precisely estimated in a carefully designed high-throughput study[5,7,25], but this is not so even for the largest literature-curated datasets because negative results, the pairs tested but not found to interact, are almost never reported.

Estimating reliability—the portion of reported interactions that are valid (and hence reproducible)—is daunting. For high-throughput datasets, the introduction of an empirical framework for interactome mapping now allows experimental estimation of reliability parameters[5]. Previously, reliability of high-throughput datasets was routinely estimated by measuring the overlap with a reference set of gold-standard positives (GSP). Several caveats must be considered when constructing GSPs. The assays used to generate a GSP have to match as closely as possible the assays used to generate the experimental dataset, especially indirect co-complex versus binary representation[7]. A GSP should be as unbiased as possible, sampling all, or at least most, parts and processes of the cell[26], and a GSP must be of the highest reliability and reproducibility[27].

Literature-curated datasets are used for appraisal of the reliability of experimental PPI datasets, for predicting PPIs, for predicting other features such as protein function and for benchmarking data-mining methodologies[28–30]. In these efforts, the superior reliability of literature-curated PPI datasets, versus high-throughput datasets, is generally presumed. High-quality reference datasets of PPIs are integral for empirical estimation of the reliability and size of interactome maps[5,7,25,27]. Confidence in literature curation is accordingly a prerequisite for generating useful reference datasets. Whether literature-curated PPI datasets really have exceptional reliability has not been thoroughly investigated.

**Table 1** | Comparison of strategies toward completing an interactome map

| Attribute | High-throughput | Literature-curated |
|---|---|---|
| Investigation | Discovery-based | Hypothesis-driven |
| Functional inference | Determinable from network? | Determinable from study design? |
| Study bias | Unbiased | Biased |
| Completeness | Estimable | Inestimable |
| Reliability | Determinable | Indeterminable |

### Completeness and replication of literature-curated datasets

As PPIs supported by multiple publications should be more reliable than those supported by only a single publication, we assessed the proportion of multiply supported PPIs for yeast. We ranked the 11,858 literature-curated yeast PPIs in BioGRID[21] (LC-all downloaded in mid-2007). Only 25% of LC-all PPIs have been described in multiple publications (**Fig. 1**), with just 5% and 2% of these pairs described in ≥3 or ≥5 publications, respectively. More than 75% of LC-all PPIs were thus described in a single publication. Consistent with this low portion of multiply supported PPIs, experimental retests have demonstrated a significantly lower quality for singly supported versus multiply supported literature-curated PPIs for yeast[7].

Similar investigations for human and for *Arabidopsis thaliana* showed comparably low proportions of multiply supported PPIs. In the initial search space of ~7,000 × 7,000 genes for a first-draft human interactome mapping project, there are 4,067 binary literature-curated interactions[31]. Only 15% of these PPIs have been described in multiple publications (**Fig. 1**), with just 5% and 1% described in ≥3 or ≥5 publications, respectively. More than 85% of human PPIs in the literature-curated set are supported by a single publication, greater than the 75% for yeast. The set of *Arabidopsis* PPIs was collected from the only two protein-interaction databases that curate *Arabidopsis* protein interactions, TAIR[15] and IntAct[20]. The *Arabidopsis* PPI dataset has fewer interactions supported by data in multiple manuscripts than yeast or human (**Fig. 1**), with just 1% and 0.1% described in ≥3 or ≥5 publications, respectively, with 93% of available *Arabidopsis* literature-curated PPIs supported by data in only a single publication. All told, the number of PPIs supported by data in multiple publications is small.

Literature-curated datasets are reported to be composed primarily of small-scale experiments[21,32]. To assess the presumption that PPI databases are small-scale, we measured the proportion of total PPIs identified in high-throughput experiments. For yeast, we ranked the 8,933 interactions supported by data in a single publication by the number of distinct PPIs reported in each corresponding publication (**Fig. 2a**). More than 60% of protein pairs were curated from manuscripts that described more than 10 interactions, all extracted from 6% of all the manuscripts curated. One-third of the total interactions came from less than 1% of all manuscripts that each describe 100 or more interactions (**Fig. 2a**), which would reasonably be considered high-throughput. Thus, the yeast literature-curated dataset of PPIs supported by a single publication record is not composed solely of validated interactions from small-scale studies but has a marked portion of PPIs derived from high-throughput experiments. We similarly analyzed a dataset of human curated PPIs[31] and found that this human PPI dataset is predominantly low-throughput (**Fig. 2b**), possibly because at the time these PPIs had been downloaded from the databases few medium- to high-throughput experiments had been published. For *Arabidopsis,* the proportion of the total literature-curated interactions derived from medium to high-throughput manuscripts is about the same as for yeast (**Fig. 2c**). In sum, many available literature-curated PPI datasets are populated widely by PPIs from high-throughput experiments.

As an assessment of the completeness for literature-curated datasets is not possible (**Table 1**), we evaluated database overlaps as a surrogate for completeness, on the argument that different PPI databases should curate from the same set of PubMed reports. BioGRID reports the greatest completeness for yeast but is not yet a participating member of the International Molecular Exchange (IMEx) consortium[33,34], so we could not use this database for this analysis. The three IMEx members that do substantial curation of yeast PPIs (MINT, IntAct and DIP) had surprisingly low overlap of curated PPIs (**Fig. 3a**). That the overlap is so small after years of intense curation of protein interactions is reason for concern. The small overlap is not due to differential curation of high-throughput data, as removal of the six largest PPI reports[35–40] still left small overlaps, especially of IntAct with the other two databases (**Fig. 3a**).

