Published online 22 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D607–D613***
doi: 10.1093/nar/gky1131

STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets

Damian Szklarczyk[1], Annika L. Gable[1], David Lyon[1], Alexander Junge[2], Stefan Wyder[1], Jaime Huerta-Cepas[3], Milan Simonovic[1], Nadezhda T. Doncheva[2,4], John H. Morris[5], Peer Bork[6,7,8,9,*], Lars J. Jensen[2,*] and Christian von Mering[1,*]

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland, [2]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen N, Denmark, [3]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM)—Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), 28223 Madrid, Spain, [4]Center for non-coding RNA in Technology and Health, University of Copenhagen, 2200 Copenhagen N, Denmark, [5]Resource on Biocomputing, Visualization, and Informatics, University of California, San Francisco, CA 94158-2517, USA, [6]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [7]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [8]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany and [9]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Received September 28, 2018; Revised October 23, 2018; Editorial Decision October 24, 2018; Accepted November 16, 2018

ABSTRACT

Proteins and their functional interactions form the backbone of the cellular machinery. Their connectivity network needs to be considered for the full understanding of biological phenomena, but the available information on protein–protein associations is incomplete and exhibits varying levels of annotation granularity and reliability. The STRING database aims to collect, score and integrate all publicly available sources of protein–protein interaction information, and to complement these with computational predictions. Its goal is to achieve a comprehensive and objective global network, including direct (physical) as well as indirect (functional) interactions. The latest version of STRING (11.0) more than doubles the number of organisms it covers, to 5090. The most important new feature is an option to upload entire, genome-wide datasets as input, allowing users to visualize subsets as interaction networks and to perform gene-set enrichment analysis on the entire input. For the enrichment analysis, STRING implements well-known classification systems such as Gene Ontology and KEGG, but also offers additional, new classification systems based on high-throughput text-mining as well as on a hierarchical clustering of the association network itself. The STRING resource is available online at https://string-db.org/.

INTRODUCTION

While an impressive amount of structural and functional information on individual proteins has been amassed (1–3), our knowledge about their interactions remains more fragmented. Some interactions are quite well documented and understood, for example in the context of three-dimensional reconstructions of large cellular machineries (4–6), while others are only hinted at so far, through indirect evidence such as genetic observations or statistical predictions. Furthermore, the space of potential protein–protein interactions is much larger, and also more context-dependent, than the space of intrinsic molecular function of individual molecules. Interactions may not only be limited to certain cell types or certain physiological conditions, but their specificity and strength may vary as well, from obligatory, highly specific and stable bindings to more fleeting and relatively unspecific encounters. From a purely functional perspective, proteins can even interact specifically without touching at all, such as when a transcription factor helps to regulate the expression and production of another pro-

*To whom correspondence should be addressed. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch
Correspondence may also be addressed to Peer Bork. Tel: +49 6221 3878526; Fax: +49 6221 387517; Email: peer.bork@embl.de
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 353 25025; Fax: +45 353 25001; Email: lars.juhl.jensen@cpr.ku.dk

tein, or when two enzymes exchange a specific substrate via diffusion.

Arguably, the common denominator of the various forms of protein–protein associations is information flow—biologically meaningful interfaces have evolved to allow the flow of information through the cell, and they are ultimately essential for implementing a functional system. Hence, it is desirable to collect and integrate all types of protein–protein interactions under one framework; this then provides support for data analysis pipelines in diverse areas, ranging from disease module identification (7,8) to biomarker discovery (9–11) and allows manual browsing, *ad hoc* discovery and annotation.

Protein–protein interactions can be collected from a number of online databases (reviewed in (12,13)), as well as from individual high-throughput efforts, e.g. (14). Primary interaction databases (3,15–18) are jointly annotating experimental interaction evidence directly from the source publications, and they are coordinating their efforts through the *IMEx* consortium (19). They provide highly valuable added services such as curating metadata, maintaining common name spaces and devising ontologies and standards. A second source of protein–protein interaction information is provided by computational prediction efforts, some of which are hosted by dedicated databases, e.g. (20,21). Lastly, a third class of databases is dedicated to protein interactions at the widest scope, integrating both primary as well as predicted interactions, often including annotated pathway knowledge, text-mining results, inter-organism transfers or other accessory information. The STRING database ('Search Tool for Retrieval of Interacting Genes/Proteins') belongs to this latter class, along with GeneMania (22), FunCoup (23), I2D (24), ConsensusPathDb (25), IMP (26) and HumanNet (27)—most of which have recently been reviewed and benchmarked in (7).

