Nucleic Acids Research Advance Access published November 24, 2012

Nucleic Acids Research, 2012, 1–6
doi:10.1093/nar/gks1147

InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation

Karin Breuer[1], Amir K. Foroushani[1,2], Matthew R. Laird[1], Carol Chen[1,3], Anastasia Sribnaia[1,3], Raymond Lo[1], Geoffrey L. Winsor[1], Robert E. W. Hancock[3], Fiona S. L. Brinkman[1] and David J. Lynn[2,*]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, V5A1S6, Canada, [2]Animal & Bioscience Research Department, AGRIC, Teagasc, Grange, Dunsany, County Meath, Ireland and [3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, V6T 1Z1, Canada

Received September 15, 2012; Accepted October 24, 2012

ABSTRACT

InnateDB (http://www.innatedb.com) is an integrated analysis platform that has been specifically designed to facilitate systems-level analyses of mammalian innate immunity networks, pathways and genes. In this article, we provide details of recent updates and improvements to the database. InnateDB now contains >196 000 human, mouse and bovine experimentally validated molecular interactions and 3000 pathway annotations of relevance to all mammalian cellular systems (i.e. not just immune relevant pathways and interactions). In addition, the InnateDB team has, to date, manually curated in excess of 18 000 molecular interactions of relevance to innate immunity, providing unprecedented insight into innate immunity networks, pathways and their component molecules. More recently, InnateDB has also initiated the curation of allergy- and asthma-related interactions. Furthermore, we report a range of improvements to our integrated bioinformatics solutions including web service access to InnateDB interaction data using Proteomics Standards Initiative Common Query Interface, enhanced Gene Ontology analysis for innate immunity, and the availability of new network visualizations tools. Finally, the recent integration of bovine data makes InnateDB the first integrated network analysis platform for this agriculturally important model organism.

INTRODUCTION

The innate immune response is a critical branch of immunity, which not only provides a first line of defence against pathogens, but also regulates and shapes subsequent adaptive responses. Innate immunity, however, can also do great harm by driving inappropriate inflammatory cascades. Therefore complex molecular networks are required to regulate innate immunity and maintain appropriate and specific responses to different pathogens, while limiting potential harm from dysregulated inflammation (1–5). The intricate interplay of a multitude of regulatory layers that initiate and coordinate the innate immune response has led to an ever-increasing interest in applying systems-oriented approaches to better understand innate immunity and its modulators (6).

InnateDB (publicly accessible at http://www.innatedb.com and mirrored at http://innatedb.teagasc.ie) is a knowledge base and analysis platform that was specifically designed to provide a system-oriented yet user-friendly tool for integrative analyses of the mammalian innate immune response (7).

INNATEDB CURATION

A key component of the InnateDB project is the contextual manual curation of innate immunity interactions, pathways and their component molecules. The curation process has previously been described in detail (8). InnateDB was first publicly released in 2008 (7). At that time, ~3500 molecular interactions had been curated. By 2010, the database contained 11 786 InnateDB-curated molecular interactions from the review of >3000 published articles. As of September 2012, our curation team

*To whom correspondence should be addressed. Tel: +353 46 9026729; Fax: +353 46 9026154; Email: david.lynn@teagasc.ie

The authors wish to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

has reviewed >4300 publications, and >18 000 interactions of relevance to innate immunity have been annotated (Figure 1; for detailed statistics see http://www.innatedb.com/statistics). More recently, as part of the AllerGen Networks of Centres of Excellence (NCE) (http://www.allergen-nce.com), InnateDB curators have also begun to annotate interactions and pathways of relevance to allergy and asthma. All interactions in InnateDB are provided with detailed contextual information according to the minimum information required for reporting a molecular interaction experiment (MIMIx) standards (9), including the evidence supporting each interaction, the tissue or cell type the interaction was reported in, the type of interaction and the method of detection. New interactions are added to the database weekly, providing up-to-date annotation on the innate immunity interactome. This resource can be mined to identify new relationships between innate immunity and other processes, to identify potential novel regulators of innate immunity and to interpret a user's own data (e.g. gene expression data) from a network biology perspective.

