Protein interaction data curation: the International Molecular Exchange (IMEx) consortium

Sandra Orchard[1], Samuel Kerrien[1], Sara Abbani[2], Bruno Aranda[1], Jignesh Bhate[3], Shelby Bidwell[4], Alan Bridge[5], Leonardo Briganti[6], Fiona Brinkman[7], Gianni Cesareni[6,8], Andrew Chatr-aryamontri[6,9], Emilie Chautard[10,11], Carol Chen[12], Marine Dumousseau[1], David Eisenberg[2], Johannes Goll[4], Robert Hancock[12], Linda I Hannick[4], Igor Jurisica[13], Jyoti Khadake[1], David J Lynn[14], Usha Mahadevan[3], Livia Perfetto[6], Arathi Raghunath[3], Sylvie Ricard-Blum[10], Bernd Roechert[5], Lukasz Salwinski[2], Volker Stümpflen[15], Mike Tyers[9,16],
Q1 Peter Uetz[17,18], Ioannis Xenarios[5,19,20] & Henning Hermjakob[1]

The International Molecular Exchange (IMEx) consortium is an international collaboration between major public interaction data providers to share literature-curation effort and make a nonredundant set of protein interactions available in a single search interface on a common website (http://www.imexconsortium.org/). Common curation rules have been developed, and a central registry is used to manage the selection of articles to enter into the dataset. We discuss the advantages of such a service to the user, our quality-control measures and our data-distribution practices.

Protein-protein interactions are a key element in our understanding of molecular biology. However, in contrast to areas of activity such as DNA sequencing or protein structural analysis, the systematic capture of published molecular interaction data into public domain repositories is still in its infancy. This is not due to lack of resources in this domain. As of December 2011, the PathGuide resource[1] listed more than 100 protein-protein interaction–related databases. Although many of these databases focus on predictions of potential interactions or on mapping interologs, rather than experimentally determined interactions, the extent of activity suggests ample resources. However, most of these resources are independently funded and pursue their goals in isolation. As a result, accessing all publicly available molecular interaction data, even on a specific biological or biomedical topic, is a challenging, time-consuming task that requires the user to query multiple resources, each with a different interface; additionally, many resources use different identifiers and often contain redundant data from overlapping sets of publications.

Efforts to address this problem began ten years ago with the development of a common file format for representing protein-interaction data. The 'minimum information about a molecular interaction experiment' (MIMIX) guidelines had been published then[2], defining a list of the information to be supplied when

[1]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [2]University of California, Los Angeles, Department of Energy Institute for Genomics and Proteomics, Los Angeles, California, USA. [3]Molecular Connections Pvt. Ltd., Bangalore, India. [4]The J. Craig Venter Institute, Rockville, Maryland, USA. [5]Swiss-Prot group, Swiss Institute of Bioinformatics, Geneva, Switzerland. [6]Department of Biology, University of Rome Tor Vergata, Rome, Italy. [7]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada. [8]Istituto di Ricovero e Cura a Carattere Scientifico–Fondazione S. Lucia, Rome, Italy. [9]Wellcome Trust Centre for Cell Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, Scotland. [10]Institut de Biologie et Chimie des Protéines, Unité Mixte de Recherche 5086 Centre National de la Recherche Scientifique. Université Lyon 1, Lyon, France. [11]Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada. [12]Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, British Columbia, Canada. [13]Ontario Cancer Institute, the Campbell Family Institute for Cancer Research, and Techna Institute, University Health Network, Toronto, Ontario, Canada. [14]Animal and Bioscience Research Department, Animal and Grassland Research and Innovation Center, Teagasc, Grange, Dunsany, County Meath, Ireland. [15]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany. [16]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [17]Proteros Biostructures, Martinsried, Germany. [18]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, USA. [19]Vital-IT group, Swiss Institute of Bioinfomatics, Lausanne, Switzerland. [20]University of Lausanne, Lausanne, Switzerland. Correspondence should be addressed to S.O. (orchard@ebi.ac.uk).

PUBLISHED ONLINE XX XXXX 2012; DOI:10.1038/NMETH.XXXX

describing experimental molecular-interaction data in a journal publication. In parallel to this, the curation strategies of a select group of molecular-interaction databases, the IMEx consortium, were coordinated to create a single non-redundant set of homogeneously curated protein-interaction data, as we discuss here.

