## Title

The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible



## Journal

Nucleic Acids Research, 45(D1)

## ISSN

0305-1048

## Authors

Szklarczyk, Damian
Morris, John H
Cook, Helen
et al.

## Publication Date

2017-01-04

## DOI

10.1093/nar/gkw937

Peer reviewed

*D362–D368* *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 18 October 2016*
*doi: 10.1093/nar/gkw937*

# The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible

**Damian Szklarczyk[1], John H Morris[2], Helen Cook[3], Michael Kuhn[4], Stefan Wyder[1], Milan Simonovic[1], Alberto Santos[3], Nadezhda T Doncheva[3], Alexander Roth[1], Peer Bork[4,5,6,7,*], Lars J. Jensen[3,*] and Christian von Mering[1,*]**

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland, [2]Resource on Biocomputing, Visualization, and Informatics, University of California, San Francisco, CA 94158-2517, USA, [3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen N, Denmark, [4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [5]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [6]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany and [7]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Received September 15, 2016; Editorial Decision October 04, 2016; Accepted October 06, 2016

## ABSTRACT

**A system-wide understanding of cellular function requires knowledge of all functional interactions between the expressed proteins. The STRING database aims to collect and integrate this information, by consolidating known and predicted protein–protein association data for a large number of organisms. The associations in STRING include direct (physical) interactions, as well as indirect (functional) interactions, as long as both are specific and biologically meaningful. Apart from collecting and reassessing available experimental data on protein–protein interactions, and importing known pathways and protein complexes from curated databases, interaction predictions are derived from the following sources: (i) systematic co-expression analysis, (ii) detection of shared selective signals across genomes, (iii) automated text-mining of the scientific literature and (iv) computational transfer of interaction knowledge between organisms based on gene orthology. In the latest version 10.5 of STRING, the biggest changes are concerned with data dissemination: the web frontend has been completely redesigned to reduce dependency on outdated browser technologies, and the database can now also be queried from inside the popular Cytoscape software framework. Further improvements include automated background analysis of user inputs for functional enrichments, and streamlined download options. The STRING resource is available online, at http://string-db.org/.**

## INTRODUCTION

The flow of information and energy through the cell proceeds along specific and evolved interfaces: across and between nucleotides, proteins, lipids, metabolites and other small molecules. Among these interfaces, those between proteins are arguably among the most important, being biochemically diverse and information-rich, and showing exquisite specificity (1–3). Apart from direct physical binding, proteins also have many other, indirect ways of cooperation and mutual regulation: they can influence each other's production and half-life transcriptionally and post-transcriptionally, exchange reaction products, participate in signal relay mechanisms, or jointly contribute toward specific organismal functions. Together, these direct and indirect interactions constitute 'functional association', a useful operational umbrella-term for specific and functionally productive interactions of any type (4–9).

Assembling all known and predicted protein functional associations for a given organism results in a protein network of genome-wide functional connectivity. These networks represent a crucial, intermediate level of information aggregation: they are placed between pathway databases at one extreme (which provide mechanistic detail but often have low coverage), and high-throughput experimental interaction discovery and *ad hoc* predictions at the other ex-

*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: 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

treme (which have high coverage but usually also high levels of false positives). As such, protein networks are ideally suited to serve as scaffolds or filters for further data integration, for visualization and for molecular discovery. They are essential for modern life sciences: protein networks are used to increase discovery power for noisy data sets by 'network smoothing' (10,11), help define drug efficiency by network-based 'drug-disease proximity measures' (12), help to interpret the results of genome-wide association screens (13–17) and enable the discovery of new molecular players through the 'guilt by association' concept (18,19).

A number of databases and online resources are dedicated to protein networks, at various levels of abstraction and each with a somewhat different focus/scope. First, individual well-supported protein–protein interactions are curated manually from the published literature, through dedicated efforts by members of the IMEx consortium (20,21), but also as part of more general annotation workflows such as within the UniProt consortium (22). Second, a number of databases assemble larger, genome-wide protein networks that are nevertheless still restricted to experimentally observed interactions only; examples include BioGRID (23), HINT (24), iRefWeb (25) and APID (26). Lastly, resources such as STRING include indirect and predicted interactions on top, aiming for inclusiveness in scope and for maximal coverage. Apart from STRING, this latter group includes GeneMANIA (27), Integrated Multi-species Prediction (28), Integrated Interactions Database (29), HumanNet (17), FunCoup (30) and others. For this group of data resources, it is particularly important to provide interaction weights (such as quality scores or confidence estimates), to allow the users to prune down these inclusive networks, as needed.