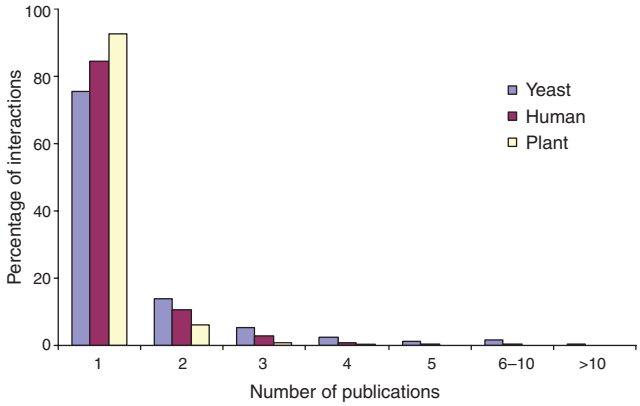


**Figure 1** | Distribution of the number of published manuscripts supporting each interaction. Data are from the dataset of yeast protein interactions downloaded from the BioGRID[21] database, the literature-curated dataset of human protein interactions, and the literature-curated dataset of *Arabidopsis* protein interactions.

Are the small overlaps due to curating different manuscripts or to differential curation of data from overlapping sets of manuscripts? The answer seems to be that vastly different sets of manuscripts are curated because the curation of PubMed reports also shows small overlap (**Fig. 3b**). For multiply supported interactions (those reported in two or more published studies), the low overlap remains (**Fig. 3c**), though the number of interactions drops greatly. Hence even the most heavily investigated interactions, those most likely to be multiply curated, do not seem to be comprehensively covered. In sum, surrogate estimates of completeness of literature-curated datasets, at least for yeast, suggest that coverage of curated literature is far from comprehensive.

These investigations suggest, but in no way demonstrate, that literature-curated PPIs may not have the high reliability often attributed to them. There has not yet been any intensive investigation of the actual reliability of literature-curated PPI datasets. To do so, we recurated representative samples of existing literature-curated PPI datasets for three model organisms—yeast (*Saccharomyces cerevisiae*), human and plant (*Arabidopsis thaliana*)—and found that the literature curation of PPI publications can be less than impeccable.

### Estimating curation reliability by recuration

For yeast, we recurated in detail 100 randomly selected pairs from the yeast dataset of singly supported interactions (**Fig. 1**). After evaluating several relevant criteria, we assigned each interaction a score of 0 (no confidence), 1 (low confidence or unsubstantiated) or 2 (substantiated or of high confidence) (see detailed protocol below).