A common data format and the IMEx consortium

The issue of the individual data resource formats maintained by the separate resources has largely been addressed by the efforts of the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI)[3]. In 2002, several providers of protein-interaction data, among them Biomolecular Interaction Network database (BIND)[4], Database of Interacting Proteins (DIP)[5], Hybrigenics[6], IntAct[7], Molecular Interaction database (MINT)[8] and Munich Information Center for Protein Sequences (MIPS)[9], set out to develop a common file format for the representation of protein-interaction data. This resulted in the creation of the HUPO-PSI-MI XML format[10], which is now widely implemented, and has since been expanded to enable the interchange of all forms of molecular-interaction data[11]. This enables the user to download, combine, visualize and analyze data in a single format from multiple resources. This format has since been supplemented by a simplified tabular format, MITAB[11].

Although a common data format is a key step in providing consistent, user-friendly access to publicly available molecular interaction data, it is only a first stage. Until recently, all interaction databases independently curated interaction-data publications, on occasion resulting in several different datasets derived from a single publication, owing to the implementation of different curation strategies. In addition to the use of scarce public funding for the duplication of expensive manual database curation, the differences in the datasets can leave the user bewildered about which to regard as the correct interpretation of data in a paper. To address this issue, five molecular interaction databases agreed in September 2005 on a long-term coordination of their curation strategies. The framework for this collaboration was the IMEx consortium, which currently comprises DIP[5], IntAct[7], MatrixDB[12], MINT[8], Microbial Protein Interaction database (MPIDB)[13], I2D[14], InnateDB[15] and Molecular Connections (http://www.molecularconnections.com/home/en/home/products/netPro/) as full members, with Biological General Repository for Interaction Datasets (BioGRID)[16] as an observer member. A full IMEx consortium member commits to producing a relevant number of records curated to a common IMEx consortium standard, whereas an observer member is a prospective IMEx consortium member, working with the full members to produce the curation rules and improve curation quality. The aims of the IMEx consortium are to coordinate curation to avoid redundant work on the same data, increase curation coverage and synchronize curation strategies to ensure consistency of data across all IMEx consortium member databases (IMEx databases). Since 2005, an increasing number of these databases have been working together to generate a single set of curation rules to ensure both the quality and consistency of annotation across the IMEx databases. As a result of many detailed IMEx consortium discussions, a single joint IMEx consortium curation manual (http://www.imexconsortium.org/curation/) has been agreed on and made publicly available. This forms the basis for the curation by all IMEx databases and at all levels uses the controlled vocabularies developed by the HUPO-PSI[10,11].

Table 1 | Current journal coverage by IMEx consortium members

Journal	Period of coverage	Database
Cancer Cell	January 2006–present	IntAct
Cell	January 2006–present	IntAct
FEBS Letters	January 2005–present	MINT
EMBO Journal	January 2006–present	MINT
EMBO Reports	January 2006–present	MINT
Journal of Bacteriology	August 2007–present	MPIDB
Journal of Molecular Signaling	November 2006–present	Molecular Connections
Matrix Biology	January 2009–present	MatrixDB
Molecular Cancer	September 2010–present	Molecular Connections
Molecular Microbiology	August 2007–August 2009	MPIDB
Nature Immunology	October 2010–present	InnateDB
Nature Structural and Molecular Biology	January 2006–present	DIP
Oncogene	September 2010	I2D
PLoS Biology	January 2003–present	DIP
Proteomics	January 2005–present	IntAct
Structure	January 2006–present	DIP

Curation strategy and coverage

Protein-interaction databases currently contain a considerable amount of redundant data, that is, the same paper curated by multiple resources, often to differing depths of curation or following different annotation strategies. As stated above, one of the major aims of the IMEx consortium is to present the user with a nonredundant dataset to search: namely, each paper should be present only once in the IMEx dataset, with the protein-protein interaction information it contains having been fully captured following consistent rules.

Initially, the IMEx consortium members agreed to share the curation workload based on journal selection. Each member selected one or more journals to curate, with the aim of representing in the database all relevant protein-interaction data published in that journal within a reasonably short time of publication, normally less than three months. The IMEx consortium members selected journal(s), which largely reflect their particular areas of interest or editorial connections (**Table 1**). There is no preselection of data from particular organisms, although, in practice, the well-studied model organisms such as *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana, Saccharomyces cerevisiae* and *Escherichia coli* also tend to be the best represented in the scientific literature available for curation.