Building a comprehensive list of innate immunity genes

Aside from annotating molecular interactions, InnateDB now also annotates which genes have a published role in innate immunity, providing a brief summary of that role and links to the relevant publications. This data set, available at http://www.innatedb.com/curatedGenes, presently contains >1500 genes (957 human, 527 mouse and 46 bovine) and is the most comprehensive list of genes involved in innate immunity that is available. This list was recently used by a group of researchers to show that human proteins, which are targeted by viruses, are highly enriched for having a role in innate immunity (10).

Contribution to the International Molecular Exchange Consortium

In 2010, InnateDB became a member of the International Molecular Exchange Consortium (IMEx) (11). This organization is dedicated to developing rules for describing molecular interaction data, actively curating these interactions from the scientific literature and making them available through a common website.

Within IMEx, InnateDB has committed to curating every issue of *Nature Immunology* from September 2010 onwards using IMEx curation standards (12). Because IMEx curation requires more annotation detail than the MIMIx level (9) currently supported by InnateDB's submission system, InnateDB curators are submitting these IMEx interactions through the IntAct interaction database (13). On submission, each IMEx interaction is thoroughly reviewed by an IntAct curator before it is

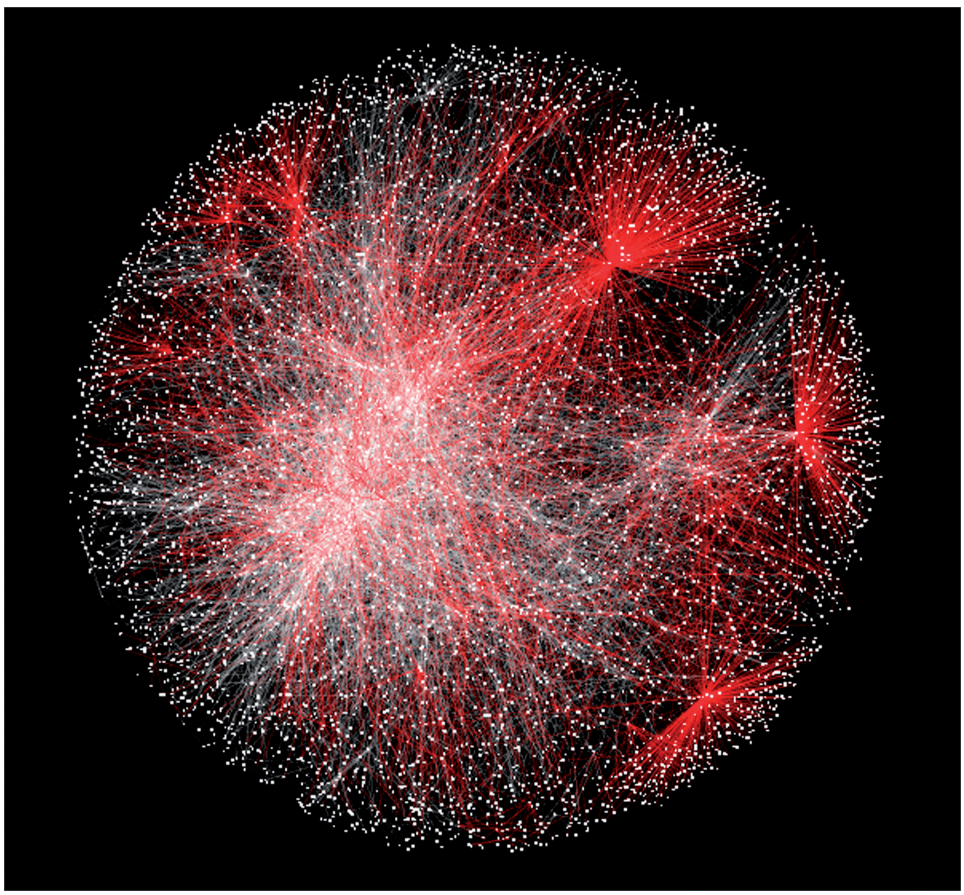

Figure 1. The InnateDB curated interactome in July 2012. Red edges represent interactions that have been added in 2011 and 2012.

