https://doi.org/10.1093/nar/gkac1000

The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest

Damian Szklarczyk[1,2], Rebecca Kirsch[3], Mikaela Koutrouli[3], Katerina Nastou[3], Farrokh Mehryary[4], Radja Hachilif[1,2], Annika L. Gable[1,2], Tao Fang[1,2], Nadezhda T. Doncheva[3], Sampo Pyysalo[4], Peer Bork[5,6,7,8,*], Lars J. Jensen[3,*] and Christian von Mering[1,2,*]

[1]Department of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland, [2]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen N, Denmark, [4]TurkuNLP lab, Department of Computing, University of Turku, 20014 Turku, Finland, [5]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [6]Yonsei Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea, [7]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany and [8]Department of Bioinformatics, Biozentrum, University of Würzburg, 97074 Würzburg, Germany

Received September 15, 2022; Revised October 10, 2022; Editorial Decision October 11, 2022; Accepted October 19, 2022

ABSTRACT

Much of the complexity within cells arises from functional and regulatory interactions among proteins. The core of these interactions is increasingly known, but novel interactions continue to be discovered, and the information remains scattered across different database resources, experimental modalities and levels of mechanistic detail. The STRING database (https://string-db.org/) systematically collects and integrates protein–protein interactions—both physical interactions as well as functional associations. The data originate from a number of sources: automated text mining of the scientific literature, computational interaction predictions from co-expression, conserved genomic context, databases of interaction experiments and known complexes/pathways from curated sources. All of these interactions are critically assessed, scored, and subsequently automatically transferred to less well-studied organisms using hierarchical orthology information. The data can be accessed via the website, but also programmatically and via bulk downloads. The most recent developments in STRING (version 12.0) are: (i) it is now possible to create, browse and analyze a full interaction network for any novel genome of interest, by submitting its complement of encoded proteins, (ii) the co-expression channel now uses variational auto-encoders to predict interactions, and it covers two new sources, single-cell RNA-seq and experimental proteomics data and (iii) the confidence in each experimentally derived interaction is now estimated based on the detection method used, and communicated to the user in the web-interface. Furthermore, STRING continues to enhance its facilities for functional enrichment analysis, which are now fully available also for user-submitted genomes.

INTRODUCTION

A dense network of functional connections among proteins has evolved to support cellular life, forming a multitude of pathways, protein complexes and cellular structures (1,2). Recent developments have further improved our ability to unravel this connectivity, through techniques such as high-throughput genetic screens (3–5), systematic co-fractionation of proteomes (6,7), *in-vivo* proteome-wide cross-linking of proteins (8–10), and deep learning-based computational prediction frameworks such as AlphaFold (11,12). These efforts complement earlier results based on focused, small-scale laboratory studies, yeast two-hybrid screens, affinity purifications, co-crystallization and computational prediction algorithms (reviewed in (13)). Together, the many approaches have begun to reveal a large part of the interaction landscape of cellular proteins, but they each

*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 Lars J. Jensen. Tel: +45 3 532 5025; Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Peer Bork. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; Email: bork@embl.de

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

have strengths and weaknesses, including potential biases, false negatives and noise. Furthermore, the resulting interaction data is scattered over a number of online resources, and available in varying namespaces and data-formats as well as varying levels of detail.

Given this, a number of meta-resources dedicated to data integration in the protein network context have been developed. These resources aim to collect and critically assess protein–protein functional linkage data, integrate it, connect it to previous knowledge, and allow users to browse, compare and retrieve organism-wide protein networks. Among frequently used and actively maintained frameworks are ConsensusPathDB (14), FunCoup (15), GeneMANIA (16), HumanBase (17), HumanNet (18), IID (19) and STRING (20–22). Within this group, STRING places its focus on comprehensiveness and ease of use—it covers >10 000 organisms, draws from a wide diversity of data sources including text mining and computational predictions, and offers many intuitive interface features including personalization, enrichment detection and programmatic access.

Researchers employ protein network meta-resources for a wide variety of purposes, broadly falling into three categories: (i) facilitating individual, small-scale molecular discoveries, (ii) facilitating large-scale data analysis and (iii) contributing to new methodologies and workflows. With respect to STRING, it has, for example, proven useful in interpreting and reducing newly acquired genetic screening datasets—these are sometimes noisy and unwieldy, and STRING can be used to distill such data into more manageable sets of observations and hypotheses. As a case in point, consider three recent screens for SARS-CoV-2 human host factors (23–25). All three studies used STRING to interpret their initial, raw lists of screening hits—placing them into network contexts and searching them for functionally enriched processes/pathways. STRING and its competitors are also often used as data providers in novel methodologies and computational resources, be it new databases, new algorithms, or community-wide competitions. As an example, consider two recent uses of STRING networks in deep-learning frameworks (26,27). Both studies use deep learning to predict protein function (i.e. Gene Ontology terms), based on amino acid sequences and protein–protein interaction network topologies derived from STRING.

Here, we provide an update on the current features of the STRING database and describe some novel developments in more detail. The latter include a complete redesign of the co-expression based interaction prediction pipeline, newly exposed sub-scores for experimental dataset confidence, as well as novel facilities allowing users to upload and analyze any newly sequenced genome of interest.