Whereas articles from targeted journals form the baseline of IMEx consortium curation, most databases curate additional publications; this choice is usually based on scientific collaborations, curator expertise or reflects the specialization of thematic databases such as MatrixDB and MPIDB. As an example, IntAct recently curated a targeted dataset on interactions of proteins that have a role in Alzheimer's disease[17]. Until recently, these targeted curation efforts were not coordinated between the IMEx consortium members. However, in 2010 we released IMExCentral, a web service that enables IMEx consortium partners to reserve any publication for curation, either manually through a web interface or through a web service directly from our curation tools. Based on this tool, we are now also coordinating curation of all individual publications outside of the journal curation commitment.

IMEx consortium members are now working on releasing a non-redundant set of all papers curated to IMEx consortium

Table 2 | Publications curated per calendar year and the number of those released to date to the IMEx dataset

Database:	MINT		IntAct		DIP		MPIDB		MatrixDB		Molecular Connections		I2D		InnateDB	
Year	Curated	Exported	Curated	Exported	Curated	Exported	Curated	Exported	Curated	Exported	Curated	Exported	Curated	Exported	Curated	Exported
2001	278	6	0	0	0[a]	0										
2002	185	0	0	0	0[a]	0										
2003	131	1	110	0	0	0										
2004	538	29	348	0	3,005[a]	0										
2005	439	120	519	1	0[a]	0										
2006	557	401	1,294	236	0[a]	0										
2007	268	259	715	87	899	0	813	0								
2008	466	251	756	138	771	771	0	0							1,038[b]	
2009	574	211	478	123	621	3	183	183							808[b]	
2010	957	152	348	130	542	542	0	0	36	36	18	6	42	27	2,596[b]	
2011	614	284	447	160	615	615	5	5	41	25	17	17	58	58	676[b]	27

[a]Records were curated before the formation of IMEx consortium; exact release date cannot be tracked. [b]Data from publications not exported to IMEx are curated to MIMIx standards.

standards by the participating databases since 2006. Key large-scale papers, such as the protein-interaction map of *Drosophila melanogaster*[18], and the human protein-protein interaction networks[19,20] have been recurated to the existing IMEx consortium standard and released to the dataset. More recent large-scale papers are routinely added to the dataset, and users are encouraged to propose additional publications for curation.

Several of the participating databases contain data curated to different depths (see below) or which were curated while the IMEx consortium rules were under development. Most participating databases have a wealth of data curated from published papers that have yet to be released to the IMEx consortium dataset (**Table 2**). A major aim in 2012 will be to identify archival data appropriate for release through the IMEx consortium website and, if necessary, recurate these to current IMEx consortium standards. Data curated by MINT and IntAct as training and test datasets for the BioCreative competitions[21,22] have already been released as part of this process. Where data from a paper have previously been redundantly curated, that is, annotated by more than one IMEx database, IMExCentral will only allow one set of data for the paper to acquire an IMEx accession number and will alert the databases if a second resource attempts to register the same publication.

IMExCentral already allows participating databases to encourage and manage the annotation of directly submitted data as an integral part of the publication process. Authors may submit data to any IMEx database. A common identifier allocated by IMExCentral (IM-xxxx), will allow data users to access the dataset, after publication, both in the original resource and via the IMEx website. Should identical data be offered to more than one member database, this will immediately be highlighted by the IMExCentral service when a database attempts to register a second copy of the same dataset.