STRING is one of the earliest efforts (28) and strives to differentiate itself mainly through (i) high coverage, (ii) ease of use and (iii) a consistent scoring system. It currently features the largest number of organisms (5090) and proteins (24.6 million), has very broad and diverse, benchmarked data sources and provides intuitive and fast viewers for online use. It also features a number of additional data access points, such as programmatic access through an API, access through a Cytoscape app (http://apps.cytoscape.org/apps/stringapp), as well as download pages covering individual species networks and associated data. The website allows users to log on and store their searches and gene sets, and contains evidence viewers to inspect the underlying evidence of any given interaction. It also provides users with high-level information regarding their input/search data, including network enrichment statistics and functional enrichment detection, using two different conceptual frameworks for the latter (see below). Many of the features of STRING have been made available and described earlier (28–31) and the website is currently accessed by around 3500 distinct users daily; its hosting facilities have recently been replicated and placed under a commercial load balancer, to provide added stability and capacity. Users can submit multiple proteins simultaneously and visualize large networks; the Cytoscape stringApp can even handle network sizes of several thousand proteins. STRING shares its genome-, protein- and name spaces with a number of sister projects, dedicated to orthology (eggNOG (32)), small molecules (STITCH (33)), protein abundances (PaxDB (34)), tissue expression (TISSUES (35)) and viruses (Viruses.STRING (36)), respectively.

Together with other online resources (including the IMEx consortium, which is one of STRING's largest primary data sources), the STRING database has recently been awarded the status of a European Core Data Resource by ELIXIR, a pan-European bioinformatics initiative dedicated to sustainable bioinformatics infrastructure (37). As a prerequisite and consequence of this status, all interaction data and accessory information in STRING are now freely available without restrictions, under the Creative Commons Attribution (CC BY) 4.0 license.

DATABASE CONTENT

The basic interaction unit in STRING is the 'functional association', i.e. a link between two proteins that both contribute jointly to a specific biological function (38–40). For two proteins to be associated this way, they do not need to interact physically. Instead, it is sufficient if at least some part of their functional roles in the cell overlap—and this overlapping function should be specific enough to broadly qualify as a pathway or functional map (in contrast, merely sharing 'metabolism' as an overlapping function would be too unspecific). By this definition, even proteins that antagonize each other can be functionally associated, such as an inhibitor and an activator within the same pathway. The desired specificity cutoff for functional associations in STRING roughly corresponds to the annotation granularity of KEGG pathway maps (41), whereby maps that largely group proteins by homology (such as 'ABC transporters') are removed from consideration.

All of the association evidence in the STRING database is categorized into one of seven independent 'channels': three prediction channels based on genomic context information (see below), and one channel each for (i) co-expression, (ii) text-mining, (iii) biochemical/genetic data ('experiments') and (iv) previously curated pathway and protein-complex knowledge ('databases'). Users can disable all channels individually or in combinations. For each channel, separate interaction scores are available as well as viewers for inspecting the underlying evidence (Figure 1). In general, the interaction scores in STRING do not represent the strength or specificity of a given interaction, but instead are meant to express an approximate confidence, on a scale of zero to one, of the association being true, given all the available evidence. The scores in STRING are benchmarked using the subset of associations for which both protein partners are already functionally annotated; for this, the KEGG pathway maps (41) are used as a gold standard and they thus implicitly also determine the granularity of the functional associations.