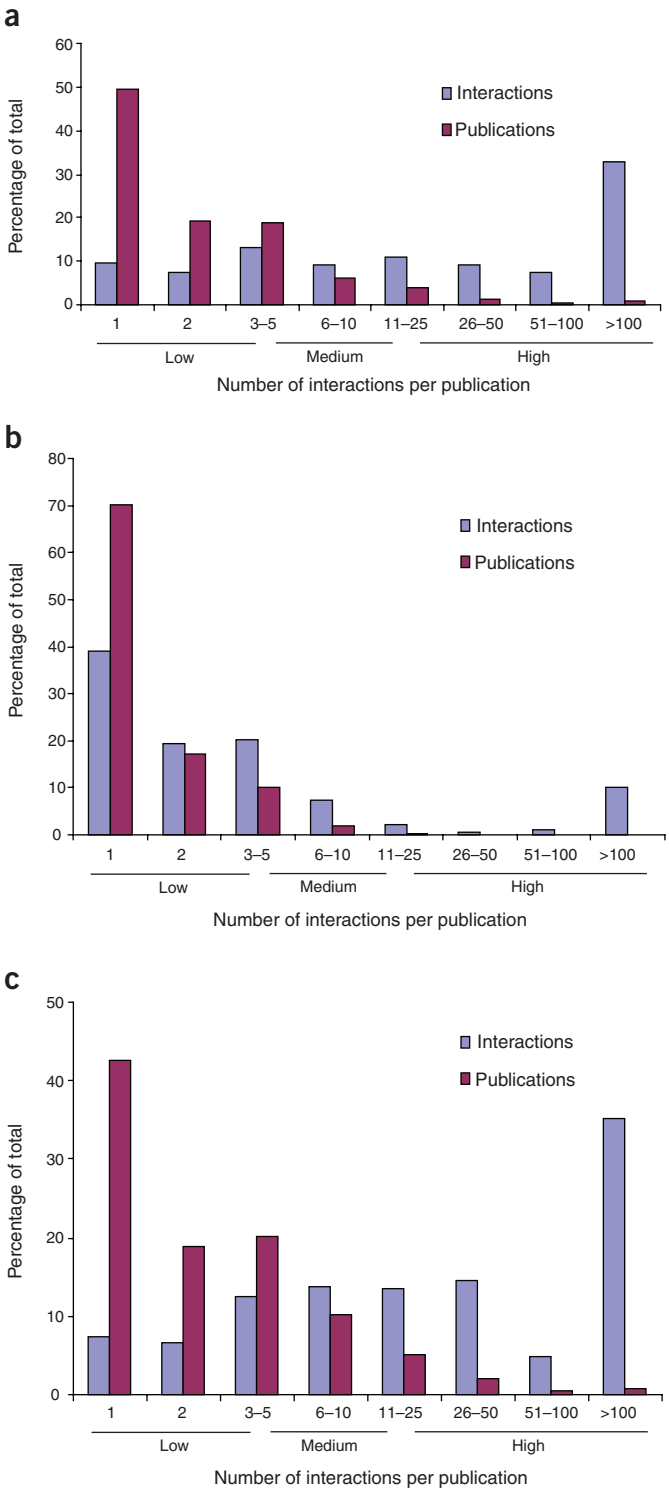


**Figure 2** | Distribution of the publications in literature-curated datasets by the number of interactions reported in the publication. (**a**–**c**) Distribution in the yeast (**a**), human (**b**) and *Arabidopsis* (**c**) literature-curated PPI datasets supported by a single publication.

The results of this recuration (**Fig. 4a** and **Supplementary Table 1** online) showed that 25% of the sampled interactions could be substantiated whereas three-quarters were not. Of the interacting pairs in the sample, 35% were incorrectly curated. These observations explain the poor reliability, relative to high-throughput datasets, of the singly supported literature-curated dataset in both computational and experimental comparative analyses[7].

For human PPI recuration, we prepared two curation datasets. One was a presumed high-confidence literature-curated dataset of interactions (LC-multiple) within the initial search space of ~7,000 × 7,000 genes for a first-draft human interactome mapping project[31] corresponding to pairs reported two or more times (two different PubMed identifiers) and curated in two or more databases (the five databases used were HPRD[22], BIND[41], MINT[19], MIPS mammalian database[16] and DIP[18]). From within this small (275 multiply supported interactions) 'hypercore' set of protein interactions[31], 188 interactions were left for recuration, after excluding homodimers.

The other dataset was a lower-confidence literature-sampled dataset of 188 interactions, generated by randomly selecting interactions from the initial search space[5]. Most of these interactions have one publication linked to the interaction, but because sampling was random, several interactions had been reported in more than one publication.

In the LC-multiple recuration set, 38% of the initial curation unit values (defined in **Table 2**) were wrong (**Fig. 4b** and **Supplementary Table 2** online). The most common errors were wrong species (assignment to a species other than human (**Box 1**)) and absence of a binding experiment supporting the interactions. Although 40% of the human LC-multiple interactions were not supported by multiple publications after recuration, most of these interactions were supported in only one manuscript instead of two or more, perhaps constituting a 'secondary' dataset of reduced confidence (**Supplementary Table 2**).

For the presumably lower confidence literature-sampled dataset of 160 interacting pairs (after removing interactions that had more than one supporting publication), 45% of interactions were not validated (**Fig. 4c** and **Supplementary Table 3** online) and 55% were validated. Almost half of the randomly sampled interactions were not supported by recuration. The most common errors here were wrong species and wrong protein name (**Fig. 4c**).

Yeast and human have the largest amount of curated literature in interaction databases[21,42]. A model organism with fewer curated interactions might yield different results. We curated 100 higher-confidence protein interactions of *Arabidopsis* from the two interaction databases that curate *Arabidopsis*, TAIR[15] and IntAct[20]. The results were improved relative to the yeast or human results, as 6 interactions and 24 curation units were scored incorrect (**Supplementary Table 4** online and **Table 2**). We scored the 24 errors as follows: 9 as 'no binding experiment'; 6 as 'no binding partner'; 6 as 'indirect'; and 3 as 'wrong protein'. The improved results for *Arabidopsis* likely reflect a smaller research community whose members can maintain uniformity in gene and protein names[15].

## Why is reliability of literature curation so low?

Our findings of large error rates in curated protein interaction databases, at least for yeast and human, are consistent with recent hints that the quality of literature-curated datasets may not be as high as widely perceived[23,29,43–45]. Perhaps occasionally curator error is responsible. However, we suggest that the errors are due not so much to curators but to the simple reality that extracting accurate information from a long free-text document can be extremely difficult. Gene name confusion is particularly thorny[30,46]. An example from our curated yeast sample illustrates the difficulties. A purification with

a tandem affinity purification tag with Vps71/Swc6 (slash separates synonymous approved names) as bait[47] pulls down a protein named Swc3, but double-checking this finds that the coresponding open reading frame is actually *SWC3* (locus name YAL011w), and not the *ALR1*/*SWC3* (locus name YOL130w) open reading frame curated in the database. A shared synonym thoroughly muddled the curation.

Common curation practice has been to score equally every interaction reported in a publication[21,48], even though common experimental practice consists of first screening for new interacting proteins, then focusing on and substantiating one or a few of the most interesting interactions while leaving the others aside. Perhaps more curator judgment is needed, applying higher ranking to verified interactions and lower ranking to unverified 'along for the ride' interactions. Users can then choose the confidence level suitable to their needs. Given the demands of systems biology, perhaps biological databases should no longer serve as mere repositories of data but should appraise data[49]. Recent small incremental steps at developing a confidence score for curated PPIs have been taken[50,51].