Within the spectrum of the above resources, STRING aims to set itself apart in three ways: (i) comprehensiveness – it covers the largest number of organisms and uses the widest breadth of input sources, including automated text-mining and computational predictions, (ii) usability – in terms of an intuitive web interface, Cytoscape integration and programmatic access options, and (iii) quality control and traceability – each interaction is annotated with benchmarked confidence scores, separately per evidence type, and the underlying evidence can be tracked to its source. STRING has been maintained continuously since the year 2000, and has already been described in several publications (31–34). Below, we provide a brief overview of the main features, and describe recent technical developments.

## DATABASE CONTENT

For each protein–protein association stored in STRING, a score is provided. These scores (i.e., the 'edge weights' in each network) represent confidence scores, and are scaled between zero and one. They indicate the estimated likelihood that a given interaction is biologically meaningful, specific and reproducible, given the supporting evidence. For each interaction, the supporting evidence is divided into one or more 'evidence channels', depending on the origin and type of the evidence. There are seven channels, and they are assembled, scored and benchmarked separately. In the network visualization on the web frontend, the evidence channels are usually delineated by edges of different color, and each of the channels can be disabled individually by the user, in case some types of evidence might not be considered suitable for a particular question that is being studied. Based on the seven channels, a combined and final confidence score is computed for each interaction, and it is this 'combined score' that is typically used as the final measure when building networks or when sorting and filtering interactions. For a given interaction, it is generally a good sign of support when not only the combined score is high, but when there is also more than one evidence channel contributing to the score. Furthermore, it is important to note that the interactions in STRING have gene-locus resolution only: we do not discriminate between different splice isoforms or post-translationally modified forms. Hence, the interacting units in STRING are actually the protein-coding gene loci (represented by their main, canonical protein isoform).

Briefly, the seven evidence channels in STRING are (i) The *experiments* channel: Here, evidence comes from actual experiments in the lab (including biochemical, biophysical, as well as genetic experiments). This channel is populated mainly from the primary interaction databases organized in the IMEx consortium, plus BioGRID. (ii) The *database* channel: In this channel, STRING collects evidence that has been asserted by a human expert curator; this information is imported from pathway databases. (iii) The *textmining* channel: Here, STRING searches for mentions of protein names in all PubMed abstracts, in an in-house collection of more than three million fulltext articles, and in other text collections (35,36). Pairs of proteins are given an association score when they are frequently mentioned together in the same paper, abstract or even sentence (relative to how often they are mentioned separately). This score is raised further when it has been possible to parse one or more sentences through Natural Language Processing, and a concept connecting the two proteins was encountered (such as 'binding' or 'phosphorylation by'). (iv) The *coexpression* channel: For this channel, gene expression data originating from a variety of expression experiments are normalized, pruned and then correlated (34). Pairs of proteins that are consistently similar in their expression patterns, under a variety of conditions, will receive a high association score. In addition to large-scale microarray data, in version 10.5 of STRING, RNAseq expression data are now also processed; this results in the inclusion of 16 previously non-covered organisms into this channel. (v) The *neighborhood* channel: This channel, and the next two, are genome-based prediction channels, whose functionality is generally most relevant for Bacteria and Archaea. In the neighborhood channel, genes are given an association score where they are consistently observed in each other's genome neighborhood (such as in the case of conserved, co-transcribed 'operons'). (vi) The *fusion* channel: Pairs of proteins are given an association score when there is at least one organism where their respective orthologs have fused into a single, protein-coding gene. Finally, (vii) The co-occurrence channel: In this channel, STRING evaluates the phylogenetic distribution of orthologs of all proteins in a given organism. If two proteins show a high similarity in this distribution, i.e. if their orthologs tend to be observed as 'present' or 'absent' in the same subsets of organisms, then an association score is as-