Within each channel, the evidence is further subdivided into two sub-scores, one of which represents evidence stemming from the organism itself, and the other represents evidence transferred from other organisms. For the latter transfer, the 'interolog' concept is applied (42,43); STRING uses hierarchically arranged orthologous group relations as

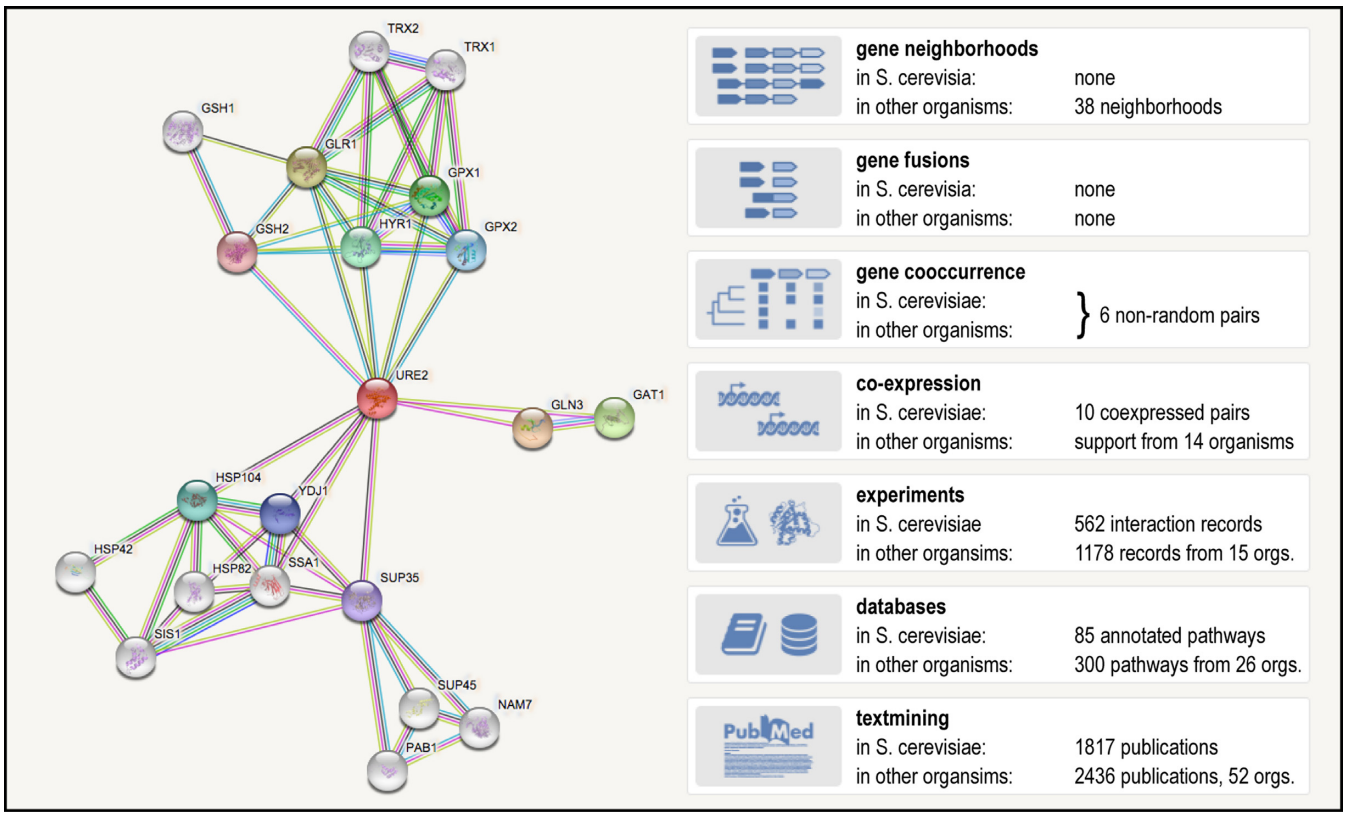

Figure 1. A typical association network in STRING. The yeast prion-like protein URE2 has been selected as input. The network has been expanded by an additional 10 proteins (via the 'More' button in the STRING interface), and the confidence cutoff for showing interaction links has been set to 'highest' (0.900). The insets at the right show how many items of the various evidence types in STRING contributed to this particular network (counts denote how many records covered at least two of the proteins in the network; not all of these records contributed high-scoring links after score calibration).

defined in eggNOG (32), in order to transfer associations between organisms where applicable (described in (29)).