The difficulty of literature curation is often underappreciated[4,21,30]. The lack of formal representation of PPIs in published manuscripts makes it difficult, if not impossible, to extract the PPI data in usable form. Designation of the species of origin of the protein interactors, an absolutely critical piece of information, is often buried or lacking altogether; protein or gene name synonyms used in a particular manuscript are hard to trace back to the canonical protein or gene names, especially in older manuscripts; and standardized descriptions, sometimes all description, of the methods used are absent. Faced with these difficulties, the curator is forced either to omit the information altogether (curated false negative) or make an educated guess, even though guesses, albeit educated ones, are often erroneous (curated false positive). The small overlaps noted between curated yeast interactions in different databases (**Fig. 3**) might be due to differential treatments of potential curated false negatives.

Our observations that literature-curated datasets have inherent reliability difficulties should influence thinking about proper generation of positive reference sets[29]. Already the human positive reference sets generated in our sampled recuration efforts have proven useful in multiple investigations[5,27,52].

It is still rarely doubted that literature-curated interactions are better than datasets generated with any high-throughput technology[6,21,53,54]. Our findings lead us to argue otherwise. If rigorously carried out, high-throughput experimental PPIs can be of higher quality than literature-curated interactions[5,25,27].



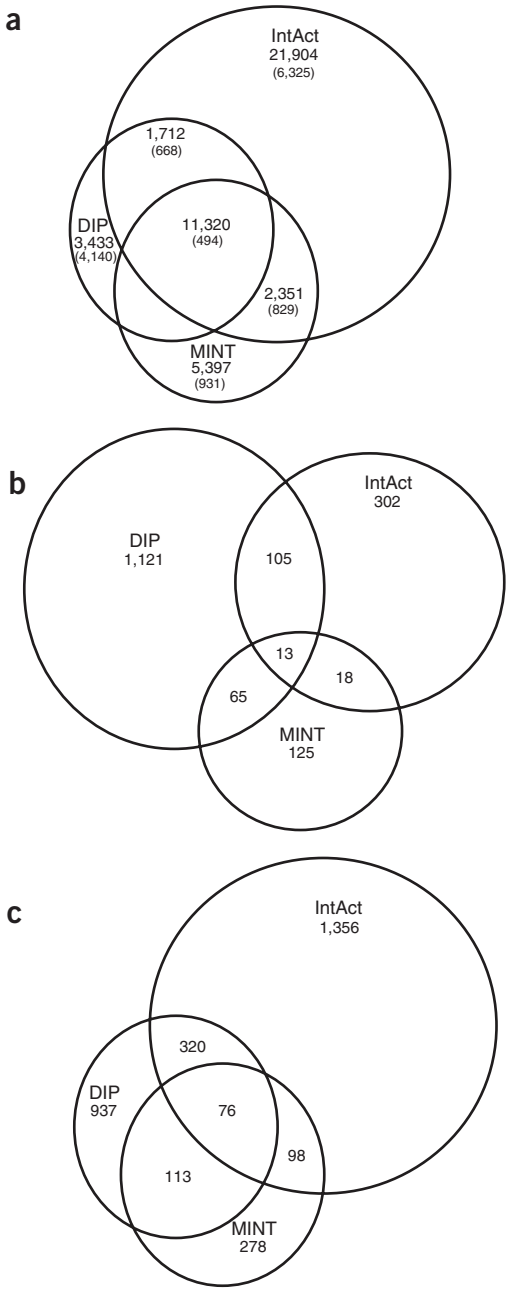


**Figure 3** | Overlaps of reported curation for yeast PPIs. (**a**) Overlaps of the total number of reported binary PPIs or after removing the largest high-throughput yeast PPI reports (numbers in parentheses). (**b**) Overlaps of the PubMed reports curated. (**c**) Overlaps after removing multiply supported interactions.

## Improving reliability of literature-curated PPI datasets

The difficulty of curation arises partly because PPI data are not submitted to databases in standardized format upon publication[55,56], unlike DNA-sequence or protein-structure data. The difficulty that curators have in extracting PPI information from manuscripts has led to the promulgation of the minimal information about a molecular interaction experiment (MIMIx) initiative[55]. MIMIx standardizes the presentation of PPI information in published manuscripts regarding species, protein names, methodological descriptions and protein identifiers, making it easier for curators to extract the pertinent information[33]. Once widely promulgated, which should come about sooner if the structured digital abstract[57,58] project gains traction, MIMIx will greatly improve curation such that the erroneous curation uncovered here will be lessened. Other minimal information initiatives for large-scale biology data are under development[59], and their development is wholeheartedly endorsed by the biocuration community so as to reduce curation error[30].

Our findings, although possibly critical of the quality of existing PPI curation, must not be used for quality evaluation of the underlying scientific literature. Actually, some PPI publications do warn of possible cross-contamination[60] or even occasionally provide heuristic confidence scores[61], warnings that should be taken into account in the curation.

**Table 2** | Summary of curation results for human and *Arabidopsis*

| Sampled dataset | Interaction units | Curation units[a] |
|---|---|---|
| Human LC-multiple | Correct: 172 (91.5%)<br>Incorrect: 16 (8.5%) | Correct: 362 (62%)<br>Incorrect: 223 (38%) |
| Human literature sampled | Correct: 88 (55%)<br>Incorrect: 72 (45%) | Correct: 88 (55%)<br>Incorrect: 72 (45%) |
| Arabidopsis | Correct: 94 (94%)<br>Incorrect: 6 (6%) | Correct: 201 (89.3%)<br>Incorrect: 24 (10.7%) |

[a]For human a curation unit is an interaction reported in one publication regardless of the number of databases curating the interaction. An interaction reported in three distinct manuscripts and curated in two databases represents three curation units. For *Arabidopsis* a curation unit is an interaction reported in one publication or one database. An interaction reported in three distinct manuscripts and with all three curated in the two *Arabidopsis* PPI databases represents six curation units.