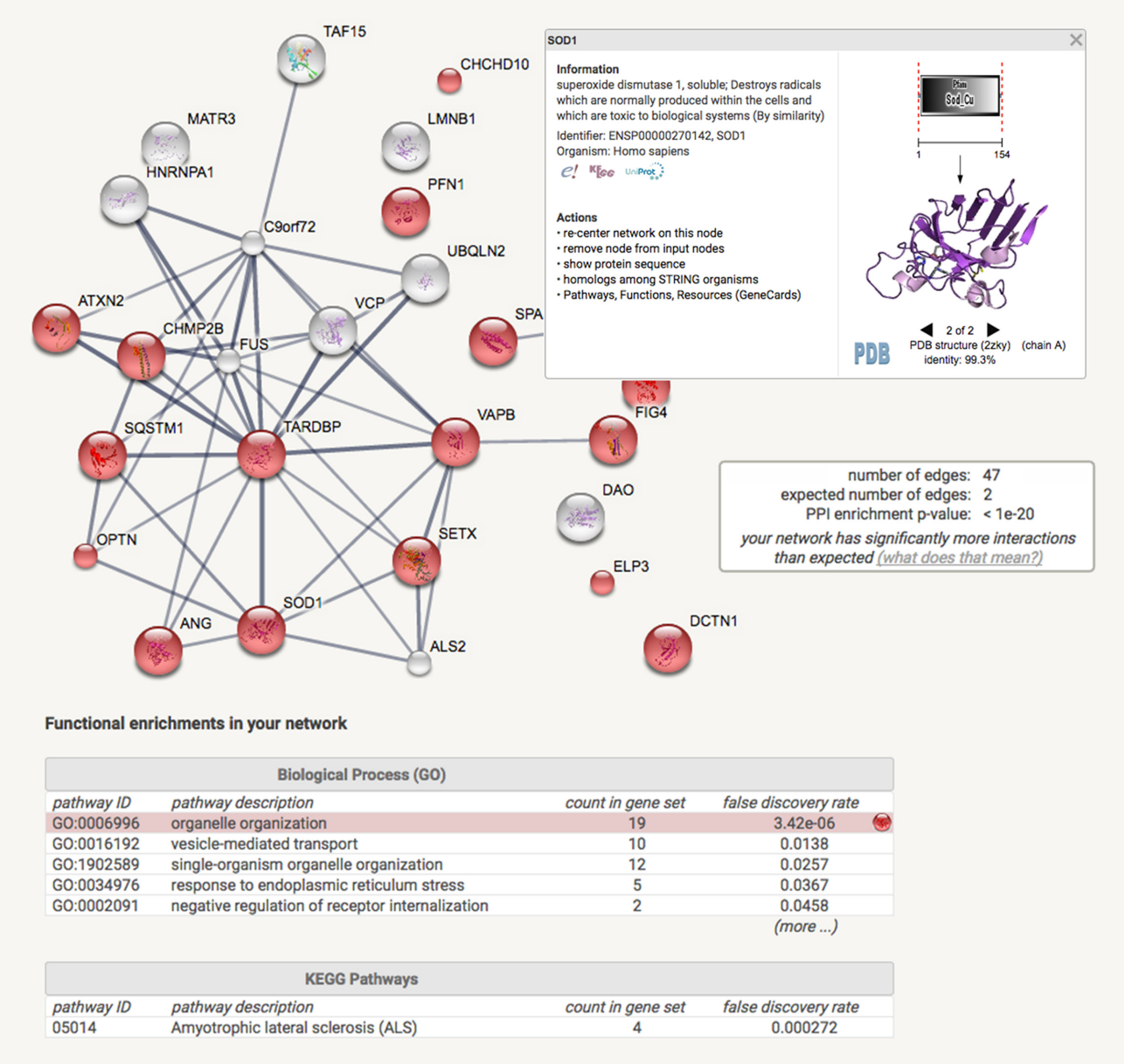


**Figure 1.** Network and Enrichment Analysis. Combined screenshots from the STRING website, showing results obtained upon entering a set of 31 proteins suspected to be involved in Amyotrophic Lateral Sclerosis (55). The insets are showing (from top to bottom): the accessory information available for a single protein, a reported enrichment of functional connections among the set of proteins, and statistical enrichments detected in functional subsystems. In the bottom inset, one enriched function has been selected, and the corresponding protein nodes in the network are automatically highlighted in color.

signed. For this channel, the details of the STRING implementation have recently been described, separately (37).