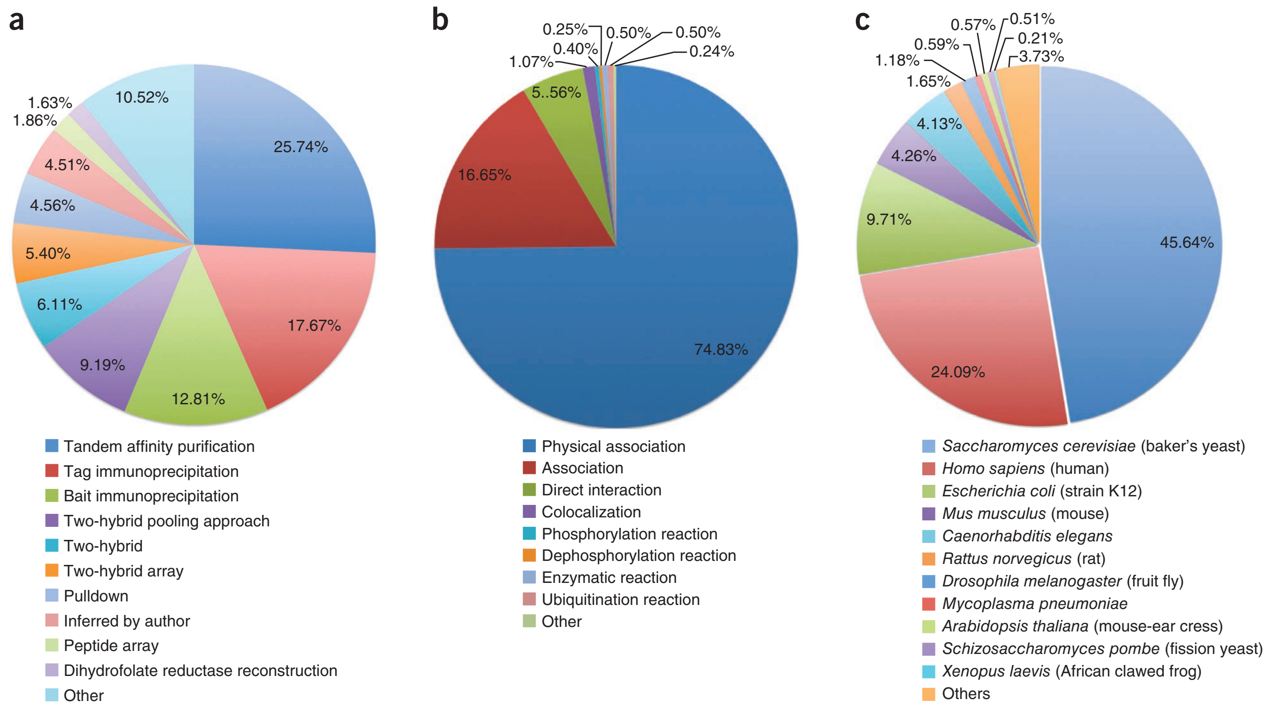
In addition to deposition of new experimental data, IMEx database users can also request curation of specific publications via the IMEx website (http://www.imexconsortium.org/), for example, if they notice a well-known interaction missing from the IMEx databases or to establish the currently known interactions for a particular research target.

Curation depth

The IMEx consortium partners have committed to a 'deep' curation model, which aims to capture the full experimental detail provided in the interaction report, as this is often essential to assess interaction context and confidence. In fact, it has become increasingly clear that minor changes in experimental detail may have dramatic effects on the outcome of an interaction experiment[23]. In **Figure 1a**, we illustrate the major interaction detection methods used to identify protein interactions represented in the IMEx dataset, as defined in the PSI-MI controlled vocabulary (http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI). IMEx consortium members refer to all interactions between two molecules as binary interactions, and these are classified by the type of binding described (**Fig. 1b**). 'Association' indicates that the interaction is from an experimental method that identifies a loose 'co-complex', in which all the members may not have been identified, typically by co-immunoprecipitation or pulldown from an *in vivo* sample. 'Physical association' indicates the interaction has been identified by a method indicating a tighter complex, but again in which all the members may not have been identified—for example, protein-complementation assays such as yeast two-hybrid. 'Direct interaction' indicates that the two molecules are known to be in actual physical contact with each other. Evidence of direct interaction is only taken from *in vitro* methodologies and does not include yeast two-hybrid assays, but we acknowledge that when performed properly yeast two-hybrid assays are strong evidence of a direct interaction. Experimental molecular features such as affinity tags, labels and functional protein modifications, including post-processing of the transcript or phosphorylation sites, are mapped to the given sequence, as are binding domains and interacting residues. Author-provided confidence scores are also documented, where these data are available. Where essential data as required by the MIMIx guidelines[2] are not available, we either obtain the data from the corresponding author or mark the manuscript as containing data that cannot be annotated. The IMEx consortium members collate experimental evidence from any species for which interaction data are available (**Fig. 1c**).