accepted and released. In addition to submitting to IntAct, all InnateDB acceptable interactions (i.e. interactions of relevance to innate immunity) from *Nature Immunology* are also deposited into InnateDB.

Integrating data from external resources

To supplement our manual curation efforts and to provide a snapshot of the entire interactome beyond known innate immunity interactions, InnateDB imports data from a wide range of genome, interaction and pathway databases (http://www.innatedb.com/resources). Currently, InnateDB contains 178 000+ imported experimentally validated interactions, 3000+ pathways and 300 000+ interactions based on Ortholuge (14) orthology predictions (interologs) in addition to the 18 000+ InnateDB manually curated interactions.

INTEGRATION OF BOVINE DATA—ORTHOLOGY-BASED PATHWAY & NETWORK RECONSTRUCTION

In February 2012, a new version of InnateDB was released that included the incorporation of bovine gene, pathway and molecular interaction annotation in addition to the existing data for human and mouse. This new version of the platform now also facilitates a systems biology approach to the investigation of the bovine innate immune response and is poised to deepen our understanding of important bovine infectious diseases associated with significant economic losses (e.g. bovine tuberculosis and mastitis), as well as enabling cross-species comparisons of innate immunity.

As bovine experimentally validated interactions and pathways are virtually non-existent, InnateDB uses an orthology-based approach to predict bovine pathways and interactions primarily from human data. One should be aware that this approach results in a humanized and frequently incomplete representation of the bovine interactome, but in the absence of widespread experimental data it provides at least a network biology framework to build on and to generate hypotheses that can be subsequently experimentally validated. InnateDB experimentally validated and predicted interactions are clearly labelled. As of September 2012, InnateDB contains >70 000 bovine interologs (interactions based on orthology) involving 10 717 bovine genes. In each case, one can link back to the orthologous human interaction to review evidence for the interaction.

The latest release of InnateDB also uses orthology predictions to transfer human and mouse pathway annotations to bovine genes in real time. Currently, pathway annotations can be assigned to 7032 bovine genes by orthology to human genes. Notably, although only ~70% of all human genes (14 316 genes) have a predicted bovine ortholog, and a significantly higher proportion (85%) of human genes with pathway annotations have a bovine ortholog. This higher prevalence of conserved genes among pathway-annotated genes indicates that many of the associated processes may be well preserved.

To further examine the appropriateness of the orthology-based annotation transfer on a per-pathway basis, we determined the 'conservation rate' (*cons*) of each pathway, defined as the ratio of pathway participants in the source organism (human/mouse) that have a putative counterpart in the target organism (cow) to the total number of participants in the source organism. As of September 2012, InnateDB contains 1536 human pathways with five or more pathway participants, 80% (1257 pathways) of these have a conservation rate of 0.8 or better. The corresponding number for a conservation rate of ≥0.7 is 93% (1442 pathways). The high prevalence of strongly conserved pathways seems to largely justify an orthology-based approach for inferring bovine pathways. Supplementary Table S1 lists the remaining 107 pathways with a relatively low conservation rate (*cons* <0.7). Notably, the list of pathways for which an orthology-based reconstruction is challenging includes >30 immunologically important pathways. In some cases, the low conservation rate can be attributed to real divergence of the underlying processes. The bovine Type I Interferon family, for example, has been shown to have undergone widespread expansion, including the divergence of a new Type I interferon (IFN) family (IFNX) in the cow from IFN alpha (15). In other cases, the conservation rate might further increase with future improvements to the quality of the bovine draft genome.