The individual protein associations in the various channels are derived, briefly, as follows:

The three genomic context prediction channels (neighborhood, fusion, gene co-occurrence) are the result of systematic all-against-all genome comparisons, aiming to assess the consequences of past genome rearrangements, gene gains and losses, as well as gene fusion events. These evolutionary events are known to be retained non-randomly with respect to the functional roles of genes, and thus allow the inference of functional associations between genes even for otherwise rarely studied organisms (genomic context techniques are reviewed in (44,45)).

The co-expression channel is based on gene-by-gene correlation tests across a large number of gene expression datasets (using both transcriptome measurements as well as proteome measurements). In the case of transcript data, STRING re-processes and maps the large number of experiments stored in the NCBI Gene Expression Omnibus (46), followed by normalization, redundancy reduction and Pearson correlation (described in (29)). For version 11, we have further improved the RNAseq co-expression inference pipeline. This was achieved by processing a higher number of RNAseq samples and using the robust biweight midcorrelation (47). In addition to NCBI Geo, for a subset of species, gene count data was downloaded from the ARCHS4 and ARCHS4 zoo collections (48).

Protein-based co-expression analysis is new in version 11 of STRING, and as of now it is restricted to one dataset imported *as is*: namely the ProteomeHD dataset of the Juri Rappsilber lab (unpublished, https://www.proteomehd.net/), covering 294 biological conditions measured using SILAC in human cells. ProteomeHD is not based on Pearson correlation, but instead uses the treeClust algorithm (49); for STRING, the results of this algorithm are recalibrated and scored using the KEGG benchmark. Each ProteomeHD-provided interaction features a cross-link through which the underlying evidence can be inspected at the ProteomeHD website.

For the experiments channel, all interaction records from the IMEx databases (plus BioGRID), are re-mapped and re-processed: first, duplicate records and datasets are removed, and then entire groups of records are benchmarked against KEGG and scored accordingly.

The database channel is based on manually curated interaction records assembled by expert curators, at KEGG (41), Reactome (50), BioCyc (51) and Gene Ontology (52), as well as legacy datasets from PID and BioCarta. STRING only retains associations between direct pathway members or within protein complexes. The database channel is the only channel for which score calibration does not apply; in-