Quality control

Curation rules are only useful if they are consistently applied, and the IMEx consortium is gradually implementing measures for mutual quality control. The PSI validator[24], a tool that executes rules based on the PSI-MI ontology to check XML files, provides not only syntactic checking of released files but also semantic checking, validating the use of the correct controlled vocabularies



Q8 Q9 **Figure 1** | Overview of the IMEx dataset. (**a**) Interaction detection methods currently represented in the IMEx dataset. (**b**) Types of interaction data represented in the IMEx dataset. (**c**) Species for which data are available in the IMEx dataset. All data were collected in December 2011.

Q4 as well as more complex, context-dependent rules. Validation rules that ensure compliance with the IMEx consortium curation manual have been developed and are now publicly available for use by consortium members, data submitters and, indeed, any user of the PSI-MI XML format.

Cross-curation exercises have been undertaken, and these will remain an ongoing regular exercise, with several 'challenging' papers being selected for annotation by all participating databases. The resulting download files are compared, and discrepancies in data capture are discussed to ensure curation rules and controlled vocabularies are used consistently across databases. Alternatively, rules or vocabularies may be modified to address challenges.

In addition, each month, members of one database select a paper for discussion by the collaborators, initially via a Wiki page, but if problems cannot be resolved, then they can be discussed in a phone conference or face-to-face meetings. In this way, rules can be generated to address new technologies or variations on accepted methodologies. Finally, ~20 papers highlighted by the iRefIndex database[25] as being curated by more than one IMEx database have been compared. The redundant curation predated the formation of the IMEx consortium, and the exercise confirmed that the current IMEx consortium curation rules and internal quality control measures would have addressed the vast majority of problems identified.

Data dissemination

Many collaborative curation projects—for example, UniProt, Gene Ontology annotation or wwPDB—exchange data on a regular basis, with the data from each partner being copied to all other partners. However, the regular full copying of complex records from multiple partners, in particular the management of the updates and deletions of both interaction records and the underlying sequences to which they are mapped, is highly resource-consuming in terms of both computational load and staff. Although IMEx consortium partners have been increasingly collaborating since 2005, we only recently entered IMEx 'production mode' with the regular release of IMEx records and required sharing of curated interaction data between partners.

Recently, a standard interface for direct computational access to standards-compliant molecular interaction data resources, the PSI Common Query Interface (PSICQUIC) was developed[26]. PSICQUIC supports simultaneous querying of multiple participating molecular-interaction databases.

IMEx consortium partners decided to use the distributed PSICQUIC system as the basis for IMEx data dissemination to minimize the data-exchange overhead[26]. IMEx interaction records are delivered to the IMEx consortium partners and individual member database websites through a tagging process. Only IMEx consortium partners may use the IMEx tag, and only records presented in a registered PSICQUIC service tagged as an IMEx record and with an IMEx accession number will be part of this IMEx dataset. Each IMEx consortium partner operates a PSICQUIC server, and a PSICQUIC client can query all partners for IMEx data matching a given query, providing an up-to-date view of all relevant data from all IMEx consortium partners.

When using the full PSICQUIC service, users can access all available interaction data including the tagged data subset provided by IMEx consortium members. However, when searching all data available through PSICQUIC, it is currently difficult to separate experimentally proven binary pairs from predicted interactions, functional associations or the results from text-mining. This data are also highly redundant, in that the manually curated data in primary databases are re-exported by several integrative databases such as iRefIndex[25], Agile Protein Interaction

DataAnalizer (APID)[27] and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)[28]. Unfortunately, much of the experimental detail may be lost during the integration process, although a link back to the primary database record is usually provided. For example, as of 27 February 2012, the data associated with one publication (PubMed identifier (PMID):17923092) appeared in six resources when searched for in PSICQUIC, and in many of these resources it is not clear that the majority of data in this paper derive from genetic interference assays (PSI-MI controlled vocabulary identifier MI:0254) as the data in integrative databases can lack the detailed information required to make this clear. In the IMEx dataset, each interaction publication appears only once, with experimental detail and the protein constructs clearly defined. Users are encouraged to access and search the IMEx dataset via PSICQUIC, either directly from http://www.imexconsortium.org/ or via member database websites.