DATABASE CONTENT

The scope of protein–protein links in STRING is that of a 'functional association' (28–30)—proteins are considered to be associated when there is evidence suggesting an evolved, specific functional partnership between the two. This definition includes proteins that are physically associated to each other in a protein complex or in a transient interaction, but also proteins that are more indirectly associated: they may work towards a common goal in a metabolic or signaling pathway, may regulate each other through intermediaries, or may jointly contribute to a common cellular structure. The granularity of what constitutes a 'common function' is not formally defined; it should not be understood too broadly, however, and operationally it roughly corresponds to the specificity of pathways 'maps' or 'diagrams' in knowledgebases such as KEGG (31) or Reactome (32). It should be noted that the definition of a functional assocation can include proteins that act antagonistically to each other, albeit in the same overall pathway.

All protein–protein association evidence in STRING is assessed and quantified, and its correspondence to the above definition is benchmarked against common membership in KEGG pathway maps (excluding maps that are largely based on homology, such as 'ABC transporters'). The result of this benchmarking/calibration is a STRING 'confidence score' for each association; confidence scores are scaled between zero and one, and correspond to the estimated likelihood of a given association being true, given the underlying evidence. Confidence scores are first computed separately per evidence type (see (33) for an example), and then integrated into a final, 'combined' confidence score. All evidence collected for a given protein pair contributes to the score, irrespective of the exact nature of these proteins in terms of alternative splicing isoforms or post-translational modifications; correspondingly, the interacting unit in STRING networks is the entire protein-coding locus, represented by its most canonical protein product (34).

All confidence scores in STRING are pre-computed and freely available for download, under a Creative Commons Attribution license (CC BY 4.0). The various evidence types are grouped into seven distinct 'evidence channels', with separate sub-scores available for each channel. These channels can also be individually viewed on the STRING user interface, together with their underlying evidence, and can be enabled or disabled separately as desired. Of the seven channels, the first three ('neighborhood', 'fusion' and 'co-occurrence') are dealing with association evidence that can be gleaned from genome sequences alone. These so-called 'genomic context' channels (reviewed in (35,36)) are based on detecting evolutionary constraints arising from functional gene-gene partnerships, and are best applicable in prokaryotic genomes. Another channel ('co-expression') is dealing with functional genomics measurements (transcripts or proteins) across a multitude of conditions, searching for evidence of common expression regulation (see also below). Next, the 'experiments' channel deals with laboratory experiments that were conducted with the expressed goal to uncover protein–protein association evidence. They are imported from primary database repositories: BioGRID (37), DIP (38), PDB (39), as well as IntAct and its partner databases in the IMEx consortium (40). The final two evidence channels are concerned with protein–protein associations that are already known. The 'database' channel imports well-established knowledge ('textbook knowledge') about protein complexes, pathways and other functional connections from dedicated knowledge resources: KEGG (31), Reactome (32), MetaCyc (41), EBI Complex Portal (42), and Gene Ontology Complexes

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

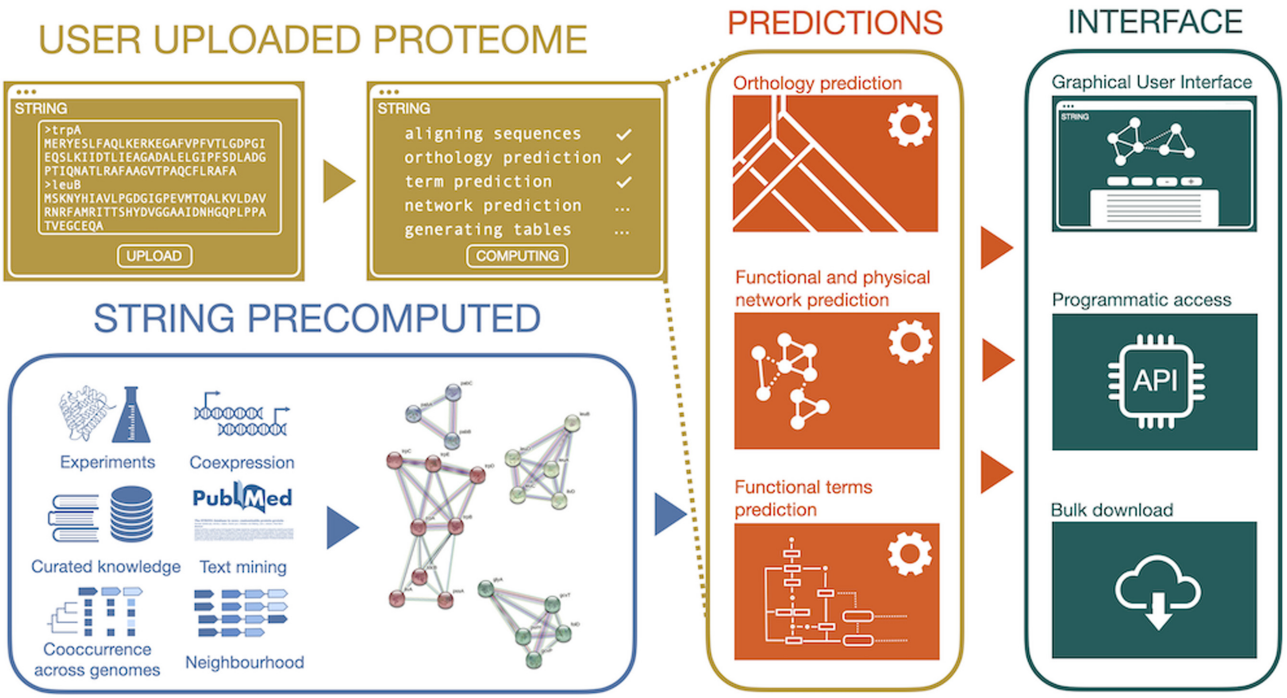

Figure 1. Extending STRING with user-submitted genomes. Submitted genomes are first searched against the existing STRING genomes, and orthology is used to transfer all relevant information (interactions, annotations) from closely related organisms. The submitted genomes then become available on the web interface, via the programmatic Application Programming Interface (API), and for bulk downloads.

