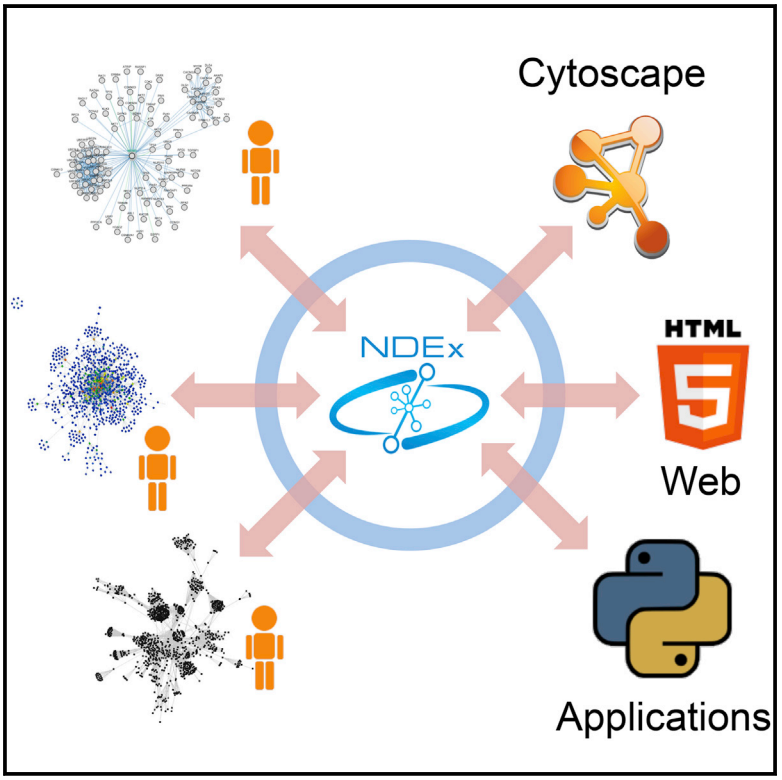
Brief Report

Cell Systems

NDEx, the Network Data Exchange

Graphical Abstract



Highlights

- NDEx (www.ndexbio.org) is an online commons for biological networks
- Users can upload, share, and distribute networks of many types, sizes, and formats
- Developers can access NDEx via a web-based programming interface
- NDEx promotes the publication of networks as dynamic, actionable data

Authors

Dexter Pratt, Jing Chen, David Welker, ..., Jan Kuentzer, Barry Demchak, Trey Ideker

Correspondence

depratt@ucsd.edu

In Brief

NDEx (www.ndexbio.org) is an online commons where scientists can upload, share, and publicly distribute biological networks of many types, sizes, and formats. It promotes the publication of networks as dynamic, actionable data and the development of applications using networks.

Pratt et al., 2015, Cell Systems *1*, 302–305
October 28, 2015
http://dx.doi.org/10.1016/j.cels.2015.10.001

CellPress

CellPress

Cell Systems
Brief Report

NDEx, the Network Data Exchange

Dexter Pratt,[1,*] Jing Chen,[1] David Welker,[1] Ricardo Rivas,[1] Rudolf Pillich,[1] Vladimir Rynkov,[1] Keiichiro Ono,[1] Carol Miello,[2] Lyndon Hicks,[3] Sandor Szalma,[4] Aleksandar Stojmirovic,[5] Radu Dobrin,[5] Michael Braxenthaler,[6] Jan Kuentzer,[7] Barry Demchak,[1] and Trey Ideker[1,8]

[1]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA
[2]Pfizer Inc., Eastern Point Road, Groton, CT 06340, USA
[3]Pfizer Inc., 1 Burtt Road, Andover, MA 01810, USA
[4]Janssen Research and Development LLC, 3210 Merryfield Row, San Diego, CA 92121, USA
[5]Janssen Research and Development LLC, 1400 McKean Road, Spring House, PA 19477, USA
[6]Roche, Pharma Research and Early Development Informatics, Roche Innovation Center, New York, NY, 10016, USA
[7]Roche, Pharma Research and Early Development Informatics, Roche Innovation Center, Penzberg 82377, Germany
[8]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA
*Correspondence: depratt@ucsd.edu
http://dx.doi.org/10.1016/j.cels.2015.10.001