## Curation protocols

**Yeast PPI recuration.** For each randomly selected protein pair, we read in detail the reporting manuscript (text, figures and supporting information), searching for all supporting information about the presumed interaction. We answered five questions for each protein pair. (i) Is there any information in the manuscript that supports the interaction? (ii) Has the experiment supporting the interaction been done at low throughput? As the perception persists that low-throughput experiments have greater reliability[21], knowing this is important. (iii) Are the interacting proteins mentioned together in the text? Lack of co-citation indicates that the authors did not actually focus on that particular interaction. (iv) Is the interaction supported by multiple methods? (v) Is the interaction likely direct? That is, did the method(s) used gauge binary interaction or membership in the same complex?Lastly, we assigned to each interacting pair an overall score of 0 (no confidence: no mention of the interacting pair, negative answer to the other four questions), 1 (low confidence: interacting pair is mentioned but the interaction is not substantiated by alternative methods) or 2 (high confidence: multiple validations by alternative methods).

Two different curators independently curated and scored all interactions. A third independent curator resolved the few scoring conflicts.

**Human PPI recuration.** We compiled the human PPI dataset as previously described[31]. First, we classified the method codes used by each database as binary (for example, two hybrid methods) or indirect (for example, co-affinity purification)[25]. Then, we selected only protein interactions with binary support for subsequent analysis[31]. The multiply supported literature-curated dataset comprised 585 curation units (**Table 2**) representing 188 PPIs, each reported in two or more publications and curated in two or more PPI databases. The dataset randomly selected from the full human literature-curated dataset[31] comprised 240 curation units representing 188 PPIs.

The types of information we collected during recuration were: the gene symbols and GeneID of each interactor; the associated PubMed identifier; the name and the identifier number of the interaction assay following the standard 'interaction detection method' vocabulary implemented in Proteomics Standards Initiative–Molecular Interactions[34]; the region of each protein used for the interaction assay (marked full-length if the entire protein sequence was used); the species for each interacting protein; and clarifying free-text comments used by the curator when needed. Interpretative fields included; an assessment of whether the interaction was bona fide, that is, not erroneous; an assessment of whether the interaction was indeed binary; and an error field, using a simple controlled vocabulary to classify erroneous curation units such as 'wrong protein', 'wrong species', 'no binding experiment', 'no binding partner' (interaction between the proteins is not shown), 'indirect' (no direct interaction is shown), 'redundant PubMed identifier' (some manuscripts (usually crystallographic structure determination manuscripts) have two distinct PubMed identifier