(43). Lastly, the 'textmining' channel is the result of parsing full-text articles from the PMC Open Access Subset (up to April 2022), PubMed abstracts (up to August 2022), as well as summary texts from OMIM (44) and SGD (45) entry descriptions. These texts are all parsed for co-mentions of protein pairs and assessed against the frequencies of all separate mentions of the respective proteins, as described in (46). On top of functional associations, the three latter channels (i.e. experiments, database and textmining) provide also the interactions for the physical sub-network of STRING. The calculation of the confidence scores for the protein pairs in this network differs from that of the functional association network and is described in detail in (22).

The deep learning-based relation extraction text mining model has received significant upgrades in the current version. Specifically, the language representation model that we use has changed from BioBERT (47) to a biomedical RoBERTa-large model (48), which has already shown better performance in relation extraction tasks (49). Moreover, the model can now detect physical interactions that span across the boundaries of a single sentence. This is made possible mainly due to the two orders of magnitude increase of manually annotated relations in the training set (from 6145 to 243 831), which led to the addition of many cross-sentence pairs during training, thus allowing the model to learn to detect such associations. These changes in the deep learning model, in combination with the increase in the size of the literature corpus compared to the previous STRING version, have led to a two-fold increase in the number of physical protein–protein interactions above the low, medium, and high confidence score cut-offs in the text mining channel (Table 1).

Table 1. Counts and relative frequencies of physically interacting protein pairs obtained via text mining. Aggregating across all organisms in STRING, the table shows counts for various frequently used score cutoff levels, for both STRING version 11.5 and STRING version 12.0. The lowest score cut-off has been used to determine what constitutes 100% for each dataset

score	Number of pairs in v. 11.5	Number of pairs in v. 12.0	Frequency of pairs above score cut-off in v. 11.5 (%)	Frequency of pairs above score cut-off in v. 12.0 (%)
0.15	253 626	401 976	100.0	100.0
0.4	70 148	143 591	27.6	35.7
0.7	22 349	45 981	8.8	11.4
0.9	0	21 689	0	5.4

All protein–protein associations assembled for the STRING database are then transferred across organisms, where applicable, based on orthology relationships with the assumption that orthologs of associated proteins are likewise associated ('interologs', (50)). For this, hierarchical orthology relationships, at various levels of taxonomic resolution, are imported from the eggNOG database (51). After interolog transfer and the final combined score integration, the resulting protein association networks can be accessed in a number of ways. Firstly, the interactive website of STRING allows browsing and searching, including evidence inspection via dedicated viewers. Users can also submit larger queries there, allowing for the construction of dedicated networks and for statistical searches for functional enrichments. Apart from the website, scientists can access STRING via a dedicated Cytoscape plugin (52), as

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

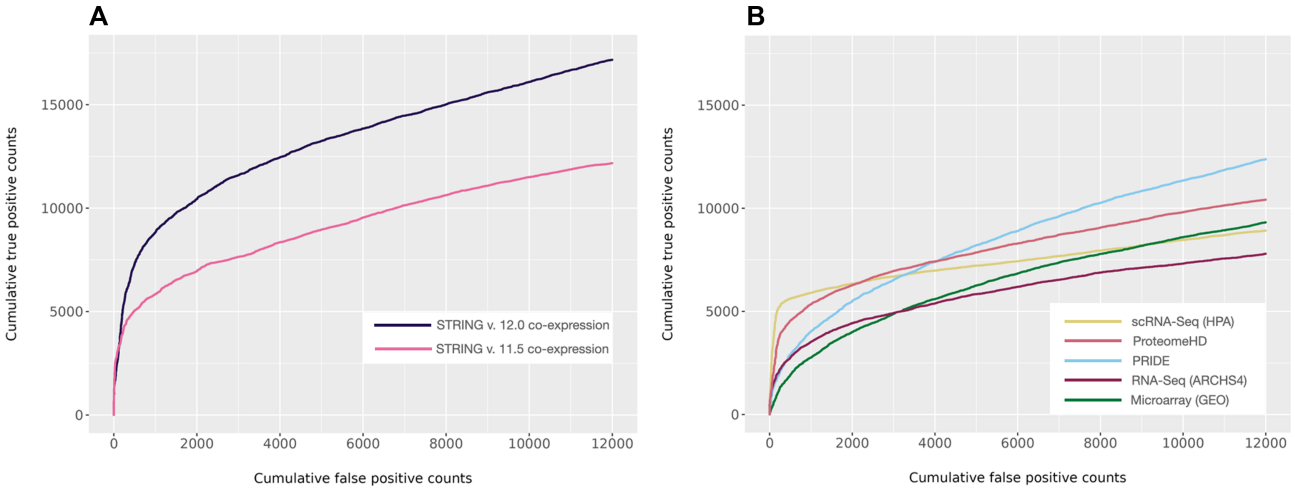

Figure 2. Improved interaction prediction based on co-expression. Interaction scores based on co-expression have been ranked and benchmarked against common KEGG pathway membership as ground truth. (**A**) Performance comparison of co-expression network between STRING version 11.5 and STRING version 12.0. (**B**) Overview of the performance of all expression datasets contributing to the STRING version 12.0 co-expression channel.

well as via a Bioconductor package and a dedicated application programming interface (REST).