Apart from direct evidence collected in the seven evidence channels, another important contribution of interactions in STRING comes from the transfer of evidence from one organism to another. This so-called 'interolog' transfer (38,39) is based on the observation that orthologs of interacting proteins in one organism are often also interacting in another organism – this inference is the more confident the better the orthology relationships can be established. STRING relies on hierarchical orthology relations imported from the eggNOG database (40), and conducts an all-against-all transfer of interactions, benchmarked separately for each evidence channel. Transfers between closely related organisms are made more confidently, whereas the existence of paralogs (i.e., implied gene duplications) will lower the transfer score. Overall, the biggest benefit of the transfers can be seen for poorly studied organisms, where the fraction of interactions supported by transfers only can be as high as 99%. In contrast, in well-studied model organisms such as *Escherichia coli*, the corresponding fraction is below 20%.

## USER INTERFACE

The protein networks stored in STRING can be accessed in a number of ways. Programmatic access is provided via a REST-API (41), via an R/Bioconductor package (34) and via a mechanism to add additional user-provided interactions, as well as protein-centric information, onto the website ('data payload') (32). Studies that require genome-wide networks can refer to the STRING download pages, where the complete interaction scores, as well as accessory information, are available (the downloads are free for academics; commercial users need a license for some of the files). As of version 10.5, the downloads can now be pruned down, prior to receiving the files, by organism (or by groups of organisms), which greatly facilitates subsequent data processing. The most important interface to STRING, however, remains the web frontend (Figure 1). In 2016, it has been completely redesigned from the ground up; this was done in order to remove dependencies on deprecated web technologies such as Adobe Flash. The new website allows easier and more intuitive browsing of the networks and the underlying evidence, and it is tightly integrated with the database backend to provide speedy responses. Users can make search results and gene sets persistent by logging in, and stable URLs are provided on each page to facilitate sharing of results.

Importantly, users are now—by default—provided with statistical analysis results for each network. The analysis is done server-side, in the background, so as not to slow down the user experience, and it produces alerts when a network is enriched in certain known functions, or has more interactions (edges) than expected. This is particularly meaningful when users arrive to the website with a set of proteins instead of just a single query protein, as it provides a functional characterization of the set (this feature is increasingly used by STRING users). The enrichment tests are done for a variety of classification systems (Gene Ontology, KEGG, Pfam and InterPro), and employ a Fisher's exact test followed by a correction for multiple testing (42,43).

## CYTOSCAPE APP INTEGRATION

The web interface of STRING is designed primarily for users interested in small- to medium-scale networks, whereas the API, R package and download files are mainly intended for bioinformaticians who want to integrate STRING with other resources or perform large-scale network analyses. To bridge the gap between the two, we have developed a so-called App for the Cytoscape software framework (44,45), which allows users to easily retrieve, visualize and analyze networks of hundreds to thousands of proteins via a GUI.

The App allows users to query STRING in three different ways from within Cytoscape: by protein names, by disease or by PubMed query. The first of these mirrors the 'Multiple proteins' query in the STRING web interface and allows users to retrieve a network for a list of up to 2000 protein names or identifiers from, for example, a proteomics or transcriptomics study. The second option is to retrieve a network for a disease of interest; it first retrieves a list of the top-N human proteins associated with the disease from the DISEASES database (46) and subsequently loads the STRING network for these proteins into Cytoscape. The third option, PubMed query, allows users to retrieve a STRING network pertaining to any topic of interest based on text mining of PubMed abstracts. The app fetches the abstracts for a user-specified query via NCBI E-utilities, counts how many of these mention each protein from the organism of interest, ranks the proteins by comparing these counts to precomputed background counts over entire PubMed and retrieves a STRING network for the top-N proteins. The underlying text mining is performed by the software also used for the text-mining channel in STRING.

