Andrea Argentini[1–3], Ludger J E Goeminne[1–5], Kenneth Verheggen[1–3], Niels Hulstaert[1–3], An Staes[1,2], Lieven Clement[3,4] & Lennart Martens[1–3]

[1]Department of Medical Protein Research, Ghent, Belgium. [2]Department of Biochemistry, Ghent University, Ghent, Belgium. [3]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium. [4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [5]Department of Plant Systems Biology, VIB, Ghent University, Zwijnaarde, Belgium.
e-mail: lennart.martens@ugent.be

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.4075).

ACKNOWLEDGMENTS

This research was supported by the Ghent University Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” (01MR0410) to L.J.E.G., L.C. and L.M., Ghent University grant BOF12/GOA/014 to N.H. and L.M., the IWT SBO grant “INSPECTOR” (120025) to A.A. and L.M., and IWT SBO grant “Differential proteomics at peptide, protein and module level” (141573) to L.J.E.G.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Vaudel, M., Sickmann, A. & Martens, L. *Proteomics* **10**, 650–670 (2010).
2. Liu, N.Q. *et al. J. Proteome Res.* **12**, 4627–4641 (2013).
3. Sandin, M., Teleman, J., Malmström, J. & Levander, F. *Biochim. Biophys. Acta* **1844** (1 Pt A), 29–41 (2014).
4. Cox, J. & Mann, M. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
5. Vizcaíno, J.A. *et al. Nat. Biotechnol.* **32**, 223–226 (2014).
6. Vaudel, M. *et al. Proteomics* **16**, 214–225 (2016).
7. Vaudel, M. *et al. Nat. Biotechnol.* **33**, 22–24 (2015).
8. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A. & Martens, L. *Proteomics* **11**, 996–999 (2011).
9. Argentini, A. et al. Protocol Exchange http://dx.doi.org/10.1038/protex.2016.085 (2016).
10. Paulovich, A.G. *et al. Mol. Cell. Proteomics* **9**, 242–254 (2010).
11. Staes, A. *et al. Anal. Chem.* **85**, 11054–11060 (2013).

OmniPath: guidelines and gateway for literature-curated signaling pathway resources

To the Editor: Resources that capture information about signaling pathways from the literature are essential for the experimental design and analysis of many biological studies. Multiple resources are available that have different focuses and levels of granularity[1], making it often unclear which should be used, either alone or in combination, in a given situation (**Supplementary Results 1–3**).

We performed a systematic analysis of public resources containing literature-curated human signaling interactions (**Supplementary Table 1**, **Supplementary Results 4** and **Supplementary Note 1**) and also generated a large integrated resource, OmniPath (http://omnipathdb.org/, **Supplementary Results 5**).

From the 55 relevant resources that we identified in our analysis, we selected 34 (see **Supplementary Methods** for selection criteria): 20 that provide causal interactions (12 activity flow and 8 enzyme–substrate), 8 that deliver undirected interactions from both literature curation and high-throughput screens, and 6 that capture biochemical reactions (process description). Of the causal resources, 16 provide information on the direction and 9 on the effect sign (stimulation or inhibition) of interactions (**Fig. 1a** and **Supplementary Table 2**). We focused a great deal of effort on integration in order to develop a uniform representation of the data that includes references, directionality, sign and additional details (for example, localization, compound effects, and mechanistic details) for each interaction as available. To provide guidelines for users from different fields, we grouped the resources on the basis of their features and provided notes to assist in selecting the most suitable resources (**Supplementary Table 2**).

A comparison of the resources highlighted inconsistencies between them at frequencies ranging from 2% to 7% in the direction and sign of interactions. This might be due to curation errors (**Supplementary Figs. 1**–**3**), but it may also represent important bidirectional connections and feedback loops that regulate signaling and are not yet captured by a single resource (**Supplementary Results 4**). We also found that while literature-based resources are enriched for disease-related and druggable proteins as well as for kinase–substrate interactions (**Supplementary Fig. 4**), they have low overlap with each other (**Fig. 1b**, **Supplementary Tables 3** and **4**), and individually they provide limited coverage of the human proteome (maximum of 13%).