USER-UPLOADED GENOMES

One of the unique features of STRING has always been its support for a large selection of non-model organisms: the current version of the database contains protein–protein interaction networks (and protein functional pathway annotations) for >10 000 distinct genomes. When selecting which genomes to include in a STRING update, key factors are the organism's research prominence, genome quality, and completeness. Subsequently, STRING utilizes genomes from authoritative sources only, including Ensembl (53), UniProtKB Reference Proteomes (54) and the 'representative genomes' set in the proGenomes database (55).

However, new genomes are sequenced and assembled daily, and existing genomes are re-sequenced or re-annotated; overall, the number of taxonomically distinct species that are being sequenced has been doubling roughly every 3 years (56). The UniProtKB database is updated with the new proteomes on an eight-week cycle, and projects like Ensembl Rapid Release (53) aim to provide annotation for the newly released genomes on a 2-week cycle. For STRING, such frequent update cycles would require heavy resources and may create complexities with data reproducibility/data archiving when the new updates supersede older datasets. On the other hand, a slow update cycle implies that the database will not incorporate newly sequenced genomes or any improvements to the gene sets of existing genomes.

To improve this situation, STRING now allows its users to upload any fully sequenced proteome (including those that are already part of the database), enabling them to browse and query the predicted protein–protein interactions in an identical manner to the genomes already natively covered by the STRING database. This includes access to the evidence viewers, homology viewer, network clustering methods, gene set enrichment analysis, bulk download, and REST API access. The outline of the proteome annotation pipeline is shown in Figure 1. The procedure for uploading a new proteome begins by choosing 'Annotate your proteome' on the STRING input page, after which the user is guided through a few simple steps of the process. All that is required for the submission is a simple FASTA-formatted proteome, as well as the taxonomic name of the species or clade that the uploaded proteome belongs to. Along with the protein sequences from the file, if provided, STRING will extract from the FASTA definition lines (headers) any standard gene names, identifiers, and free-text protein descriptions; these will later be searchable from the input page and used throughout the webpage. For this, STRING automatically recognizes several formats of FASTA headers including those from RefSeq (57), UniProtKB (54) and Ensembl (53), and checks for any apparent errors such as duplicate sequence identifiers. After the proteome is uploaded, STRING directly aligns the sequences to all sequences in its database using DIAMOND (58) (with the –iterate option). Each protein is then assigned to its orthologs via the hierarchical orthology database eggNOG (51) based on its highest scoring alignment (best hit). The taxonomic level at which the protein is placed in the group hierarchy is defined by the last common ancestor between the user-specified taxon and the taxon of the best-scoring hit. If a protein cannot be placed via the hierarchical orthology groups, it is considered a direct one-to-one ortholog with its best-scoring hit. As the user can submit a proteome of an organism already included in the database, the proteins of the matching proteomes will then directly map to each other, without a need for mapping to any of the existing orthologous groups.

The network prediction based on orthology is then performed as described previously (46). In parallel with the network prediction, STRING also attempts to assign the submitted proteins to their corresponding pathways and functional subsystems, as imported for STRING from the three Gene Ontology branches (Biological process, Molecular Function and Cellular Component) (43), KEGG pathways (31), UniProtKB Keywords (54), COMPARTMENTS (59)

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

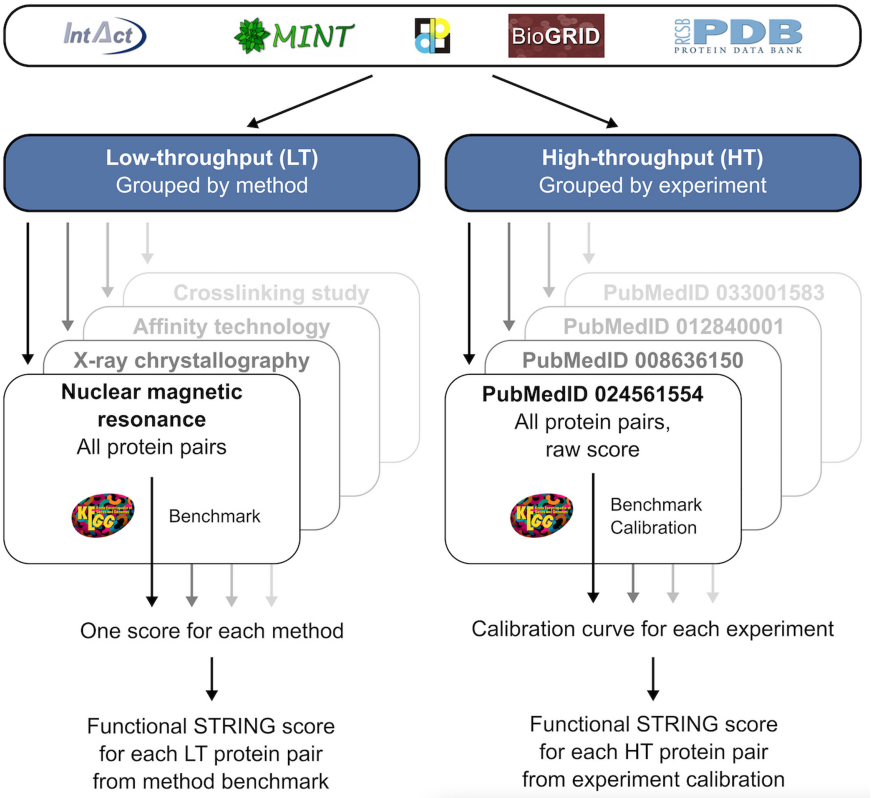

Figure 3. Processing and scoring of experimental interaction evidence. Experimental evidence is retrieved from several public databases. Protein pairs from low-throughput (LT) experiments are grouped by detection method and pairs coming from high-throughput (HT) experiments are grouped by experiment. Within each group, pairs are benchmarked against the KEGG pathway database to assess the confidence of identifying functional associations. All LT pairs are assigned the benchmark score derived for the particular detection method. HT pairs are scored based on calibration on the experiment level.