SUMMARY

Networks are a powerful and flexible methodology for expressing biological knowledge for computation and communication. Network-encoded information can include systematic screens for molecular interactions, biological relationships curated from literature, and outputs from analyses of Big Data. NDEx, the Network Data Exchange (www.ndexbio.org), is an online commons where scientists can upload, share, and publicly distribute networks. Networks in NDEx receive globally unique accession IDs and can be stored for private use, shared in pre-publication collaboration, or released for public access. Standard and novel data formats are accommodated in a flexible storage model. Organizations can use NDEx as a distribution channel for networks they generate or curate. Developers of bioinformatic applications can store and query NDEx networks via a common programmatic interface. NDEx helps expand the role of networks in scientific discourse and facilitates the integration of networks as data in publications. It is a step toward an ecosystem in which networks bearing data, hypotheses, and findings flow easily between scientists.

Networks are a precise and computable form in which biologists can express many kinds of information, including models of biological mechanisms, experimental facts, and relationships derived by systematic data analysis. When pathway diagrams evolved into repositories of small pathway networks (Croft et al., 2014; Kanehisa and Goto, 2000; Ogata et al., 1999), they became searchable resources and the basis for data interpretation and collaborative pathway editing applications (van Iersel et al., 2008). The emergence of repositories of large networks of molecular relationships (Franceschini et al., 2013; Orchard et al., 2014; Stark et al., 2011; Warde-Farley et al., 2010) in both simple and complex formats (Demir et al., 2010; Le Novère et al., 2006; Mi et al., 2010; OpenBEL, 2011) condensed collections of data into structured findings useful for hypothesis generation and computational prediction (Bandyopadhyay et al., 2006; QIAGEN, 2015; Vandin et al., 2012). In recent years, there has been rapid progress in the construction of networks inferred by the systematic processing of genome-scale information, providing an important avenue for interpretation and a counterpoint to literature curation (Califano et al., 2012; Chuang et al., 2007; Hofree et al., 2013). By providing a flexible computable medium for biological knowledge, networks are also becoming a critical element for new models of scientific publication in which data and their derivatives are as important as text (CyNetShare, 2014).

NDEx, the Network Data Exchange, is an open-source (Data S1, Open Source) software framework that facilitates the sharing of networks of many types and formats, the publication of networks as data, and the use of networks in modular software. In comparison with repositories such as IntAct (Orchard et al., 2014) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), where network content in specific formats is managed by the organization maintaining the resource, NDEx is a data commons where users manage the sharing and publication of their own networks (Data S1, Related Resources). It accommodates networks of any type, from pathway models and interaction maps to novel data-driven knowledge, handling diverse formats including simple interaction format (SIF), extensible graph markup and modeling language (XGMML), BioPAX3 and OpenBEL (Data S1, Network Formats). It promotes scientific publication and reproducibility by enabling the tracking of accession and provenance of networks. Finally, NDEx provides a flexible, programmatically accessible storage service that promotes modular software development and workflows in which networks that are output from one application can be input to another. NDEx does not perform biological analysis and visualization itself but, instead, enables the interchange of networks between applications that do.

As a data commons, NDEx enables scientists and organizations that create accounts on the NDEx server to upload and save networks and to create communities of users, much like Google+ Circles or LinkedIn Groups. They can manage access to their networks, making them private, public, or shared with selected users and community groups, similar to shared document systems such as Google Docs or DropBox (Figure 1A; Data S1, NDEx Basics). The shared networks preserve the