In addition to the interactive PSICQUIC access, all IMEx data are also available for full download in PSI-MI XML or MITAB tabular formats. All IMEx data from all partners are freely available without any restrictions.

In the future, we expect a substantial increase in the coverage of IMEx records, in particular through ongoing curation and the acquisition of new IMEx consortium partners but also through an 'upgrade' of existing archival records to IMEx records as discussed above. In particular, we will validate and where necessary recurate widely used large-scale interaction datasets as IMEx records. Most importantly, however, we aim to shift our focus from curation after publication to curation before publication, in collaboration with all relevant stakeholders. Curation of data before publication, in direct dialog with the authors, ensures data representation that is both factually correct and optimally aligned with the authors' view of the data. Through inclusion of IMEx accession numbers in the publication and data release synchronized with the publication of the paper, both data producers and databases benefit from increased visibility, and users benefit from timely access to this comprehensive, annotated and accurate protein-interaction data.

Why is the IMEx consortium necessary?

As previously stated, many interaction databases exist, which attempt to capture protein-interaction data from the literature using different curation strategies. In addition to this, there are now several 'composite' databases, which contribute no new manual curation but instead merge the work of other resources. Other databases take a median strategy, importing selected data from curated resources and adding to this their own annotation. There are also datasets of predicted protein interactions, using a variety of information sources. Attempts to merge data are often hampered by the differing strategies adopted by the databases, in particular when mapping ambiguous protein descriptions given in the text to identifiers in sequence databases. Even when both gene name and species are stated, which is often not the case[2], authors rarely clearly define which isoform of the protein they are dealing with, even when this information is known. Databases deal with this ambiguity in several ways, either by mapping the data to a gene identifier and sacrificing all ability to map to a specific isoform (BioGRID) or by selecting one transcript, usually the longest (BIND), which makes it impossible to indicate when this is an ambiguous or a specific mapping, or by using the canonical sequence displayed by UniProtKB (IntAct, MINT, DIP, MatrixDB, I2D and MPIDB).

Another cause of apparent differences between databases is their varying policies to describe interactions demonstrated between protein constructs from different species, for example, human and mouse. Most databases report the data to the exact protein species used in the experiment, others choose to model this onto a single organism such as human (Human Protein Reference Database; HPRD)[29]. Additionally, databases may only partially curate a publication, extracting only content that relates to their specific area of interest (InnateDB, HPRD and MPIDB). Whereas none of these policies are in any way wrong, they do create difficulties when attempting to reconcile redundancies between databases. A recent report suggested agreement between databases may be only 54% for curated interactions and 71% in protein identifications, and attributed much of this to the difficulties described above[30]. The effect of curation errors cannot be ignored but a recuration exercise showed this, in fact, to range from only 2% to 9% for several different databases[31].

We firmly believe that the policy followed by the IMEx consortium of taking a coordinated, collaborative rather than competing approach to the integration of protein-interaction data provides the best possible service to the user community. We not only achieve a much broader coverage of the interaction literature published each year than a single database working in isolation can achieve, but we also provide the research community with a single point of access to the data, removing the need to combine records from different databases. The quality-control measures, both internal and cross-database, being developed by the consortium minimize curation error and by supplying data consistently mapped to external reference resources, eliminate errors potentially introduced when identifiers are remapped by third-party resources.

To maintain consistency of mapping we map the IMEx records to the UniProtKB canonical sequence[32] when the isoform is ambiguous and to the specific isoform identifier when it is known, with the corresponding entity in RefSeq[33], mapped at the sequence level, also referenced. To facilitate coordination among resources, we use a scientific publication as the basic unit of IMEx curation. If a publication is curated in the IMEx dataset, it is curated in full, collecting all reported protein-protein interactions into the database, rather than, for example, only those relevant for a specific disease. This enables full data traceability, and, where possible, we provide even more fine-grained data source information by annotating figure or supplement numbers from which the data have been extracted. The quality control measures currently being implemented will also bring down the curation error rates cited above and improve data quality. Turinsky and colleagues concluded[30]: "Many of the discrepancies we identified should in the future be eliminated if the IMEx guidelines are widely followed."