When a network is retrieved by the App, it comes associated with a large number of node attributes for each protein and edge attributes for each interaction, which can subsequently be used within Cytoscape. These include STRING and UniProt accessions to facilitate cross-linking with other resources, a human-readable name for display purposes and the protein sequence. If a protein was retrieved through a protein name query, we store also the exact query term with which the protein was found. This is helpful when querying for proteins identified in a proteomics or transcriptomics study, since it facilitates subsequent import of tabular data from the study (Figure 2). If available for the organism in question, the App also fetches information on the subcellular localization and tissue expression of each protein from the COMPARTMENTS (47) and TISSUES (48) databases as well as drug target information from Pharos (http://pharos.nih.gov/). For each interaction, the edge attributes include the overall confidence score and the subscores from each individual evidence channel.

Cytoscape and its hundreds of apps provide numerous ways for users to interact with, visualize and analyze STRING networks (49), including integrating additional data from public repositories or their own experiments, changing visual styles and applying algorithms for network layout, clustering (50), enrichment analysis (51,52) and network analysis (53). In addition to these, the STRING App allows users to modify an already retrieved network in three different ways. First, the confidence cutoff for the imported evidence channels can be increased or decreased, which in the latter case involves fetching additional interactions from STRING. Second, users can expand the network by a user-specified number of interactors that are most closely associated with all network nodes or a selected subset of them. Third, any number of additional nodes can be queried by name and added to the existing network. Furthermore, the App provides a results panel with links to related databases such as UniProt (22), GeneCards (54), Pharos, COMPARTMENTS, TISSUES and DISEASES.

## OUTLOOK

The availability of completely sequenced genomes, and of protein–protein interaction data, continues to grow quickly. Hence, the data importing and processing for STRING will be further streamlined in order to accommodate this. The upcoming version 11 of STRING will cover more than 4000 organisms, and will contain pre-computed protein networks for all of them. We are also developing a separate and distinctive interface specifically for the investigation of virus-