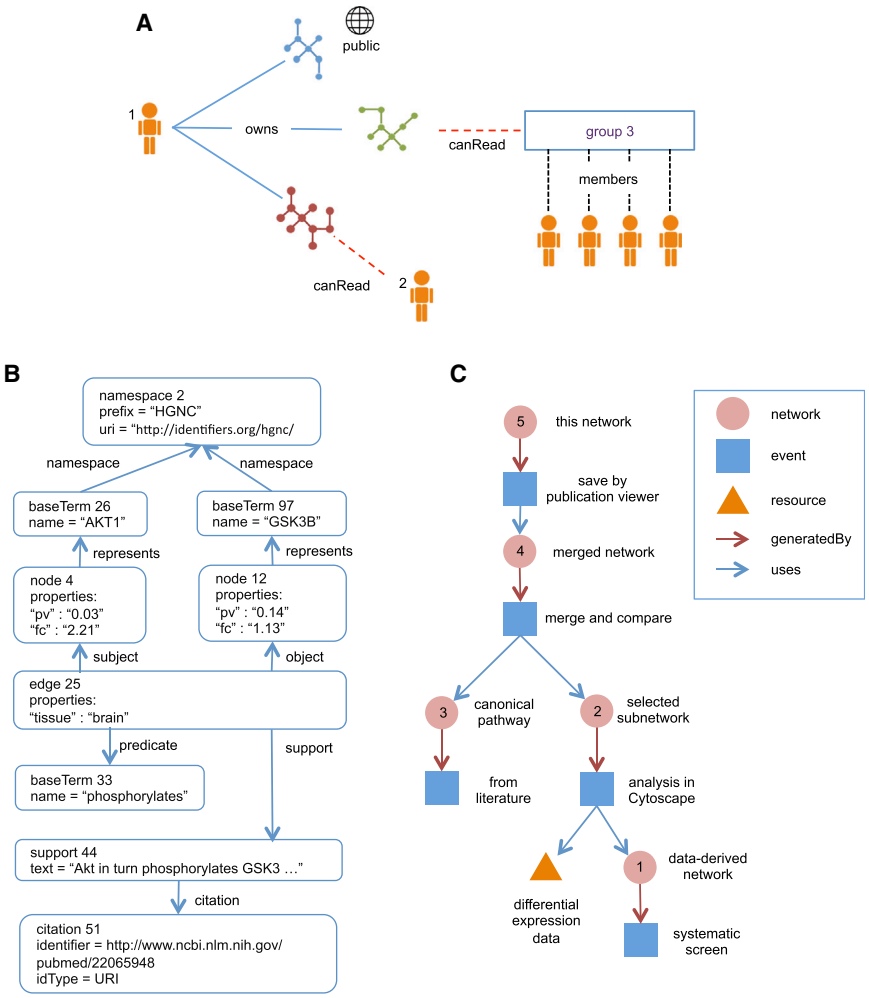


Figure 1. Access Control, Network Data Structures, and Provenance History in NDEx

(A) Examples of access control relationships for networks in NDEx. User 1 owns the red, green, and blue networks. She shares the red network directly with user 2, the green network with the members of community group 3, and makes the blue network a public network available to any user or by anonymous query.

(B) Example of one edge represented in the NDEx network data model (Data S1, Data Model). Each box in the diagram is a network element, labeled with its type and ID. Edge 25 connects nodes 4 and 12 by the subject and object relationships. The meaning of edge 25 is set by the predicate relationship to baseTerm 33, "phosphorylates." BaseTerm objects define the vocabulary used by the network, and the primary meaning of node 4 is set by the represents relationship to baseTerm 26, "AKT1." Node 12 represents baseTerm 97, "GSK3B." Both baseTerm 97 and baseTerm 26 are associated with namespace 2, indicating that they are standard human gene symbols. Both nodes have the user-defined properties "fc" and "pv" associated with them, used to record differential expression data that was mapped onto the network. Edge 25 has a user-defined property, "tissue" = "brain," used by the authors to indicate the tissue context. The edge is also annotated with evidence text by support 44 associated with citation 51, the article from which the text was derived.