In addition to orthology-based annotation transfer, the tissue expression profile of a gene can provide some insight into its potential function (16). Through collaboration with colleagues at the United States Department of Agriculture, InnateDB now integrates bovine tissue expression data for >13 000 genes. This data was sourced from the Bovine Gene Atlas (17), which has profiled gene expression across 87 different bovine tissues using a next generation sequencing approach. A graphical tissue expression profile is available on the Gene Card page of bovine genes.

INNATEDB DATA ANALYSIS AND VISUALIZATION

InnateDB can serve as a knowledge base where users can search for particular genes or proteins of interest and their associated interactions and pathways, using a variety of search fields. Alternatively, InnateDB can be queried in a more high-throughput fashion, where users can upload a list of genes/proteins and associated quantitative data from up to 10 different conditions (e.g. gene expression data) and carry out more complex searches and analyses (Figure 2). After uploading a list of gene IDs (human, mouse and bovine Ensembl, RefSeq, Entrez Gene, UniProt and InnateDB IDs are all accepted), users can quickly find which pathways, Gene Ontologies (including enhanced innate immunity gene annotation) or transcription factor binding sites are statistically over-represented in their dataset. Users can also use InnateDB to build, visualize and analyse molecular interaction networks consisting of their uploaded genes and their encoded products. One can, for example, construct a network of how differentially expressed genes interact