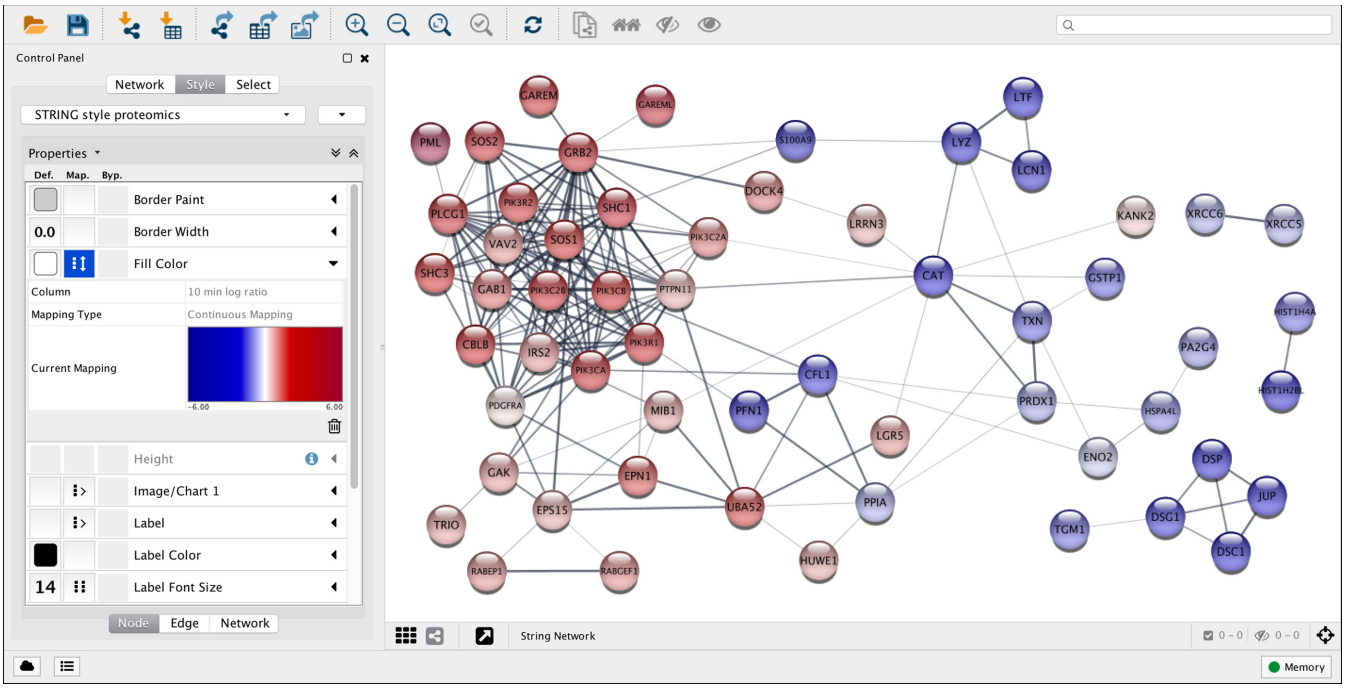


**Figure 2.** STRING network visualization within Cytoscape. Using the Cytoscape STRING app, a network was retrieved for 78 proteins interacting with TrkA (tropomyosin-related kinase A) 10 min after stimulating neuroblastoma cells with NGF (nerve growth factor) (56). With a confidence cutoff of 0.4, the resulting network contains 182 functional associations between 57 of the proteins; the 21 proteins with no associations to other proteins in the network were removed. Nodes are colored according to the protein abundance (log ratio) compared to the cells before NGF treatment. The confidence score of each interaction is mapped to the edge thickness and opacity.

host protein–protein interactions, which will incorporate many of the evidence channels present in STRING. This specialized database will enable querying for a whole virus or for specific viral proteins and will superimpose the viral interaction network onto that of the host.

Furthermore, we plan to extend the analysis options for user-provided gene set input, addressing a frequently expressed user need. This will include the possibility to report statistical enrichments for ranked genes lists, even genome-wide rankings. Together with the up-to-date network information, this will allow users to extract the maximum functional information from their queries, for any organism of interest.

## ACKNOWLEDGEMENTS

The authors are indebted to Yan P. Yuan (EMBL Heidelberg) for IT support, and to Dr. Thomas Rattei (University of Vienna) for producing and sharing systematic, all-against-all protein-protein similarity data.

## FUNDING

Core funding for STRING comes from the Swiss Institute of Bioinformatics (Lausanne), the Novo Nordisk Foundation (Copenhagen, NNF14CC0001), and the European Molecular Biology Laboratory (EMBL Heidelberg). J.H.M. has been funded by NIHGMS grant P41-GM103311. Funding for Open Access charges: University of Zurich.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **22**, 1317–1321.
2. Gao,M. and Skolnick,J. (2010) Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 22517–22522.
3. Garma,L., Mukherjee,S., Mitra,P. and Zhang,Y. (2012)How many protein-protein interactions types exist in nature? *PLoS One*, **7**, e38913.
4. 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.
5. Snel,B., Bork,P. and Huynen,M.A. (2002)The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 5890–5895.
6. Rives,A.W. and Galitski,T. (2003)Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 1128–1133.
7. De Las Rivas,J and de Luis,A (2004)Interactome data and databases: different types of protein interaction. *Comp. Funct. Genomics*, **5**, 173–178.
8. Dannenfelser,R., Clark,N.R. and Ma'ayan,A. (2012) Genes2FANs: connecting genes through functional association networks. *BMC Bioinformatics*, **13**, 156–168.
9. Studham,M.E., Tjärnberg,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.
10. Cun,Y. and Frohlich,H. (2013) Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One*, **8**, e73074.
11. Hofree,M., Shen,J.P., Carter,H., Gross,A. and Ideker,T. (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.
12. Guney,E., Menche,J., Vidal,M. and Barábasi,A.L. (2016) Network-based in silico drug efficacy screening. *Nat. Commun.*, **7**, 10331–10343.
13. Hillenmeyer,S., Davis,L.K., Gamazon,E.R., Cook,E.H., Cox,N.J. and Altman,R.B. (2016) STAMS: STRING-Assisted Module Search