Since coverage is essential to capture biological information, and our analysis pointed to the highly complementary nature of existing resources (**Supplementary Figs. 4a**–**g** and **5a**,**b**), we developed a high-confidence combined resource called OmniPath. From the 34 literature-based resources that were analyzed, we included interactions from 27 interaction resources, some causal and some undirected, based on specific criteria (see **Supplementary Methods**). The conversion of reactions from process description resources into binary interactions resulted in many indirect relationships, for which it was often not possible to assign references unambiguously (**Supplementary Methods**); hence, we kept these resources in separate categories.

OmniPath covers approximately three times more proteins (7,984) and four times more interactions (36,557) than the largest causal resource it contains. It covers ~39% of the human proteome, 61% of disease–gene associations, >80% of cancer-related genes and 54% of druggable proteins (as compared to 13%, 42%, 55% and 22%, respectively, in the largest casual resource; **Supplementary Fig. 4**, **Supplementary Results 2** and **Supplementary Methods**). OmniPath encompasses 41,237 references from 1,132 journals. On average, each interaction is supported by 2.88 references. OmniPath integrates additional information on the structure and mechanism of the interactions, drug targets, functional annotation, tissue-specific expression and mutations to increase its applicability (**Fig. 1a**).

We provide a free, annually updated and ready-to-use web resource (http://omnipathdb.org/), as well as a complementary open-source, feature-rich Python module called pypath (**Supplementary Software** and http://github.com/saezlab/pypath), which offers advanced possibilities for pathway analysis with unprecedented coverage and detail. pypath automatically updates its content, which is dynamically gathered from the resources. pypath is also capable of compiling networks from custom sets of interaction resources and can integrate additional annotations.

Previous comparisons of consistency, coverage, network topology and biological properties (for example, refs. 2–4) cover fewer aspects than our analysis, while other integrative efforts[5,6] include fewer literature-based resources than OmniPath and lack our detailed data structure, in particular the integrated information about directionality and references (**Supplementary Note 2**) that enables analyses such as those presented here.

Our analysis and guidelines (see **Supplementary Note 1**) are provided to help researchers find the most appropriate set of