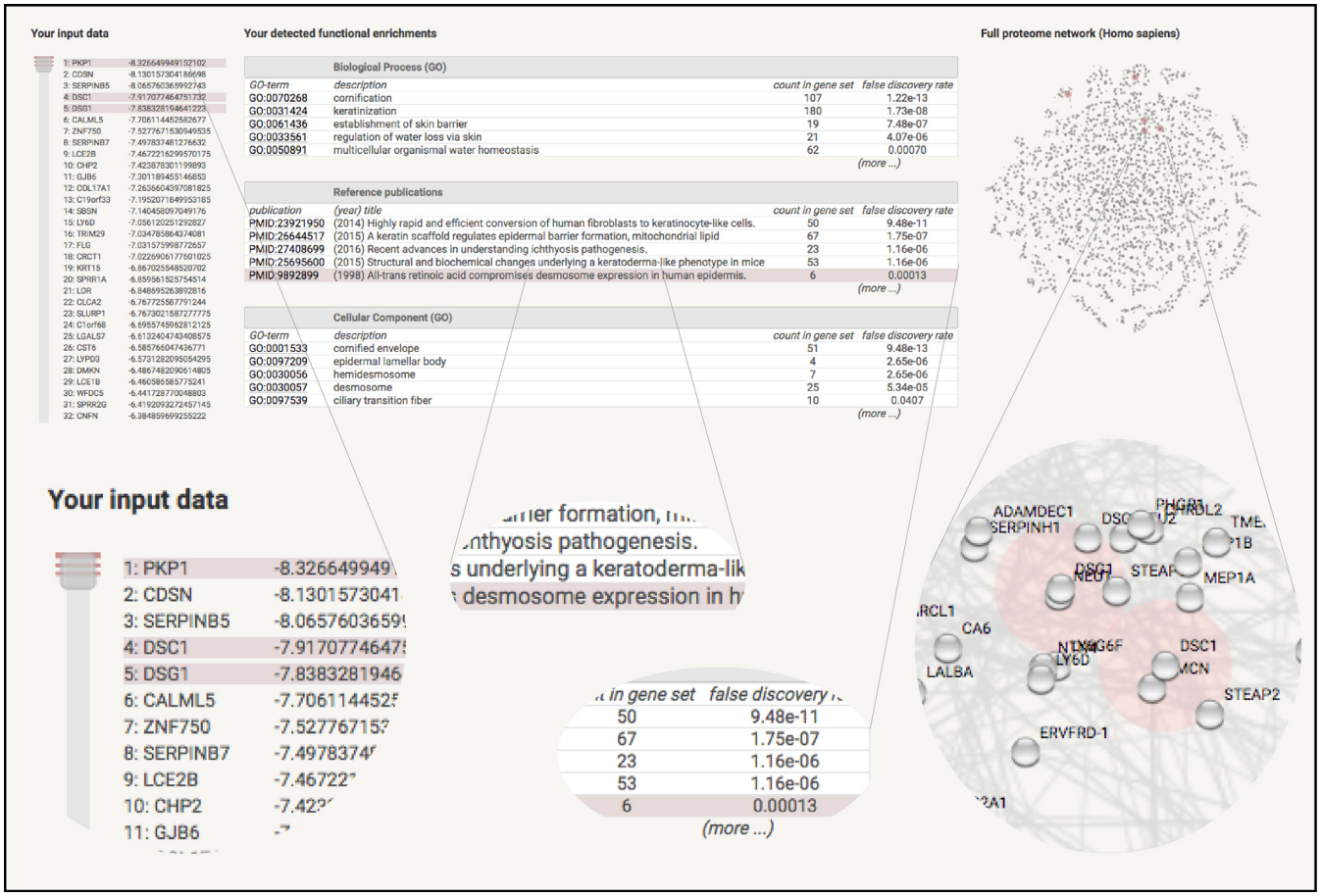

Figure 2. Functional enrichment analysis of a genome-sized input set. An expression dataset comparing metastatic melanoma cells with normal skin tissue (62) has been submitted to STRING, with average log fold change values associated to each gene (negative values signify depletion in the melanoma cells). The screenshot shows how STRING presents and groups statistical enrichment observations for a number of pathways and functional subsystems. When hovering with the mouse, the website highlights the corresponding proteins both in the input data on the left side, as well as in the organism-wide network on the right side. The latter can be interactively zoomed until individual proteins and their neighbors become discernible. Here, the highlighted observation shows that the desmosome is downregulated in melanoma cells—this stands out by way of several publications in PubMed whose discussed proteins (desmosome proteins) are strongly enriched at one end of the user input.

stead, all associations in this channel receive a high, uniform score (0.900).

At last, for the text-mining channel, STRING conducts statistical co-citation analysis across a large number of scientific texts, including all PubMed abstracts as well as OMIM (53). Since version 10.5 of STRING, the text corpus also contains a subset of full-text articles. For version 11.0, the Medline abstracts (last updated on 9 June 2018) were complemented with open access as well as author-manuscript full text articles available from PMC in BioC XML format (https://arxiv.org/abs/1804.05957) (last updated on 17 April 2018). Full-text articles that were not classified as English-language articles were removed (using fastText and a pretrained language identification model for 176 languages (https://arxiv.org/abs/1607.01759)), as were those that could not be mapped to PubMed. We also removed highly unspecific articles that mention more than 200 relevant biomedical entities such as proteins, chemicals, diseases or tissues. The final corpus consists of 28 579 637 scientific publications, of which 2 106 542 are available as full-text articles and the remainder as abstracts. While the text-mining pipeline itself has remained unchanged (last described in (29)), its dictionary of gene and protein names has been updated to the new set of genomes and the stopword list improved to increase precision, especially for human proteins.

NEW ENRICHMENT DETECTION MODE

For users that query the STRING database with a set of proteins (as opposed to a single query protein only), the website computes a functional enrichment analysis in the background; this can then be inspected and browsed by the user, and includes interactive projections of the results onto the user's protein network. This functionality has been available since version 9.1, and is based on straightforward over-representation analysis using hypergeometric tests.