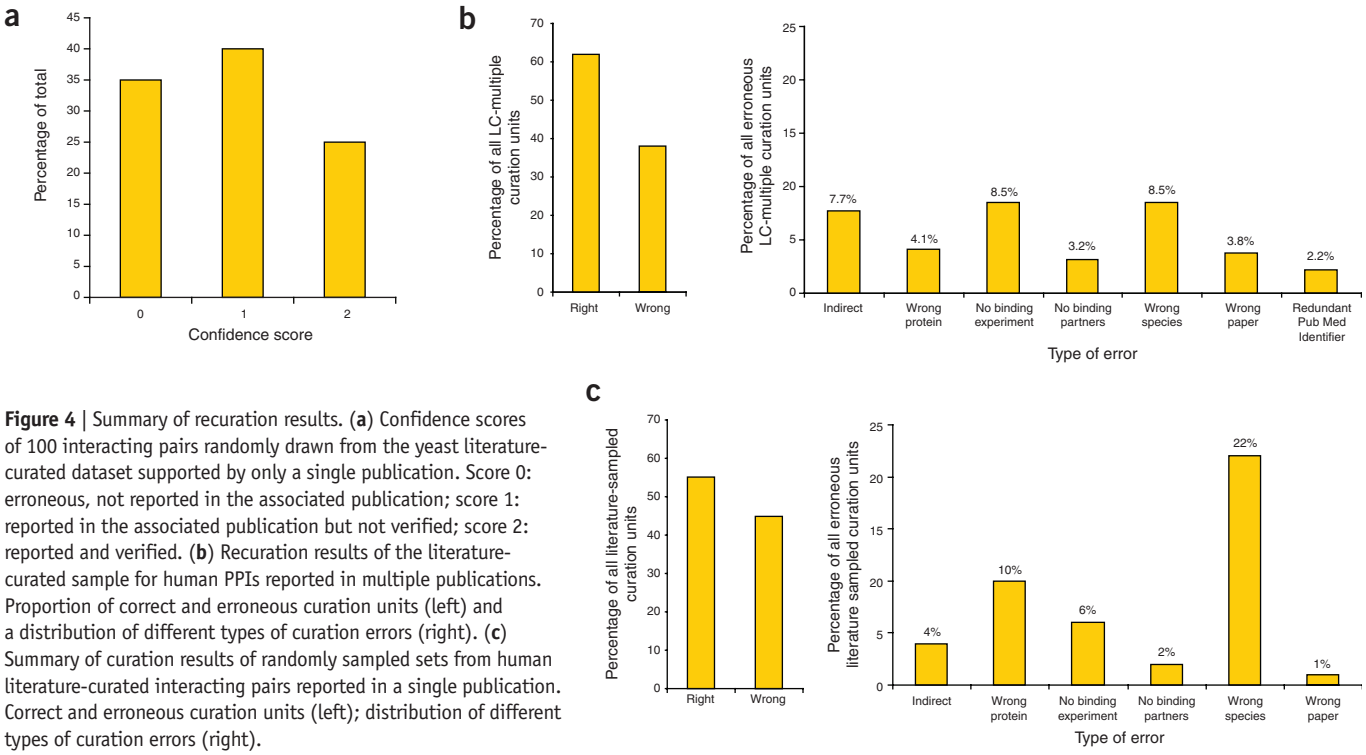


**Figure 4** | Summary of recuration results. (**a**) Confidence scores of 100 interacting pairs randomly drawn from the yeast literature-curated dataset supported by only a single publication. Score 0: erroneous, not reported in the associated publication; score 1: reported in the associated publication but not verified; score 2: reported and verified. (**b**) Recuration results of the literature-curated sample for human PPIs reported in multiple publications. Proportion of correct and erroneous curation units (left) and a distribution of different types of curation errors (right). (**c**) Summary of curation results of randomly sampled sets from human literature-curated interacting pairs reported in a single publication. Correct and erroneous curation units (left); distribution of different types of curation errors (right).

numbers in PPI databases, and thus do not constitute two distinct manuscripts supporting an interaction).

If there was no information about the region of the protein responsible for the interaction, then the default we used was the full-length protein. If the species of the interacting proteins was not stated in a manuscript, a distressingly common occurrence, our default was to record the species as human. Thus, many interactions that did not involve human proteins might have been curated as human, so we may have underestimated the actual error rate. If the interaction was legitimate but one or the other protein partner was a species other than human, then we did not call this interaction bona fide. An interaction supported by multiple methods had to have just one bona fide and binary method to be recorded as legitimate; other methods apart from this one could be not binary or erroneous and not affect the final scoring.

Generally, we labeled yeast two-hybrid and other protein complementation assays as well as structural determinations as binary. We considered immunoprecipitation and co-affinity purification methodologies done *in vivo* that assess membership in the same complex not binary, whereas we considered those done *in vitro* with, for example, recombinant proteins as binary. If a tagged protein was heterologously expressed in a cell to pull down endogenous proteins, then we called such an interaction not binary. However, if both proteins of an interacting pair were heterologously expressed in a cell and shown to interact by, for example, pull down, then we considered such an interaction binary, as it is unlikely that an endogenous host protein mediates the interaction between the two heterologous proteins. If a co-immunoprecipitation done *in vivo* occurred in both orientations (protein A immunoprecipitation pulls down protein B and protein B immunoprecipitation pulls down protein A), then we judged this interaction as binary. As experimental procedures are often not described in sufficient detail to allow judgment of binary interaction, consistent policies in this regard were difficult to achieve.

Usually, we considered structural determinations to be binary, except for protein complexes of more than two proteins in which the interacting protein pairs did not actually contact each other in the solved structure. We scored solved structures that required a small third entity for crystal formation (for example, GTP, phosphatidylinositol) as binary, even though the interaction does not occur unless the small molecule is present.

A particular curation unit could have more than one error, though we counted only the most prominent error.

***Arabidopsis* PPI recuration.** For *Arabidopsis*, we defined high-confidence interactions as those supported by two manuscripts or by two databases. In the initial search space of an ongoing *Arabidopsis* interactome mapping project, we collected 100 such interactions. We chose the union (or) instead of the intersection (and) for *Arabidopsis*, in contrast to human, so that a sufficiently sized sample of interactions was available. Otherwise, curation policies were as for human, including the error codes, but adding the name and the identifier of the 'participant identification method' vocabulary implemented in PSI-MI[34].

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

**ACKNOWLEDGMENTS**

This work was supported by US National Human Genome Research Institute grants R01 HG001715 to M.V. and F.P.R., P50 HG004233 to M.V. and R01 HG003224 to F.P.R. by funds from the W.M. Keck Foundation to M.V. by an award (DBI-0703905) from the National Science Foundation to M.V., J.R.E. and D.E.H. and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative to M.V. and CCSB. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, French Community of Belgium). We thank all members of CCSB for constructive discussions.