Outlook

We believe that by establishing a network of closely collaborating interaction data resources with a common data representation, query interface and shared curation rules, we are creating a new, reliable and highly visible infrastructure for protein-interaction data collection that will motivate data producers, funding agencies and journals to increasingly make interaction data deposition an integral part of the publication process. Enforcing quality-control checks across the partner databases will improve data quality, and clear statements of our curation policies will make these transparent to users and ensure consistency across the entire IMEx dataset. Regular meetings among IMEX members enable the review of these curation rules and will allow us to rapidly respond to new data types such as quantitative data and dynamic interactions. The IMEx consortium is open to the participation of new partners, and all data producers are encouraged to submit their data to one of the IMEx consortium partners before publication. Detailed information on IMEx consortium membership and data deposition is available at http://www.imexconsortium.org/.

Q5 ACKNOWLEDGMENTS
This work was supported by the European Commission under PSIMEx, contract FP7-HEALTH-2007-223411, and in part by AntiPathoGN, contract number
Q6 HEALTH-F3-2009-223101, SLING, contract number 226073, APO-SYS, contract number FP7-HEALTH-2007-200767), the Ontario Research Fund (GL2-01-030) and Canada Foundation for Innovation (CFI #12301 and CFI #203383).

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

Q7 1. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, d504–d506 (2006).
2. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25**, 894–898 (2007).
3. Orchard, S. & Hermjakob, H. The HUPO proteomics standards initiative - easing communication and minimizing data loss in a changing world. *Brief. Bioinform.* **9**, 166–173 (2008).
4. Alfarano, C. *et al.* The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* **33**, d418–d424 (2006).
5. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
6. Rain, J.C. *et al.* The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211–215 (2001).
7. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, d841–d846 (2012).
8. Ceol, A. *et al.* MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* **38**, d533–d539 (2009).
9. Guldener, U. *et al.* MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34**, d436–d441 (2006).
10. 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).
11. Kerrien, S. *et al.* Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5**, 44 (2007).
12. Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* **39**, d235–d240 (2011).
13. Goll, J. MPIDB: the microbial protein interaction database. *Bioinformatics* **24**, 1743–1744 (2008).
14. Brown, K.R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* **8**, R95 (2007).
15. Lynn, D.J. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4**, 218 (2008).
16. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* **36**, d637–d640 (2008).
17. Perreau, V.M. *et al.* A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease. *Proteomics* **10**, 2377–2395 (2010).
18. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
19. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
20. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
21. 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** (Suppl. 2), s5 (2008).
22. Leitner, F. *et al.* The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. *Nat. Biotechnol.* **28**, 897–899 (2010).
23. Chen, Y.C., Rajagopala, S.V., Stellberger, T. & Uetz, P. Exhaustive benchmarking of the yeast two-hybrid system. *Nat. Methods* **7**, 667–668 (2010).
24. Montecchi-Palazzi, L. *et al.* The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* **9**, 5112–5119 (2009).
25. Turner, B. *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* baq023 (2010).
26. Aranda, B. *et al.* PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* **8**, 528–529 (2011).
27. Prieto, C. & De Las Rivas, J. APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.* **34**, W298–W302 (2006).
28. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, d561–d568 (2011).
29. Keshava Prasad, T.S. *et al.* Human Protein Reference Database—2009 update. *Nucleic Acids Res.* **37**, d767–d772 (2009).
30. Turinsky, A.L. *et al.* Literature curation of protein interactions: measuring agreement across major public databases. *Database* baq026 (2010).
31. Salwinski, L. *et al.* Recurated protein interaction datasets. *Nat. Methods* **6**, 860–861 (2009).
32. UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* **39**, d214–d219 (2011).
33. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38**, d5–d16 (2010).

QUERY FORM

Nature Methods	
Manuscript ID	[Art. Id: 1931]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q1	Please carefully check the spelling and numbering of all author names and affiliations
Q2	Define I2D
Q3	OK?
Q4	As meant? Or explain (rephrase) what you meant by validation
Q5	Please check that all funders have been appropriately acknowledged and that all grant numbers are correct
Q6	Define SLING and APO-SYS
Q7	The disclaimer "however all work reported is freely available has been removed per journal policy. Also reference annotations are not included in Perspectives and have been removed.
Q8	If not correct as edited, clarify or define here what you meant by "anti tag" and "anti bait". Immunoprecipitation via tag and bait?
Q9	Provide common names for all species or remove all
Q10	correct?