However, this analysis uses only a small part of the information that the user might have about his or her protein list. First, the original list of proteins might have been much longer, and the user would have had to truncate it (thus far, STRING enforced an upper limit on the number of query items). Second, the list might have had a biologically mean-

ingful ranking, which would have been lost during submission to STRING. Third, each protein might have been associated with some numerical information from the underlying experiment or study (such as a log fold change, a measured abundance, a phenotypic outcome, etc.). For this type of genome-wide measurements, simple overlap-based overrepresentation analysis is not the best choice (54–56).

Thus, beginning with version 11.0, STRING offers such users a second option for conducting enrichment analysis. It specifically asks for genome-scale input, with each protein or gene having an associated numerical value (a measurement or statistical metric). Of the available methods for searching functional enrichments in such a set, we chose a permutation-based, non-parametric test that performs well in a number of settings, termed 'Aggregate Fold Change' (56). Briefly, this test works by computing, for each gene set to be tested, the average of all values provided by the user for the constituent genes. This average is then compared against averages of randomized gene sets of the same size. Multiple testing correction is applied separately within each functional classification framework (GO, KEGG, InterPro, etc.), according to Benjamini and Hochberg (57), but not across these frameworks as there is significant overlap between them. For large gene sets, the AFC randomization method becomes prohibitively slow; these gene sets are instead tested after converting the user-provided gene values to ranks, using two-sided Kolmogorov–Smirnov testing. In addition to the usually applied functional classification frameworks, STRING uses two additional systems, thus giving users more options and potentially more novelty for discovery. The first is based on a hierarchical clustering of the STRING network itself. This assumes that tightly connected modules within the network broadly correspond to functional units, and has the advantage that it covers a broader scope and potentially also novel modules that may not yet be annotated as pathways. The clustering is based on a confidence diffusion state distance matrix (58,59) computed on the full, organism-wide STRING network, which is clustered hierarchically using HPC-CLUST with average linkage (60). To compute the DSD matrix, the final, combined STRING-score between proteins is used, and the DSD algorithm is run with default parameters and the '-c' flag (confidence). Following the clustering procedure, all clusters with sizes between 5 and 200 are included in the functional enrichment testing, and reported under their own, separate classification category. The second additional set for enrichment testing consists of all published papers mapping to the genes in the user's input. This takes advantage of STRING's text-mining channel, for which all of PubMed's abstract and some additional scientific text are already mapped onto STRING's protein space (based on identifier matches in the text). Detecting publications that are enriched in the user-input ranking provides yet another complementary way of interpreting the input, often with a more fine-grained view.

Following the computation of the entire new enrichment option, users are presented with a three-panel view of the results (Figure 2). There, each enriched functional subset can be highlighted, and tracked back to the user's input as well as to a pre-rendered, organism-wide STRING network. The layout of the latter is based on a t-SNE-visualization of the network (61) and can be zoomed and panned interactively.

OUTLOOK

Over the coming years, the STRING team aims to continue tracking all available protein association evidence types and prediction algorithms. One particular focus will be to expand the protein-based co-expression channel, where advances in proteomics throughput and scope lead us to expect growing data support for association searches. With regard to the STRING website, we expect to provide tighter integration of functional enrichment and network search results, and are exploring options to provide more context on the various networks (such as cell type, tissues, organelles). We will also strive to provide better interoperability options and increase our list of partnered, crosslinked resources as well as applicable direct data import options to facilitate our regular data updates.

ACKNOWLEDGEMENTS

We are indebted to Juri Rappsilber and his team for sharing ProteomeHD data prior to publication, and to Yan P. Yuan for excellent IT support at EMBL. Thomas Rattei and his SIMAP project at University of Vienna provided essential protein similarity data for our very large sequence space. We thank Tudor Oprea and the Illuminating the Druggable Genome project for help in improving the text mining, and Daniel Mende and Sofia Forslund for their help in selecting a non-redundant set of high-quality genomes.