for Genome Wide Association Studies and Application to Autism. *Bioinformatics*, doi:10.1093/bioinformatics/btw530.
14. Leiserson,M.D., Eldridge,J.V., Ramachandran,S. and Raphael,B.J. (2013) Network analysis of GWAS data. *Curr. Opin. Genet. Dev.*, **23**, 602–610.
15. Jia,P. and Zhao,Z. (2014) Network.assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Hum. Genet.*, **133**, 125–138.
16. Tasan,M., Musso,G., Hao,T., Vidal,M., MacRae,C.A. and Roth,F.P. (2015) Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nat. Methods*, **12**, 154–159.
17. 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.
18. Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, **29**, 150–159.
19. Tian,W., Zhang,L.V., Taşan,M., Gibbons,F.D., King,O.D., Park,J., Wunderlich,Z., Cherry,J.M. and Roth,F.P. (2008) Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biol.*, **9**(Suppl. 1), S7.
20. 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.
21. 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.
22. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
23. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
24. Das,J. and Yu,H. (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.*, **6**, 92–103.
25. Turner,B., Razick,S., Turinsky,A.L., Vlasblom,J., Crowdy,E.K., Cho,E., Morrison,K., Donaldson,I.M. and Wodak,S.J. (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, doi:10.1093/database/baq023.
26. Alonso-Lopez,D., Gutiérrez,M.A., Lopes,K.P., Prieto,C., Santamaría,R. and De Las Rivas,J. (2016) APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucleic Acids Res.*, **44**, W529–W535.
27. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA prediction server update. *Nucleic Acids Res*, **41**, W115–W122.
28. 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–W133.
29. 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.
30. Schmitt,T., Ogris,C. and Sonnhammer,E.L. (2014) FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res.*, **42**, D380–D388.
31. 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.
32. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
33. 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.
34. 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.
35. 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.
36. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
37. Franceschini,A., Lin,J., von Mering,C. and Jensen,L.J. (2016) SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, **32**, 1085–1087.
38. 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.
39. 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.
40. 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.
41. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
42. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
43. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, **57**, 289–300.
44. 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.
45. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat.Protoc.*, **2**, 2366–2382.
46. Pletscher-Frankild,S., Pallejà,A., Tsafou,K., Binder,J.X. and Jensen,L.J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.
47. Binder,J.X., Pletscher-Frankild,S., Tsafou,K., Stolte,C., O'Donoghue,S.I., Schneider,R. and Jensen,L.J. (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)*, doi:10.1093/database/bau012.
48. Santos,A., Tsafou,K., Stolte,C., Pletscher-Frankild,S., O'Donoghue,S.I. and Jensen,L.J. (2015) Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*, **3**, e1054.
49. Saito,R., Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L., Lotia,S., Pico,A.R., Bader,G.D. and Ideker,T. (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.
50. Morris,J.H., Apeltsin,L., Newman,A.M., Baumbach,J., Wittkop,T., Su,G., Bader,G.D. and Ferrin,T.E. (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, **12**, 436–449.
51. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
52. Bindea,G., Mlecnik,B., Hackl,H., Charoentong,P., Tosolini,M., Kirilovsky,A., Fridman,W.H., Pagès,F., Trajanoski,Z. and Galon,J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

53. Scardoni,G., Tosadori,G., Faizan,M., Spoto,F., Fabbri,F. and Laudanna,C. (2014) Biological network analysis with CentiScaPe: centralities and experimental dataset integration. *F1000Res*, **3**, 139–146.
54. Fishilevich,S., Zimmerman,S., Kohn,A., Iny Stein,T., Olender,T., Kolker,E., Safran,M. and Lancet,D. (2016) Genic insights from integrated human proteomics in GeneCards. *Database (Oxford)*, doi:10.1093/database/baw030.
55. Abel,O., Powell,J.F., Andersen,P.M. and Al-Chalabi,A. (2012) ALSoD: A user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.*, **33**, 1345–1351.
56. Emdal,K.B., Pedersen,A.K., Bekker-Jensen,D.B., Tsafou,K.P., Horn,H., Lindner,S., Schulte,J.H., Eggert,A., Jensen,L.J., Francavilla,C. *et al.* (2015) Temporal proteomics of NGF-TrkA signaling identifies an inhibitory role for the E3 ligase Cbl-b in neuroblastoma cell differentiation. *Sci. Signal*, **8**, ra40.