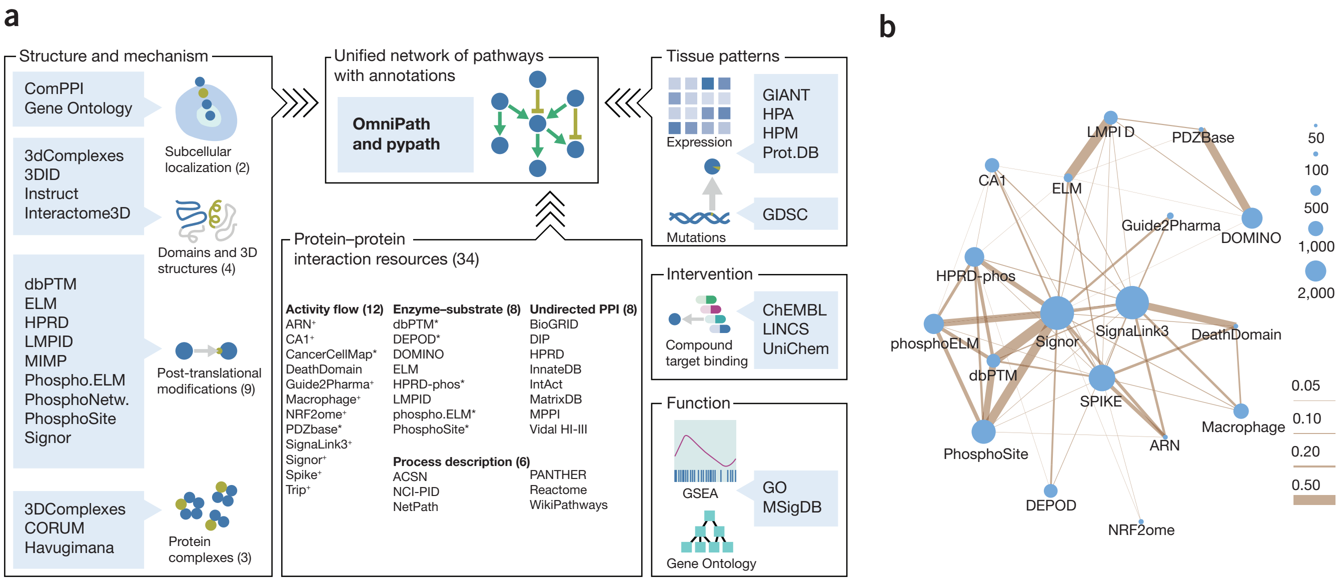

Figure 1 | Resources featured in OmniPath and pypath. (**a**) Overview of OmniPath and pypath. Causal resources (including activity-flow and enzyme–substrate resources) can provide direction (*) or sign and direction ($^+$) of interactions. Data types provided by the listed resources can be integrated with pathways in different ways (**Supplementary Results 1**, **Supplementary Methods**). GO, Gene Ontology; HPA, Human Protein Atlas; HPM, Human Proteome Map; PhosphoNetw, PhosphoNetworks; Prot.DB, ProteomicsDB. (**b**) Overlap of interactions across causal resources. Circle size denotes the number of interactions per resource, and line widths show the overlap of interactions between them, as measured by the Simpson index (see equation S3 in **Supplementary Methods**). For clarity, only links with Simpson index >0.05 are shown. MatrixDB and TRIP are not shown because they have no overlaps above this threshold.

resources for their research. While we did our best to include all relevant resources, any that were missed, as well as future resources, can be easily added to OmniPath and to the pypath tool.

Dénes Türei[1], Tamás Korcsmáros[2,3] & Julio Saez-Rodriguez[1,4]

[1]European Molecular Biology Laboratory–European Bioinformatics Institute, Hinxton, UK. [2]Earlham Institute, Norwich, UK. [3]Institute of Food Research, Norwich, UK. [4]RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine, Aachen, Germany.
e-mail: saezrodriguez@gmail.com

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.4077).

ACKNOWLEDGMENTS

We thank members of the Saez laboratory, especially L. Tobalina for useful feedback on the manuscript and for testing pypath and A. Zinovyev, G. Cesareni, H. Hermjakob and L. Perfetto for useful discussions. D.T. was supported by the EMBL Interdisciplinary Postdoc Programme (EIPOD) under Marie Skłodowska-Curie COFUND Actions (grant number 291772). T.K. was funded by a fellowship in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Institute of Food Research (Norwich, UK), and strategically supported by Biotechnological and Biosciences Research Council, UK (BB/J004529/1). We thank the developers of all the very useful resources we have used in this study and their funding agencies as described in their respective publications and webpages.

AUTHOR CONTRIBUTIONS

D.T. designed and performed the analyses, coded the pypath software and wrote the manuscript. J.S.R. conceived and supervised the project. T.K. and J.S.R. contributed to the design of the analyses, interpretation of results, and writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi: 10.1038/nmeth.4077).

1. Bader, G.D., Cary, M.P. & Sander, C. *Nucleic Acids Res.* **34**, D504–D506 (2006).
2. Cusick, M.E. *et al. Nat. Methods* **6**, 39–46 (2009).
3. Klingström, T. & Plewczynski, D. *Brief. Bioinform.* **12**, 702–713 (2011).
4. Kirouac, D.C. *et al. BMC Syst. Biol.* **6**, 29 (2012).
5. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. *Nucleic Acids Res.* **41**, D793–D800 (2013).
6. Cerami, E.G. *et al. Nucleic Acids Res.* **39**, D685–D690 (2011).