and TISSUES (60). This assignment is done based on the consensus of the pathway annotations of the pre-existing STRING proteins in the most specific orthologous group that a new protein is assigned to. If these pre-existing proteins carry no pathway annotation, the consensus is attempted in the next-higher, parental orthologous group in the orthology hierarchy; this is repeated until it is successful or until the root level is reached. STRING does not assign pathways or functional subsystems outside their previously known taxonomic scope. A pathway's scope is defined by the last common ancestor of all organisms that have in the past been annotated with this pathway. At the end of the computation process, all predictions are automatically uploaded to the internal STRING database, bulk download files are generated, and the user is given a unique STRING proteome identifier. This identifier functions as an organism name on the input page, making the newly submitted genome browsable and searchable. Submitters can share this identifier with other users, but if they choose not to, the submitted genomes remain private.

IMPROVED CO-EXPRESSION ANALYSIS

The degree of co-expression between RNAs (or between proteins) across different conditions provides an essential insight into the functional protein–protein interaction network of a cell (61–64). STRING collects gene expression evidence from RNA expression arrays and RNA-Seq datasets processed by the GEO database (65) as well as co-regulation evidence from the ProteomeHD database (7). In version 12.0 of STRING, the co-expression network is being extended with evidence from two novel sources: single-cell RNA-Seq data from the Human Protein Atlas (66) and proteomics datasets from the PRIDE database (67).

The expression data tends to be sparse, high-dimensional and highly redundant. These attributes decrease the performance of correlations using Pearson Correlation Coefficient. For previous versions of STRING we have reduced the redundancy by removing highly correlated expression matrices before correlating the gene expression which, in turn, increased the recovery of the functional associations derived from these matrices. However, this did not fully address all of the challenges of such datasets.

To address that, in the latest version we have utilized a novel method called FAVA (Functional Associations using Variational Autoencoders) (68) to build STRING's co-expression network. This deep-learning model reduces the dimensionality of the data into lower-dimensional latent spaces using variational autoencoders (VAE). The benefit of encoding the matrices into fewer dimensions is 2-fold: it reduces the overall sparsity of the data and limits redundancy by essentially compressing the data. The predictions obtained from all the sources are combined in a