(C) Abstract representation of the provenance history for network 5 in Figure 2. The provenance history records the workflow that led to the network as a tree structure of events, NDEx networks, and other resources.

distinct semantics of their original formats while standardizing the treatment of identifiers, citations, properties, and network topology (Figure 1B; Data S1, Data Model). NDEx therefore differs in approach from WikiPathways (Pico et al., 2008), a pioneering collaborative platform for the curation of biological pathways, where all documents are edited publically and use a single format. Organizations that publish network content can use NDEx as a channel for distribution. Networks from the NCI Pathway Interaction Database (Schaefer et al., 2009), Pathway Commons (Cerami et al., 2011), and the OpenBEL Consortium are among those available in NDEx (Data S1, Metrics).

To support the publication of networks as data, it must be possible to unambiguously specify the identity of the network and trust that the content of a published network will remain constant. NDEx provides accession identifiers for every network, assigning a universal unique identifier (UUID) that distinguishes it from all other networks across all servers. The owner of an NDEx network can set its status to be read-only, preventing further edits. These features enable networks to be reliable, consistent references suitable as inputs to further research.

When networks expressing data, hypotheses, and findings are both inputs and outputs of analysis and are referenced in publications, it becomes important to know how and when a network was created and which inputs and algorithms would be required to reproduce it. NDEx addresses these needs by including the "provenance history" (Figure 1C; Data S1, Provenance) with each network. The provenance history captures the workflow leading to the current network by describing prior events, networks, and other resources. The history grows as networks are created, modified, used, or copied. It incorporates concepts and vocabulary from ongoing work in provenance annotation (Ciccarese et al., 2013; DublinCore, 2012; PROV-O, 2013) but adopts a strategy where referenced resources are described rather than simply linked, preserving information in cases where the resources are later removed, altered, or become unavailable.

NDEx promotes the development of new network analysis algorithms and applications by expanding access to networks as inputs by facilitating immediate sharing of network results and by providing a path to publication. An NDEx server stores and manages networks and all related information and can be accessed by applications via a web-based relational state transfer application programming interface (REST API) (Fielding and Taylor, 2002). NDEx client libraries for this API have been created in the Java, Python, and R languages to facilitate easy use by scientists, although NDEx can be accessed via any language capable of internet communications (Data S1, REST API). The