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

1. Cusick, M.E., Klitgord, N., Vidal, M. & Hill, D.E. Interactome: Gateway into systems biology. *Hum. Mol. Genet.* **14**, R171–R181 (2005).
2. Bader, S., Kuhner, S. & Gavin, A.C. Interaction networks for systems biology. *FEBS Lett.* **582**, 1220–1224 (2008).
3. Vidal, M. Interactome modeling. *FEBS Lett.* **579**, 1834–1838 (2005).
4. Roberts, P.M. Mining literature for systems biology. *Brief. Bioinform.* **7**, 399–406 (2006).
5. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2008).
6. Stumpf, M.P. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959–6964 (2008).
7. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
8. Parrish, J.R., Gulyas, K.D. & Finley, R.L. Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17**, 387–393 (2006).
9. Ito, T. *et al.* Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteomics* **1**, 561–566 (2002).
10. Köcher, T. & Superti-Furga, G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat. Methods* **4**, 807–815 (2007).
11. Suter, B., Kittanakom, S. & Stagljar, I. Interactive proteomics: what lies ahead? *Biotechniques* **44**, 681–691 (2008).
12. Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
13. Garrels, J.I. YPD–A database for the proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **24**, 46–49 (1996).
14. Hong, E.L. *et al.* Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* **36**, D577–D581 (2008).
15. Swarbreck, D. *et al.* The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2007).
16. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
17. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
18. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
19. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
20. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).
21. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**, 11 (2006).
22. Mishra, G.R. *et al.* Human protein reference database—2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).
23. Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C. & Troyanskaya, O.G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
24. Mika, S. & Rost, B. Protein-protein interactions more conserved within species than across species. *PLoS Comput. Biol.* **2**, e79 (2006).
25. Simonis, N. *et al.* Empirically-controlled mapping of the *Caenorhabditis elegans* protein-protein interaction network. *Nat. Methods* **6**, 47–54 (2008).
26. Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7**, 535–545 (2004).
27. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97 (2008).
28. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
29. Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T. & Albrecht, M. Computational analysis of human protein interaction networks. *Proteomics* **7**, 2541–2552 (2007).
30. Howe, D. *et al.* The future of biocuration. *Nature* **455**, 47–50 (2008).
31. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

32. Peri, S. *et al.* Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
33. Orchard, S. *et al.* Submit your interaction data the IMEx way. A step by step guide to trouble-free deposition. *Proteomics* **7**, 28–34 (2007).
34. Kerrien, S. *et al.* Broadening the horizon - Level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
35. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
36. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
37. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
38. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae. Nature* **440**, 637–643 (2006).
39. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
40. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**, 623–627 (2000).
41. Alfarano, C. *et al.* The Biomolecular Interaction Network Database (BIND) and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424 (2005).
42. Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7**, S19 (2006).
43. Gentleman, R. & Huber, W. Making the most of high-throughput protein-interaction data. *Genome Biol.* **8**, 112 (2007).
44. Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M. & Matthews, J.M. Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007).
45. Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M. & Matthews, J.M. Response to Chatr-aryamontri *et al.*: Protein interactions: to believe or not to believe? *Trends Biochem. Sci.* **33**, 242–243 (2008).
46. Nelson, D.R. Gene nomenclature by default, or BLASTing to Babel. *Hum. Genomics* **2**, 196–201 (2005).
47. Krogan, N.J. *et al.* A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol. Cell* **12**, 1565–1576 (2003).
48. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
49. Philippi, S. & Kohler, J. Addressing the problems with life-science databases for traditional uses and systems biology. *Nat. Rev. Genet.* **7**, 482–488 (2006).
50. Kiemer, L., Costa, S., Ueffing, M. & Cesareni, G. WI.-PHI a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7**, 932–943 (2007).
51. Chatr-Aryamontri, A., Ceol, A., Licata, L. & Cesareni, G. Protein interactions: integration leads to belief. *Trends Biochem. Sci.* **33**, 241–242 (2008).
52. Boxem, M. *et al.* A protein domain-based interactome network for *C. elegans* early embryogenesis. *Cell* **134**, 534–545 (2008).
53. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
54. Batada, N.N., Hurst, L.D. & Tyers, M. Evolutionary and physiological importance of hub proteins. *PLoS Comput. Biol.* **2**, e88 (2006).
55. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).
56. Hermjakob, H. *et al.* The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
57. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett.* **582**, 1171–1177 (2008).
58. Gerstein, M., Seringhaus, M. & Fields, S. Structured digital abstract makes text mining easy. *Nature* **447**, 142 (2007).
59. Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
60. Stevens, S.W. *et al.* Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol. Cell* **9**, 31–44 (2002).
61. Fromont-Racine, M., Rain, J.C. & Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* **16**, 277–282 (1997).
62. Walhout, A.J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
63. 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).
64. Yu, H. *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
65. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
66. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
67. Levy, E.D. & Pereira-Leal, J.B. Evolution and dynamics of protein interactions and networks. *Curr. Opin. Struct. Biol.* **18**, 349–357 (2008).
68. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008).
69. Fuxreiter, M., Tompa, P. & Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23**, 950–956 (2007).
70. Beltrao, P. & Serrano, L. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.* **3**, e25 (2007).

# Addendum: Literature-curated protein interaction datasets

Michael E Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, Jean-François Rual, Heather Borick, Pascal Braun, Matija Dreze, Jean Vandenhaute, Mary Galli, Junshi Yazaki, David E Hill, Joseph R Ecker, Frederick P Roth & Marc Vidal

*Nat. Methods* **6**, 39–46 (2009); published online 30 December 2008; addendum published after print 25 November 2009.



We assessed literature-curated protein-protein interaction (PPI) datasets for the parameters of completeness, coverage and quality by several means, concluding that such datasets might be "possibly of lower quality than commonly assumed." A Correspondence[71] by members of the International Molecular Exchange Consortium (IMEx), while accepting many of our points, objected to our recuration exercise to assess quality, finding our criteria "subjective." We argue that the criteria were commonsensical and essentially capture how these databases are often described.