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

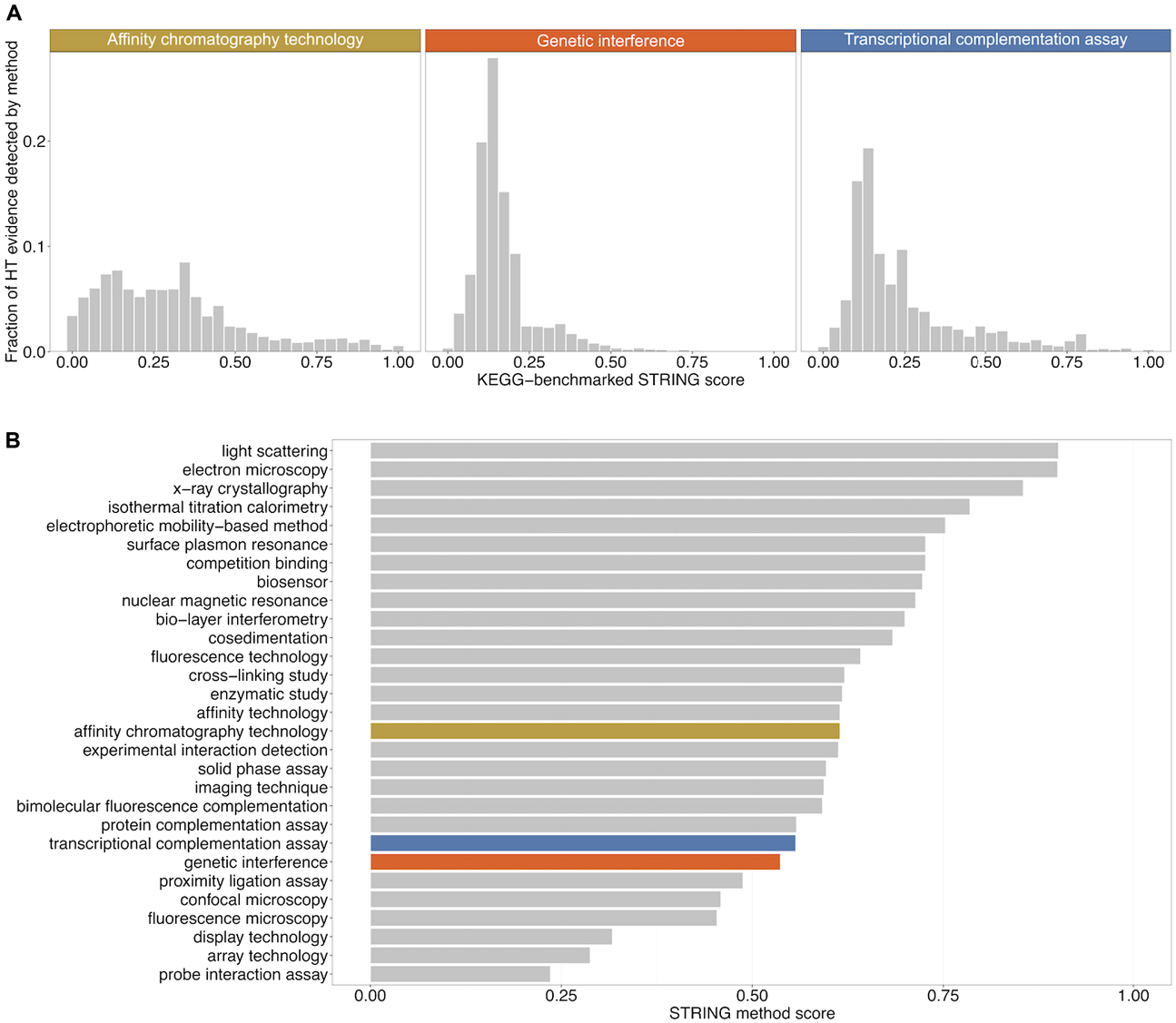

Figure 4. Reliability estimates for protein–protein interaction detection assays. The top three most prolific experimental interaction assay types are shown, ranked by the number of protein pairs to which they contribute in STRING. Benchmarked on KEGG pathways, they yield (on average) lower confidence scores when derived from the high-throughput assays (**A**), in contrast, for the equivalent low-throughput experiments (shown in color), the predicted confidence is substantially higher (**B**). The distributions shown in (A) encompass all HT interactions of a given assay type (for simplicity); in the actual scoring computations in STRING, each HT interaction datasets is scored separately.

probabilistic fashion (46) and the resulting scores are re-calibrated to negate the effect that the non-independence of the sources has on the STRING network. New methods together with the additional sources result in significant improvement in the performance of STRING version 12.0 over STRING version 11.5 co-expression network (Figure 2A). The performance of each individual source contributing to the combined network is shown in Figure 2B. In addition, the STRING interface now reports the associated fractional score and the source of each piece of evidence in the co-expression network.

RNA-Seq datasets have increased sensitivity for genes exhibiting low levels of expression (69,70). This combined with improved performance and the high-throughput nature of the experimental data reduces the inherent literature bias of the STRING network.

EXPERIMENT-LEVEL CONFIDENCE SCORING

For the experiments channel, STRING integrates pairwise experimental interaction evidence from BioGRID, IntAct, MINT, and others ((37–40), see Figure 3). Each reported protein pair supported by a specific publication is considered an independent piece of evidence and scored individually. During import, duplicate records from different sources are removed. Each 'experiment', defined as all protein–protein pairs supported by a common PubMed identifier and a given detection method, is classified as a high-throughput (HT) experiment if at least 25 unique interactions are reported, otherwise as a low-throughput (LT) experiment.

For interactions detected by HT experiments, benchmarking against KEGG molecular pathway maps and scor-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

ing is performed separately for each experiment: each pair is assigned a raw score based on the number of shared and non-shared interactors for both proteins within this experiment. The more shared and the fewer non-shared interactors, the higher the raw quality score of the protein pair from this experiment. Based on the raw score, all pairs detected in the experiment are ranked and benchmarked against KEGG. A typical STRING calibration function is derived for each HT experiment, which is then used to assign a functional association score to each protein pair identified in the experiment (33).

For the benchmarking and scoring of LT experiments, STRING makes use of the Molecular Interaction Controlled Vocabulary (PSI-MI CV, https://www.ebi.ac.uk/ols/ontologies/mi, (71)). Specifically, STRING considers the methods annotated as 'experimental interaction detection' (MI:0045) and its children but excludes those that are inferred by the author/curator (MI:0363 and MI:0364) or predicted (MI:0063). Among the included methods, evidence generated by 'genetic interference' methods (MI:0254) is considered as strictly functional associations, while all other methods provide physical interaction evidence, which inherently is evidence for functional associations as well. Protein pairs are grouped by experimental interaction detection method, to obtain groups that are large enough for benchmarking. Because protein pairs detected by LT experiments tend to focus on specific pathways or known proteins, no curve-based score calibration is done. Instead, STRING directly translates the overall true-positive rate of each interaction detection method into the confidence score for that method. Each protein pair detected in an LT experiment by this method is then assigned the corresponding method's score. By aggregating all scores for a particular protein pair across experiments, one experimental functional association score for that pair is derived.

The highest number of experimentally detected associations has been derived by affinity chromatography technologies (40%), followed by genetic inference (34%) and transcriptional complementation assays (11%). Out of these, 82% came from HT experiments. The associations derived from these three HT experiments, on average, place below or around a confidence score of 0.25 (Figure 4A), while, in comparison, the LT experiments for the same methods score more than twice as high with a medium confidence score of around 0.6 (Figure 4B). The top three methods by confidence score for LT-derived associations are assays determining the 3D structure of protein complexes, which not surprisingly are excellent predictors of functional and physical protein–protein interactions.

In STRING's 'experimental' evidence viewer, the user can now better appreciate the reliability of each experimental prediction, as the individual confidence for every HT and LT dataset is now communicated to the user in a form of three-tier confidence grade ('high', 'medium' and 'exploratory').

OUTLOOK

Version 12.0 of STRING covers a phylogenetically diverse collection of 12 535 high-quality genomes. Beyond these, the system will record which genomes are frequently submitted by users—and these will then be prioritized for inclusion into subsequent releases. In addition, the results of an ongoing online user survey will be analyzed (350 users have already participated). This way, STRING will keep concentrating its resources on those areas that are of most interest to its users.

DATA AVAILABILITY

STRING is freely available, under a Creative Commons Attribution license (CC BY 4.0).

