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).

Corrigendum: Nanoscale live-cell imaging using hopping probe ion conductance microscopy

Pavel Novak, Chao Li, Andrew I Shevchuk, Ruben Stepanyan, Matthew Caldwell, Simon Hughes, Trevor G Smart, Julia Gorelik, Victor P Ostanin, Max J Lab, Guy W J Moss, Gregory I Frolenkov, David Klenerman & Yuri E Korchev
***Nat. Methods* 6, 279–281 (2009); published online 1 March, 2009; corrected after print 3 September 2009.**

In the version of this paper originally published, references to previous work on pulse mode SICM should have been included (Mann, S.A. *et al. J. Neurosci. Methods* **116**, 113–117, (2002) and Happel, P. *et al. J. Microsc.* **212**, 144–151 (2003)). These references were removed during shortening of the paper for publication and have been added back to the PDF and HTML versions of this article. The pulse mode technique reported in these previous papers has conceptual similarity to our hopping mode SICM, in that distance feedback control is not continuous; thus, it also solves the problem of probe-sample collision for large cellular structures. However, the pulse mode technique is considerably slower owing to a different feedback mechanism and does not perform at nanoscale resolution.

Erratum: 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog

Matija Dreze, Benoit Charloteaux, Stuart Milstein, Pierre-Olivier Vidalain, Muhammed A Yildirim, Quan Zhong, Nenad Svrzikapa, Viviana Romero, Géraldine Laloux, Robert Brasseur, Jean Vandenhaute, Mike Boxem, Michael E Cusick, David E Hill & Marc Vidal
***Nat. Methods* 6, 843–849 (2009); published online 25 October, 2009; corrected after print 16 November 2009.**

In the version of this article initially published, the schematic in Figure 5a was misaligned. The error has been corrected in the HTML and PDF versions of the article.

Erratum: What's in a test?

Anonymous
***Nat. Methods* 6, 783 (2009); published online 29 October 2009; corrected after print 16 November 2009**

In the version of this article initially published, the name of Robert Cook-Deegan was misspelled. The error has been corrected in the HTML and PDF versions of the article.