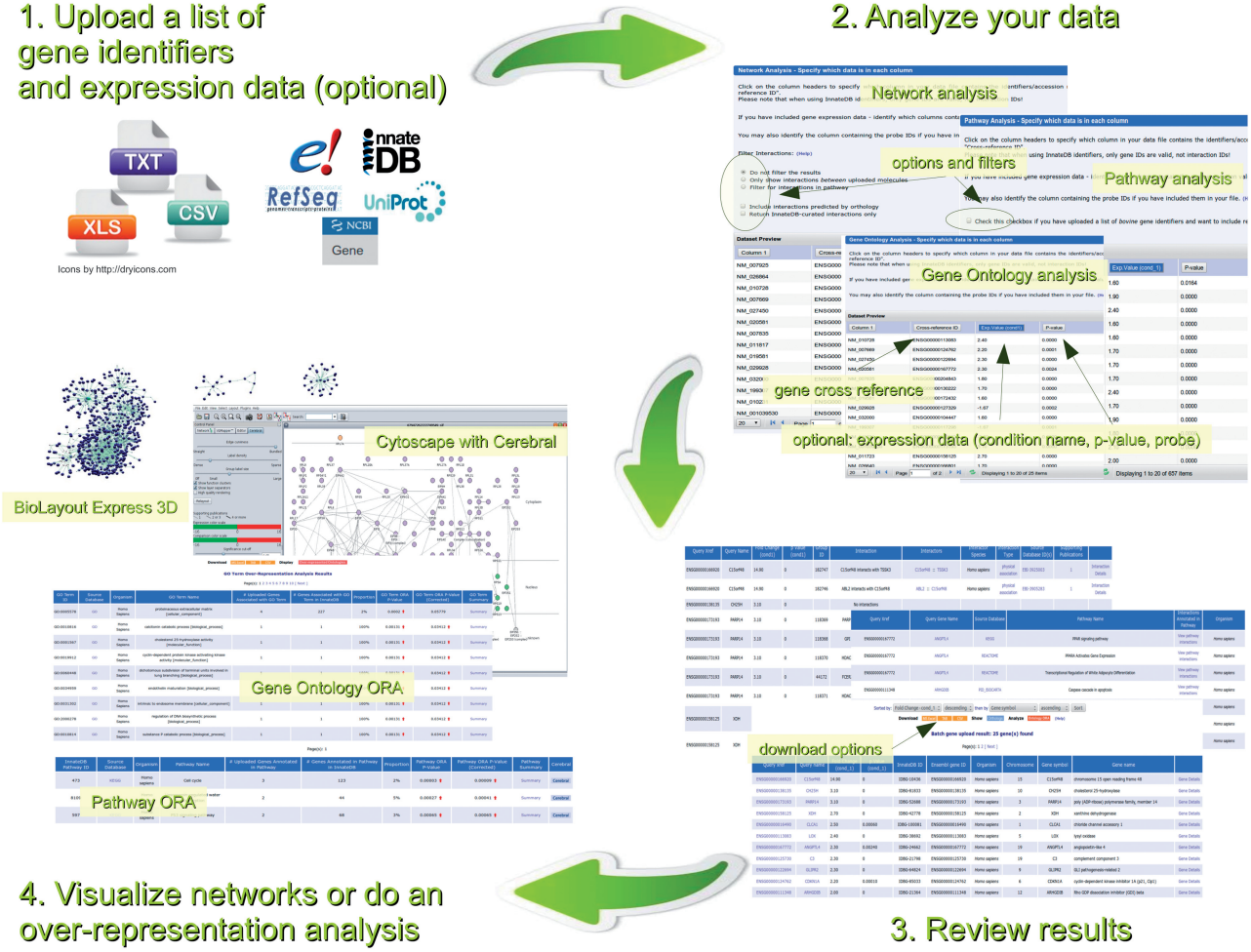

Figure 2. Data analysis workflow in InnateDB.

with one another. Quantitative data uploaded by the user is automatically overlaid on these networks. Recent improvements to InnateDB include the option to incorporate interactions based on orthology in the construction of molecular interaction networks and the option to restrict the networks to contain only InnateDB manually curated interactions. Further enhancements to the web-interface include more intuitive page layouts, faster searches and analyses, and a variety of other changes (see http://www.innatedb.com/news).

Network visualization tools

All interactions in InnateDB may be downloaded in several standardized formats including text-based formats (tab, csv, xls), the simple interaction format (sif) and the PSI-MI XML 2.5 and MITAB formats (18). Additionally, interaction networks may also be visualized in our Cerebral program (19), a Java plugin for the Cytoscape network visualization software (20,21), which uses subcellular localization information to orientate interaction networks in a more biologically intuitive pathway-like layout. Networks can also be visualized in other third-party software including the CyOog plugin (22), which uses Power Graph analysis to reduce network complexity by explicitly representing re-occurring network motifs. Recently, we have also integrated BioLayout Express 3D 2.2 (23), an application designed for the visualization, clustering and analysis of large networks in 2D and 3D space.

Proteomics Standards Initiative Common Query Interface implementation

Interaction data in InnateDB can now also be queried using web services implementing The Proteomics Standards Initiative Common Query Interface (PSICQUIC) (24). PSICQUIC is an effort from the Human Proteome Organization Proteomics Standards Initiative (http://www.hupo.org/research/psi/) to standardize programmatic access to molecular interaction databases based on the PSI standard formats (PSI-MI XML and MITAB) (18). It defines standard web services and also a query syntax for powerful and flexible searches.

All data sources implementing PSICQUIC can be queried in the exact same way, i.e. the same query can be used to retrieve the relevant data from many different interaction data sources. Independently published

observations of an experimental system, curated by independent databases, are then integrated in response to a single user query (see http://www.ebi.ac.uk/intact/imex). PSICQUIC web services are RESTful (REpresentational State Transfer) but can also be accessed through SOAP (Simple Object Access Protocol). A list of available services for InnateDB can be found at http://imex.innatedb.com/psicquic-ws/webservices. InnateDB updates the data files for the PSICQUIC web services weekly and additionally provides them for download in a compressed format at http://www.innatedb.com/downloads.