ACKNOWLEDGEMENTS

The authors wish to thank Yan P. Yuan (EMBL Heidelberg), Joao F. Matias-Rodrigues (University of Zurich) and Dandan Xue (University of Copenhagen) for IT support. Thomas Rattei (University of Vienna) is thanked for extensive computational work towards resolving orthology relations. We thank the CSC-IT Center for Science, Finland, for generous computational resources.

FUNDING

Swiss Institute of Bioinformatics; Novo Nordisk Foundation [NNF14CC0001, NNF20SA0035590]; European Molecular Biology Laboratory (EMBL Heidelberg); K.N. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie [101023676]; F.M. and S.P. have received funding from the Academy of Finland [332844]. Funding for open access charges: University of Zurich.
Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M. (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Sci.*, **28**, 1947–1951.
2. Schaffer,L.V. and Ideker,T. (2021) Mapping the multiscale structure of biological systems. *Cell Syst.*, **12**, 622–635.
3. Costanzo,M., Hou,J., Messier,V., Nelson,J., Rahman,M., VanderSluis,B., Wang,W., Pons,C., Ross,C., Ušaj,M. *et al.* (2021) Environmental robustness of the global yeast genetic interaction network. *Science*, **372**, eabf8424.
4. Przybyla,L. and Gilbert,L.A. (2022) A new era in functional genomics screens. *Nat. Rev. Genet.*, **23**, 89–103.
5. Mateus,A., Hevler,J., Bobonis,J., Kurzawa,N., Shah,M., Mitosch,K., Goemans,C.V., Helm,D., Stein,F., Typas,A. *et al.* (2020) The functional proteome landscape of escherichia coli. *Nature*, **588**, 473–478.
6. Drew,K., Wallingford,J.B. and Marcotte,E.M. (2021) hu.MAP 2.0: integration of over 15, 000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Mol. Syst. Biol.*, **17**, e10016.
7. Kustatscher,G., Grabowski,P., Schrader,T.A., Passmore,J.B., Schrader,M. and Rappsilber,J. (2019) Co-regulation map of the human proteome enables identification of protein functions. *Nat. Biotechnol.*, **37**, 1361–1371.
8. Wheat,A., Yu,C., Wang,X., Burke,A.M., Chemmama,I.E., Kaake,R.M., Baker,P., Rychnovsky,S.D., Yang,J. and Huang,L. (2021) Protein interaction landscapes revealed by advanced in vivo cross-linking-mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2023360118.
9. Pfeiffer,C.T., Paulo,J.A., Gygi,S.P. and Rockman,H.A. (2022) Proximity labeling for investigating protein–protein interactions. *Methods Cell Biol.*, **169**, 237–266.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

10. Graziadei,A. and Rappsilber,J. (2022) Leveraging crosslinking mass spectrometry in structural and cell biology. *Structure*, **30**, 37–54.
11. Humphreys,I.R., Pei,J., Baek,M., Krishnakumar,A., Anishchenko,I., Ovchinnikov,S., Zhang,J., Ness,T.J., Banjade,S., Bagde,S.R. *et al.* (2021) Computed structures of core eukaryotic protein complexes. *Science*, **374**, eabm4805.
12. Tunyasuvunakool,K. (2022) The prospects and opportunities of protein structure prediction with aI. *Nat. Rev. Mol. Cell Biol.*, **23**, 445–446.
13. Elhabashy,H., Merino,F., Alva,V., Kohlbacher,O. and Lupas,A.N. (2022) Exploring protein–protein interactions at the proteome level. *Structure*, **30**, 462–475.
14. Kamburov,A. and Herwig,R. (2022) ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Res.*, **50**, D587–D595.
15. Persson,E., Castresana-Aguirre,M., Buzzao,D., Guala,D. and Sonnhammer,E.L.L. (2021) FunCoup 5: functional association networks in all domains of life, supporting directed links and tissue-specificity. *J. Mol. Biol.*, **433**, 166835.
16. 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.
17. Greene,C.S., Krishnan,A., Wong,A.K., Ricciotti,E., Zelaya,R.A., Himmelstein,D.S., Zhang,R., Hartmann,B.M., Zaslavsky,E., Sealfon,S.C. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.
18. Kim,C.Y., Baek,S., Cha,J., Yang,S., Kim,E., Marcotte,E.M., Hart,T. and Lee,I. (2022) HumanNet v3: an improved database of human gene networks for disease research. *Nucleic Acids Res.*, **50**, D632–D639.
19. Kotlyar,M., Pastrello,C., Ahmed,Z., Chee,J., Varyova,Z. and Jurisica,I. (2022) IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. *Nucleic Acids Res.*, **50**, D640–D647.
20. 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.
21. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
22. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
23. Wei,J., Alfajaro,M.M., DeWeirdt,P.C., Hanna,R.E., Lu-Culligan,W.J., Cai,W.L., Strine,M.S., Zhang,S.-M., Graziano,V.R., Schmitz,C.O. *et al.* (2021) Genome-wide CRISPR screens reveal host factors critical for SARS-CoV-2 infection. *Cell*, **184**, 76–91.
24. Schneider,W.M., Luna,J.M., Hoffmann,H.-H., Sánchez-Rivera,F.J., Leal,A.A., Ashbrook,A.W., Le Pen,J., Ricardo-Lax,I., Michailidis,E., Peace,A. *et al.* (2021) Genome-Scale identification of SARS-CoV-2 and Pan-coronavirus host factor networks. *Cell*, **184**, 120–132.
25. Biering,S.B., Sarnik,S.A., Wang,E., Zengel,J.R., Leist,S.R., Schäfer,A., Sathyan,V., Hawkins,P., Okuda,K., Tau,C. *et al.* (2022) Genome-wide bidirectional CRISPR screens identify mucins as host factors modulating SARS-CoV-2 infection. *Nat. Genet.*, **54**, 1078–1089.
26. Kulmanov,M., Khan,M.A., Hoehndorf,R. and Wren,J. (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.
27. Zhang,F., Song,H., Zeng,M., Li,Y., Kurgan,L. and Li,M. (2019) DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics*, **19**, e1900019.
28. 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.
29. 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.
30. Guala,D., Ogris,C., Müller,N. and Sonnhammer,E.L.L. (2020) Genome-wide functional association networks: background, data & state-of-the-art resources. *Brief Bioinform*, **21**, 1224–1237.
31. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
32. Gillespie,M., Jassal,B., Stephan,R., Milacic,M., Rothfels,K., Senff-Ribeiro,A., Griss,J., Sevilla,C., Matthews,L., Gong,C. *et al.* (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, **50**, D687–D692.
33. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.
34. Morales,J., Pujar,S., Loveland,J.E., Astashyn,A., Bennett,R., Berry,A., Cox,E., Davidson,C., Ermolaeva,O., Farrell,C.M. *et al.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, **604**, 310–315.
35. 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.
36. Skrabanek,L., Saini,H.K., Bader,G.D. and Enright,A.J. (2008) Computational prediction of protein–protein interactions. *Mol. Biotechnol.*, **38**, 445–468.
37. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B.-J., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F. *et al.* (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.
38. 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–D51.
39. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
40. 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.
41. Caspi,R., Billington,R., Keseler,I.M., Kothari,A., Krummenacker,M., Midford,P.E., Ong,W.K., Paley,S., Subhraveti,P. and Karp,P.D. (2020) The metacyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.*, **48**, D445–D453.
42. Meldal,B.H.M., Perfetto,L., Combe,C., Lubiana,T., Ferreira Cavalcante,J.V., Bye-A-Jee,H., Waagmeester,A., Del-Toro,N., Shrivastava,A., Barrera,E. *et al.* (2022) Complex portal 2022: new curation frontiers. *Nucleic Acids Res.*, **50**, D578–D586.
43. Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
44. Amberger,J.S., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.*, **47**, D1038–D1043.
45. 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.
46. 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.
47. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
48. Lewis,P., Ott,M., Du,J. and Stoyanov,V. (2020) Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022