FUNDING

The Swiss Institute of Bioinformatics (Lausanne) provides long-term core funding for STRING, as do the Novo Nordisk Foundation (Copenhagen, NNF14CC0001) and the European Molecular Biology Laboratory (EMBL Heidelberg). N.D.T. received funding from the Danish Council for Independent Research (DFF-4005-00443), and A.J. from the National Institutes of Health (NIH) Illuminating the Druggable Genome Knowledge Management Center (U54 CA189205 and U24 224370). J.H.M. was funded by the NIH (NIGMS P41 GM103504), by grant number 2018-183120 from the Chan Zuckerberg Initiative DAF, and by the advised fund of the Silicon Valley Community Foundation. Incorporation into the German bioinformatics infrastructure has been enabled by the BMBF (de.nbi grant #031A537B). Funding for Open Access charges: University of Zurich.

Conflict of interest statement. None declared.

REFERENCES

1. Xie,L. and Bourne,P.E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol.*, **1**, e31.
2. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a Knowledge-Based human protein atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
3. UniProt Consortium, T (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

4. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, **289**, 905–920.
5. Schuller,J.M., Falk,S., Fromm,L., Hurt,E. and Conti,E. (2018) Structure of the nuclear exosome captured on a maturing preribosome. *Science*, **360**, 219–222.
6. Marsh,J.A. and Teichmann,S.A. (2015) Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.*, **84**, 551–575.
7. Huang,J.K., Carlin,D.E., Yu,M.K., Zhang,W., Kreisberg,J.F., Tamayo,P. and Ideker,T. (2018) Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst.*, **6**, 484–495.
8. Khurana,V., Peng,J., Chung,C.Y., Auluck,P.K., Fanning,S., Tardiff,D.F., Bartels,T., Koeva,M., Eichhorn,S.W., Benyamini,H. *et al.* (2017) Genome-scale networks link neurodegenerative disease genes to alpha-Synuclein through specific molecular pathways. *Cell Syst.*, **4**, 157–170.
9. Hayashida,M. and Akutsu,T. (2016) Complex network-based approaches to biomarker discovery. *Biomark. Med.*, **10**, 621–632.
10. 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.
11. Liu,X., Chang,X., Liu,R., Yu,X., Chen,L. and Aihara,K. (2017) Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PLoS Comput. Biol.*, **13**, e1005633.
12. Gemovic,B., Sumonja,N., Davidovic,R., Perovic,V. and Veljkovic,N. (2018) Mapping of Protein-Protein interactions: Web-Based resources for revealing interactomes. *Curr. Med. Chem.*, doi:10.2174/0929867325666180214113704.
13. Sowmya,G. and Ranganathan,S. (2014) Protein-protein interactions and prediction: a comprehensive overview. *Protein Pept. Lett.*, **21**, 779–789.
14. Drew,K., Lee,C., Huizar,R.L., Tu,F., Borgeson,B., McWhite,C.D., Ma,Y., Wallingford,J.B. and Marcotte,E.M. (2017) Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.*, **13**, 932.
15. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
16. 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.
17. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
18. Ammari,M.G., Gresham,C.R., McCarthy,F.M. and Nanduri,B. (2016) HPIDB 2.0: a curated database for host-pathogen interactions. *Database (Oxford)*, **2016**, baw103.
19. 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.
20. Zhang,Q.C., Petrey,D., Garzon,J.I., Deng,L. and Honig,B. (2013) PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.
21. McDowall,M.D., Scott,M.S. and Barton,G.J. (2009) PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res.*, **37**, D651–D656.
22. Franz,M., Rodriguez,H., Lopes,C., Zuberi,K., Montojo,J., Bader,G.D. and Morris,Q. (2018) GeneMANIA update 2018. *Nucleic Acids Res.*, **46**, W60–W64.
23. Ogris,C., Guala,D. and Sonnhammer,E.L.L. (2018) FunCoup 4: new species, data, and visualization. *Nucleic Acids Res.*, **46**, D601–D607.
24. Kotlyar,M., Pastrello,C., Sheahan,N. and Jurisica,I. (2016) Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.*, **44**, D536–D541.
25. Herwig,R., Hardt,C., Lienhard,M. and Kamburov,A. (2016) Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protoc.*, **11**, 1889–1907.
26. Wong,A.K., Krishnan,A., Yao,V., Tadych,A. and Troyanskaya,O.G. (2015) IMP 2.0: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.*, **43**, W128–D133.
27. Lee,I., Blom,U.M., Wang,P.I., Shim,J.E. and Marcotte,E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**, 1109–1121.
28. Snel,B., Lehmann,G., Bork,P. and Huynen,M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.*, **28**, 3442–3444.
29. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
30. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.
31. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
32. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M. *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*, **44**, D286–D293.
33. Szklarczyk,D., Santos,A., von Mering,C., Jensen,L.J., Bork,P. and Kuhn,M. (2016) STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.*, **44**, D380–D384.
34. Wang,M., Herrmann,C.J., Simonovic,M., Szklarczyk,D. and von Mering,C. (2015) Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**, 3163–3168.
35. Palasca,O., Santos,A., Stolte,C., Gorodkin,J. and Jensen,L.J. (2018) TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database (Oxford)*, **2018**, bay003.
36. Cook,V.H., Doncheva,N.T., Szklarczyk,D., von Mering,C. and Jensen,L.J. (2018) Viruses.STRING: A virus–host protein–protein interaction database. *Viruses*, **10**, 519.
37. Durinx,C., McEntyre,J., Appel,R., Apweiler,R., Barlow,M., Blomberg,N., Cook,C., Gasteiger,E., Kim,J.H., Lopez,R. *et al.* (2016) Identifying ELIXIR core data resources [version 2; referees: 2 approved]. *F1000Res*, **5**, 2422.
38. Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.*, **2**, RESEARCH0034.
39. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *PNAS*, **99**, 5890–5895.
40. Studham,M.E., Tjarnberg,A., Nordling,T.E., Nelander,S. and Sonnhammer,E.L. (2014) Functional association networks as priors for gene regulatory network inference. *Bioinformatics*, **30**, i130–i138.
41. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
42. Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. and Vidal,M. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, **287**, 116–122.
43. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
44. Huynen,M., Snel,B., Lathe,W. 3rd and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
45. Skrabanek,L., Saini,H.K., Bader,G.D. and Enright,A.J. (2008) Computational prediction of protein-protein interactions. *Mol. Biotechnol.*, **38**, 1–17.
46. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M.,

Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
47. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
48. Lachmann,A., Torre,D., Keenan,A.B., Jagodnik,K.M., Lee,H.J., Wang,L., Silverstein,M.C. and Ma'ayan,A. (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.*, **9**, 1366.
49. Buttrey,S.L. and Whitaker,L.R. (2015) treeClust: an R package for Tree-Based clustering dissimilarities. *TheR Journal*, **7**, 227–236.
50. Fabregat,A., Sidiropoulos,K., Garapati,P., Gillespie,M., Hausmann,K., Haw,R., Jassal,B., Jupe,S., Korninger,F., McKay,S. *et al.* (2016) The reactome pathway Knowledgebase. *Nucleic Acids Res.*, **44**, D481–D487.
51. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.
52. The Gene Ontology, C. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
53. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
54. Garcia-Campos,M.A., Espinal-Enriquez,J. and Hernandez-Lemus,E. (2015) Pathway analysis: state of the art. *Front. Physiol.*, **6**, 383.
55. Tarca,A.L., Bhatti,G. and Romero,R. (2013) A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One*, **8**, e79217.
56. Yu,C., Woo,H.J., Yu,X., Oyama,T., Wallqvist,A. and Reifman,J. (2017) A strategy for evaluating pathway analysis methods. *BMC Bioinformatics*, **18**, 453.
57. Benyamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
58. Cao,M., Zhang,H., Park,J., Daniels,N.M., Crovella,M.E., Cowen,L.J. and Hescott,B. (2013) Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PLoS One*, **8**, e76339.
59. Cao,M., Pietras,C.M., Feng,X., Doroschak,K.J., Schaffner,T., Park,J., Zhang,H., Cowen,L.J. and Hescott,B.J. (2014) New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, **30**, i219–i227.
60. Matias Rodrigues,J.F. and von Mering,C. (2014) HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics*, **30**, 287–288.
61. van der Maaten,L.J.P. and Hinton,G.E. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
62. Riker,A.I., Enkemann,S.A., Fodstad,O., Liu,S., Ren,S., Morris,C., Xi,Y., Howell,P., Metge,B., Samant,R.S. *et al.* (2008) The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med. Genomics*, **1**, 13.