ONGOING DEVELOPMENTS

InnateDB will maintain its curation efforts to annotate interactions and genes of relevance to innate immunity, with weekly updated annotation, thus continuing to provide a comprehensive platform for systems and network biology analyses of innate immune-associated responses. Continued incorporation of data from external resources, encompassing the wider human, mouse and bovine interactomes, will also continue to facilitate analyses beyond innate immunity by a wide range of researchers. Additionally, InnateDB intends to expand beyond the curation of innate immunity relevant networks, incorporating more adaptive immunity information. We are currently developing a first version of an Allergy and Asthma Portal that will further integrate data on allergy and associated immune interactions from both the literature and researchers from AllerGen. This portal will be built on InnateDB and will provide an analysis platform for more sophisticated network biology-based investigations of allergy and asthma responses. These interactions will be identifiable from innate immunity interactions, so that users can continue to have focused analyses on the innate immunity interactome.

Further future developments will include improvements to InnateDB pathway analysis tools. The over-representation–based methods for pathway analysis that are currently available through InnateDB's data analysis interface are widely established and considered a 'gold standard'; yet, they neglect the fact that many components are shared between seemingly unrelated pathways. To address this issue, we have used the InnateDB collection of pathway annotations as a basis to identify pairs of genes that co-occur only in a single pathway and developed a novel pathway analysis method [signature over-representation analysis (SIGORA)] that focuses on the over-representation of such gene pairs in a list of genes of interest (Foroushani et al, submitted). SIGORA is currently implemented as an R package [available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/sigora/index.html] and will be integrated into the future releases of InnateDB.

Finally, together with the PSICQUIC development team and other IMEx members, we are working on an improved reference implementation of PSICQUIC. We are also preparing to export our data in MITAB 2.7 format.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Matthew Whiteside for providing orthology predictions in InnateDB. The authors would also like to acknowledge Tad Sonstegard, Lakshmi Matukumalli and other members of the Bovine Gene Atlas project for providing the bovine tissue expression data. We would also like to thank all the previous members of the InnateDB development and curation teams (http://www.innatedb.com/about). Thanks also to the various interaction, pathway and annotation databases that have been integrated into InnateDB for freely providing their data to the public. Grateful thanks also go to the many researchers who have taken the time to respond to our queries regarding curation of their publications.

FUNDING

This work was supported by Genome BC through the Pathogenomics of Innate Immunity (PI2) project and by the Foundation for the National Institutes of Health and the Canadian Institutes of Health Research under the Grand Challenges in Global Health Research Initiative [Grand Challenges ID: 419]. Further funding was also provided by AllerGen grants 12ASI1 and 12B&B2. D.J.L. was funded in part during this project by a post-doctoral trainee award from the Michael Smith Foundation for Health Research (MSFHR). F.S.L.B. is a MSFHR Senior Scholar and R.E.W.H. holds a Canada Research Chair (CRC). Funding to enable bovine systems biology in InnateDB is provided by Teagasc [RMIS6018] and the Teagasc Walsh Fellowship scheme. IMEx is funded by the European Commission under the PSIMEx project [contract number FP7-HEALTH-2007-223411]. Funding for open access charge: Teagasc [RMIS6018].

Conflict of interest statement. None declared.

REFERENCES

1. Delano,M.J., Thayer,T., Gabrilovich,S., Kelly-Scumpia,K.M., Winfield,R.D., Scumpia,P.O., Cuenca,A.G., Warner,E., Wallet,S.M., Wallet,M.A. *et al.* (2011) Sepsis induces early alterations in innate immunity that impact mortality to secondary infection. *J. Immunol.*, **186**, 195–202.
2. Karin,M., Lawrence,T. and Nizet,V. (2006) Innate immunity gone awry: linking microbial infections to chronic inflammation and cancer. *Cell*, **124**, 823–835.
3. Lin,W.W. and Karin,M. (2007) A cytokine-mediated link between innate immunity, inflammation, and cancer. *J. Clin. Invest.*, **117**, 1175–1183.
4. Shi,F.D., Ljunggren,H.G. and Sarvetnick,N. (2001) Innate immunity and autoimmunity: from self-protection to self-destruction. *Trends Immunol.*, **22**, 97–101.
5. Wen,L., Ley,R.E., Volchkov,P.Y., Stranges,P.B., Avanesyan,L., Stonebraker,A.C., Hu,C., Wong,F.S., Szot,G.L., Bluestone,J.A. *et al.* (2008) Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature*, **455**, 1109–1113.