49. Miranda,A., Mehryary,F., Luoma,J., Pyysalo,S., Valencia,A. and Krallinger,M. (2021) Overview of drugprot biocreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. *BioCreative*. https://biocreative.bioinformatics.udel.edu/media/store/files/2021/Track1_pos_1_BC7_overview.pdf.
50. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.-D.J., 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.
51. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
52. Doncheva,N.T., Morris,J.H., Gorodkin,J. and Jensen,L.J. (2019) Cytoscape stringapp: network analysis and visualization of proteomics data. *J. Proteome Res.*, **18**, 623–632.
53. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Bennett,R. *et al.* (2022) Ensembl 2022. *Nucleic Acids Res.*, **50**, D988–D995.
54. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
55. Mende,D.R., Letunic,I., Maistrenko,O.M., Schmidt,T.S.B., Milanese,A., Paoli,L., Hernández-Plaza,A., Orakov,A.N., Forslund,S.K., Sunagawa,S. *et al.* (2020) proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res.*, **48**, D621–D625.
56. Nasko,D.J., Koren,S., Phillippy,A.M. and Treangen,T.J. (2018) RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.*, **19**, 165.
57. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,M.K., Durkin,A.S. *et al.* (2021) RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
58. Buchfink,B., Reuter,K. and Drost,H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.
59. 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*, **2014**, bau012.
60. 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*, **2018**, bay003.
61. Zhong,W. and Sternberg,P.W. (2006) Genome-wide prediction of c. elegans genetic interactions. *Science*, **311**, 1481–1484.
62. Raina,P., Lopes,I., Chatsirisupachai,K., Farooq,Z. and de Magalhães,J.P. (2021) GeneFriends 2021: updated co-expression databases and tools for human and mouse genes and transcripts. bioRxiv doi: https://doi.org/10.1101/2021.01.10.426125, 10 January 2021, preprint: not peer reviewed.
63. Harris,B.D., Crow,M., Fischer,S. and Gillis,J. (2021) Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Syst.*, **12**, 748–756.
64. Kanonidis,E.I., Roy,M.M., Deighton,R.F. and Le Bihan,T. (2016) Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: case of a glioblastoma multiforme study. *PLoS One*, **11**, e0161828.
65. 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.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
66. Sjöstedt,E., Zhong,W., Fagerberg,L., Karlsson,M., Mitsios,N., Adori,C., Oksvold,P., Edfors,F., Limiszewska,A., Hikmet,F. *et al.* (2020) An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science*, **367**, eaay5947.
67. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M. *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic. Acids. Res.*, **50**, D543–D552.
68. Koutrouli,M., Líndez,P.P., Bouwmeester,R., Martens,L. and Jensen,L.J. (2022) FAVA: high-quality functional association networks inferred from scRNA-seq and proteomics data. bioRxiv doi: https://doi.org/10.1101/2022.07.06.499022, 07 July 2022, preprint: not peer reviewed.
69. Zhao,S., Fung-Leung,W.-P., Bittner,A., Ngo,K. and Liu,X. (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated t cells. *PLoS One*, **9**, e78644.
70. Rai,M.F., Tycksen,E.D., Sandell,L.J. and Brophy,R.H. (2017) Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. *J. Orthop. Res.*, **36**, 484–497.
71. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1000/6825349 by EMBL user on 21 November 2022