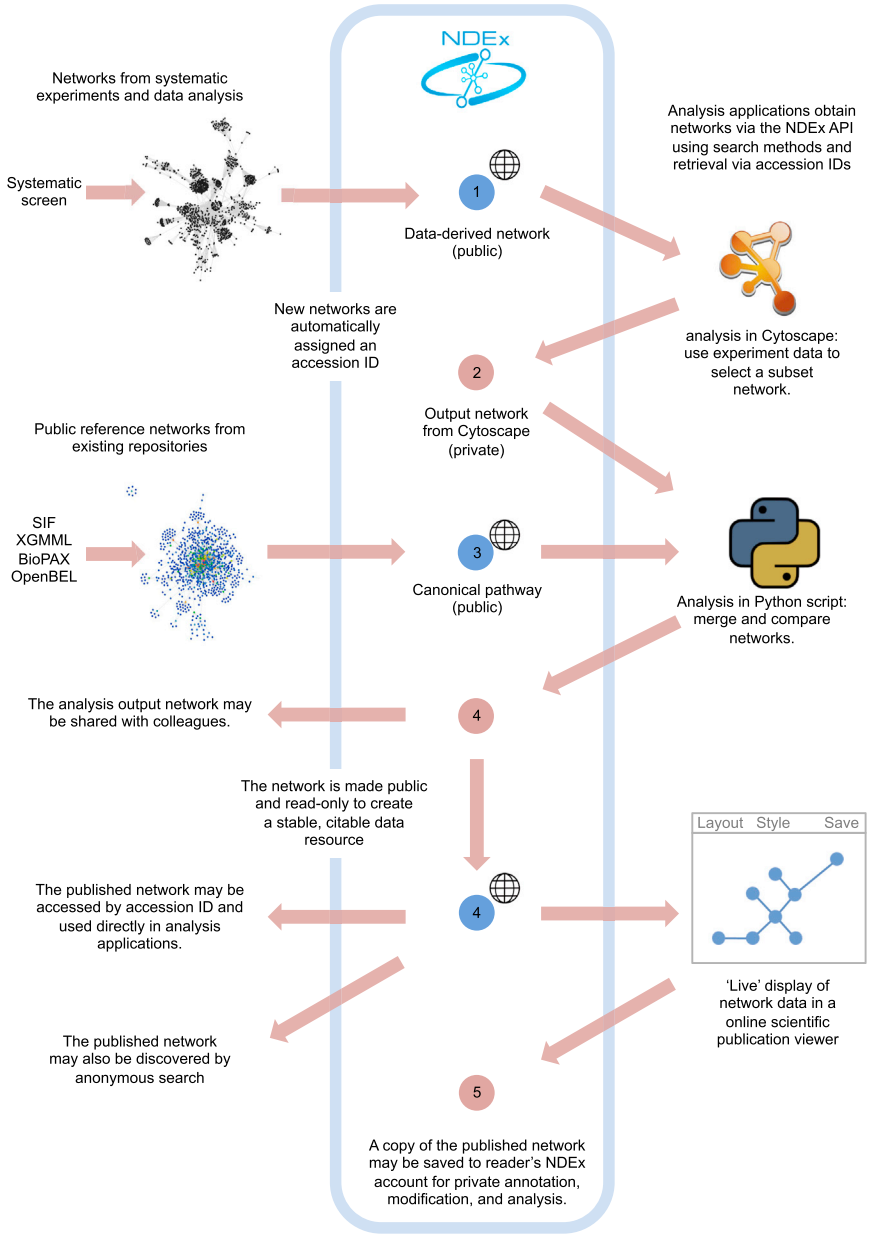


Figure 2. NDEx Workflow

Example workflow in which network 1 is created by systematic analysis of genome scale data and stored in NDEx, network 2 is produced by a Cytoscape analysis that takes network 1 as an input, network 3 represents a canonical pathway uploaded from literature, and network 4 is the output of a bioinformatic script that operates on networks 2 and 3. Network 4 is made public and read-only and becomes part of a publication. Network 4 is viewed by readers of the publication using an NDEx-capable web application that enables them to directly act on the network data, such as saving a private copy to an NDEx account as network 5.

NDEx website (www.ndexbio.org) is the most comprehensive example of an NDEx-enabled application that accesses the public NDEx server via the REST API. The website enables visitors to anonymously search, browse, and query networks and logged-in users to upload content, manage groups, and share networks. Simple analysis scripts (e.g., Python or R) can also query NDEx via the API to obtain input networks and then save analysis result networks directly to NDEx (Figure 2). This storage service model enables the researcher to focus on the core data analysis or algorithm rather than on the management, storage, and publication of networks.

The rich Cytoscape biological analysis and visualization environment (Shannon et al., 2003) can also access NDEx via the REST API, enabling Cytoscape users to search, import, and export networks. Under Cytoscape, a workflow might start by importing a transcriptional regulatory network from NDEx, after which the user could annotate the network with a differential mRNA expression dataset and process it to find subnetworks enriched for genes with significant changes in mRNA expression. The user could then export the subnetworks back to NDEx for review by collaborators or for use as input to further analyses. The CyNDEx App (Data S1, CyNDEx) implements access to NDEx, and upcoming releases of Cytoscape are expected to incorporate its functionality into the main application, making NDEx networks immediately available to users.

An important use of NDEx is to enable new models of scientific publication via network visualization applications in which live data structures replace static diagrams and supplemental files. Readers can immediately act on networks published via NDEx as data that can be dynamically visualized, inspected, and manipulated. For example, a biologist might save selected portions of a published network to their NDEx account to capture a mechanism of interest. Both the original and saved networks would be accessible to other NDEx-capable applications for analysis and visualization. This integration of viewing, annotation, sharing, and action can accelerate and enrich the process of scientific communication.