6. Gardy,J.L., Lynn,D.J., Brinkman,F.S. and Hancock,R.E. (2009) Enabling a systems biology approach to immunology: focus on innate immunity. *Trends Immunol.*, **30**, 249–262.
7. Lynn,D.J., Winsor,G.L., Chan,C., Richard,N., Laird,M.R., Barsky,A., Gardy,J.L., Roche,F.M., Chan,T.H., Shah,N. *et al.* (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.*, **4**, 218.
8. Lynn,D.J., Chan,C., Naseer,M., Yau,M., Lo,R., Sribnaia,A., Ring,G., Que,J., Wee,K., Winsor,G.L. *et al.* (2010) Curating the innate immunity interactome. *BMC Syst. Biol.*, **4**, 117.
9. Orchard,S., Salwinski,L., Kerrien,S., Montecchi-Palazzi,L., Oesterheld,M., Stumpflen,V., Ceol,A., Chatr-aryamontri,A., Armstrong,J., Woollard,P. *et al.* (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **25**, 894–898.
10. Pichlmair,A., Kandasamy,K., Alvisi,G., Mulhern,O., Sacco,R., Habjan,M., Binder,M., Stefanovic,A., Eberle,C.A., Goncalves,A. *et al.* (2012) Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature*, **487**, 486–490.
11. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S., Cesareni,G. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
12. Orchard,S., Kerrien,S., Jones,P., Ceol,A., Chatr-Aryamontri,A., Salwinski,L., Nerothin,J. and Hermjakob,H. (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, **7(Suppl.1)**, 28–34.
13. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
14. Fulton,D.L., Li,Y.Y., Laird,M.R., Horsman,B.G., Roche,F.M. and Brinkman,F.S. (2006) Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, **7**, 270.
15. Walker,A.M. and Roberts,R.M. (2009) Characterization of the bovine type I IFN locus: rearrangements, expansions, and novel subfamilies. *BMC Genomics*, **10**, 187.
16. Dezso,Z., Nikolsky,Y., Sviridov,E., Shi,W., Serebriyskaya,T., Dosymbekov,D., Bugrim,A., Rakhmatulin,E., Brennan,R.J., Guryanov,A. *et al.* (2008) A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.*, **6**, 49.
17. Harhay,G.P., Smith,T.P., Alexander,L.J., Haudenschild,C.D., Keele,J.W., Matukumalli,L.K., Schroeder,S.G., Van Tassell,C.P., Gresham,C.R., Bridges,S.M. *et al.* (2010) An atlas of bovine gene expression reveals novel distinctive tissue characteristics and evidence for improving genome annotation. *Genome Biol.*, **11**, R102.
18. Kerrien,S., Orchard,S., Montecchi-Palazzi,L., Aranda,B., Quinn,A.F., Vinod,N., Bader,G.D., Xenarios,I., Wojcik,J., Sherman,D. *et al.* (2007) Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
19. Barsky,A., Gardy,J.L., Hancock,R.E. and Munzner,T. (2007) Cerebral: a cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, **23**, 1040–1042.
20. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
21. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
22. Royer,L., Reimann,M., Andreopoulos,B. and Schroeder,M. (2008) Unraveling protein networks with power graph analysis. *PLoS Comput. Biol.*, **4**, e1000108.
23. Theocharidis,A., van Dongen,S., Enright,A.J. and Freeman,T.C. (2009) Network visualization and analysis of gene expression data using BioLayout express(3D). *Nat. Protoc.*, **4**, 1535–1550.
24. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.