A wide swath of the scientific community, from computer scientists and engineers to physicists, systems biologists and molecular biologists, use literature-curated datasets as 'gold-standard' positive controls with the tacit understanding that this information is nearly perfect. Whether user impressions were formed from statements made by database authors[18–21] or not, belief that database entries accurately correspond to high-quality, direct physical interactions is widespread[6,72]. The standards we used to assess quality are generally accepted by the IMEx members, but one that remains problematic is the definition of binary interactions. A meaningful fraction of database users is under the impression that 'binary interaction' means direct pairwise PPIs, and that is the definition we tried to apply. The definition that the IMEx databases apply is that of 'binary representation', meaning any pairwise association between two entities, direct or indirect. Although technically correct from an informatics viewpoint, binary representation likely does not accurately reflect biophysical reality. To better match user expectations, one IMEx database has adjusted their website presentation to allow users to filter 'spoke expanded co-complexes' from binary interactions, although all reported interactions are initially classified as 'binary'.

Another widespread perception is that curated databases contain predominantly low-throughput interactions, whereas the reality is that curated databases have a substantial portion of interactions derived from high-throughput experiments (Fig. 2 in our Perspective). The point is not whether high-throughput interaction experiments are of worse or better quality than low-throughput experiments, but that greater transparency should be provided so that users can filter the data according to their needs.

As a result of applying the criteria that we did, based on the observations above, the error rates we reported reflected not only errors in curation but also how well the underlying data meet the standards set forth. The details for the yeast, human and plant recurations are available in the **Supplementary Note**.

Our efforts are aimed at alerting the scientific community that literature-curated interactions may need further scrutiny or classification to qualify as a 'gold standard' for users who are specifically interested in direct pairwise PPIs. Closer inspection will allow the community to be the ultimate judge of how useful these curation units turn out to be.

We updated our original Supplementary Table 2 on LC-multiple human recurated dataset to show the databases from which each interaction came (**Supplementary Table 1**). Almost 90% of interactions, and 95% of the problematic curation units, came from non-IMEX databases (HPRD[22] and BIND[17]). We had been requested to omit this information originally, but for IMEX databases there is minimal difference in error rates between our recuration and that of Salwinski *et al.*[71]. A download discrepancy, which IntAct has now mended so that it cannot recur, necessitated the recuration of the errors for the *Arabidopsis* curation (Supplementary Table 4 in our original Perspective). We now score the 24 curation errors as: 3 'no binding experiment' (formerly 9); 6 'no binding partner' (formerly 6); 11 'indirect' (formerly 6); 3 'wrong protein' (formerly 3); and 1 'wrong species' (formerly 0).

Unfortunately the download dates for the interaction data in our original Perspective were unclear or missing. The download date for the yeast interaction data was originally reported as mid-2007 but is actually early 2006. Human interaction data were downloaded from HPRD, BIND, MINT, MIPS and DIP in mid-2005, as described in ref. 31. *Arabidopsis* interaction data from IntAct and TAIR were first downloaded in February 2008. The second download, which we used in the analysis above, occurred in March 2009 when the download inconsistencies were pointed out to us.

Our contentions that literature-curated datasets are imperfect were corroborated by a paper published concurrently[73]. Especially telling was the observation in that paper that many "databases lack a substantial portion of PPIs, emphasizing the need to integrate multiple PPI databases"[73], a concern fully echoed by our original finding of low overlaps between curated PPI databases (Fig. 3 in our original Perspective). The problem of low overlaps should be mitigated once the IMEx exchange of curation between databases becomes implemented[33].

Other investigators have reported that literature-curated interaction datasets are less perfect than is widely presumed. In papers in *Trends in Biochemical Sciences*[44,45,51] the authors argued over a distressing lack of reproducibility of curated interactions and contended that "protein interactions reported in the literature and curated in interaction databases might not occur as presented." Other reports have questioned the presumed perfection of curated PPIs[23,29,43,74], even one report by several authors of Salwinski *et al.*[71]: "a comparison of publications curated by both MINT and IntAct between 2003 and 2005 revealed that the two databases annotated exactly the same interaction pairs in only 6 out of 52 publications"[75]. BioGRID now grants that provisions are not made for quality assessment in curation: "We make no judgement calls on the methods or even, within reason, the quality of the data themselves"[76]. Perhaps quality of the underlying data should in some way begin to be assessed, to match community expectations of curated data.

Curation to extract protein-protein interactions from the literature is absolutely critical to the advancement of systems biology and proteomics. Increased transparency and appropriate communication of what is currently available in curated datasets will ultimately help these efforts. Preliminary steps toward generating confidence scores have been reported for curated[50], predicted[77] and experimental[27] PPI datasets. These measures go in the right direction and their further development should be encouraged and appropriately funded.

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


71. Salwinski, L. *et al.* Recurated protein interaction datasets. *Nat. Methods* **6**, 860–861 (2009).
72. Lee, I., Li, Z. & Marcotte, E.M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS ONE* **2**, e988 (2007).
73. Wu, J. *et al.* Integrated network analysis platform for protein-protein interactions. *Nat. Methods* **6**, 75–77 (2009).
74. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006).
75. Chatr-aryamontri, A. *et al.* MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biol.* **9**, S5 (2008).
76. Blow, N. Systems biology: untangling the protein web. *Nature* **460**, 415–418 (2009).
77. Geisler-Lee, J. *et al.* A predicted interactome for *Arabidopsis*. *Plant Physiol.* **145**, 317–329 (2007).