Finally, users can download and deploy the NDEx server software for private uses that would be impractical or unsupported on the shared public server. An NDEx server can be installed behind a firewall to handle cases where strong security is required, enabling storage of proprietary networks developed for the health sciences industry or those that incorporate patient information subject to privacy standards (e.g., the Health Insurance Portability and Accountability Act [HIPAA]). A private NDEx can also be deployed on local servers or on a scientist's desktop for applications that store very large networks or perform frequent, large transactions. Applications can

simultaneously access both public and private NDEx servers, or users can coordinate private NDEx instances with a public server using NDEx Sync (Data S1, NDEx Sync), a command line utility that can copy and update selected networks between servers.

In summary, NDEx provides distinctive capabilities as a data commons to further the use of biological networks in scientific discourse. It promotes the development of modular applications and the re-use of research products by creating a network exchange where the outputs from one project or application can readily become inputs to another. The NDEx platform is enabling new forms of publication and collaboration in which network information can be immediately analyzed, visualized, annotated, and shared.

SUPPLEMENTAL INFORMATION

Supplemental Information includes NDEx Data Exchange information and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2015.10.001.

AUTHOR CONTRIBUTIONS

D.P., J.C., D.W., R.R., V.R., and R.P. designed, implemented, tested, and documented the NDEx software. K.O. wrote the Network Publication Viewer. D.P. and T.I. wrote the paper. B.D., C.M., L.H., S.S., A.S., J.K., R.D., and M.B. provided valuable guidance, review, and technical support. T.I., S.S., and D.P. conceived the project, and D.P. and T.I. directed its execution.

ACKNOWLEDGMENTS

NDEx is supported in part by F. Hoffmann-La Roche Ltd, Janssen Research and Development, LLC, Pfizer, Inc. C.M. and L.H. are employees of Pfizer, Inc. S.S., A.S., and R.D. are employees of Janssen Research and Development, LLC. M.B. and J.K. are employees of F. Hoffmann-La Roche Ltd. The industry funding for NDEx is managed as a project of the Cytoscape Consortium, a non-profit corporation. T.I. is a board member and officer of the Cytoscape Consortium. B.D. is an officer of the Cytoscape Consortium. NDEx is also supported in part by the National Cancer Institute under award U24 CA-184427.

Received: March 19, 2015
Revised: September 14, 2015
Accepted: October 6, 2015
Published: October 28, 2015

REFERENCES

Bandyopadhyay, S., Kelley, R., and Ideker, T. (2006). Discovering regulated networks during HIV-1 latency and reactivation. Proceedings of the 2006 Pacific Symposium on Biocomputing, pp. 354-366.

Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat. Genet. *44*, 841–847.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. *39*, D685–D690.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. Mol. Syst. Biol. *3*, 140.

Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J.G., Goble, C., and Clark, T. (2013). PAV ontology: provenance, authoring and versioning. J. Biomed. Semantics *4*, 37.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Res. *42*, D472–D477.

CyNetShare (2014). http://cynetshare.ucsd.edu.

Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., et al. (2010). The BioPAX community standard for pathway data sharing. Nat. Biotechnol. *28*, 935–942.

DublinCore (2012). Dublin Core Metadata Element Set, Version 1.1 http://www.dublincore.org/documents/dces/.

Fielding, R.T., and Taylor, R.N. (2002). Principled design of the modern Web architecture. ACM Trans. Internet Technol. *2*, 115–150.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L.J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808–D815.

Hofree, M., Shen, J.P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. Nat. Methods *10*, 1108–1115.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., et al. (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. *34*, D689–D691.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P.D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. *38*, D204–D210.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *27*, 29–34.

OpenBEL (2011). OpenBEL http://www.openbel.org.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363.

Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R., and Evelo, C. (2008). WikiPathways: pathway editing for the people. PLoS Biol. *6*, e184.

PROV-O (2013). PROV-O: The PROV Ontology http://www.w3.org/TR/prov-o/.

QIAGEN (2015). Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity).

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res. *37*, D674–D679.

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.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. *39*, D698–D704.

van Iersel, M.P., Kelder, T., Pico, A.R., Hanspers, K., Coort, S., Conklin, B.R., and Evelo, C. (2008). Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics *9*, 399.

Vandin, F., Clay, P., Upfal, E., and Raphael, B.J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. Proceedings of the 2012 Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, pp. 55-66.

Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. *38*, W214–W220.