Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa348

CoCoCoNet: conserved and comparative co-expression across a diverse set of species

John Lee†, Manthan Shah†, Sara Ballouz, Megan Crow and Jesse Gillis*

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Blvd., Woodbury, NY 11797, USA

Received March 12, 2020; Revised April 21, 2020; Editorial Decision April 24, 2020; Accepted April 24, 2020

ABSTRACT

Co-expression analysis has provided insight into gene function in organisms from *Arabidopsis* to zebrafish. Comparison across species has the potential to enrich these results, for example by prioritizing among candidate human disease genes based on their network properties or by finding alternative model systems where their co-expression is conserved. Here, we present CoCoCoNet as a tool for identifying conserved gene modules and comparing co-expression networks. CoCoCoNet is a resource for both data and methods, providing gold standard networks and sophisticated tools for on-the-fly comparative analyses across 14 species. We show how CoCoCoNet can be used in two use cases. In the first, we demonstrate deep conservation of a nucleolus gene module across very divergent organisms, and in the second, we show how the heterogeneity of autism mechanisms in humans can be broken down by functional groups and translated to model organisms. CoCoCoNet is free to use and available to all at https://milton.cshl.edu/CoCoCoNet, with data and R scripts available at ftp://milton.cshl.edu/data.

INTRODUCTION

How a gene's expression level changes across conditions is a rich source of information about its function, a fact that gene co-expression networks aim to capture in a general framework (1). Gene co-expression networks link genes by their similarity in expression pattern, yielding connected subnetworks that are likely to share biological functions (2). One of the most important uses of co-expression networks is to test whether a newly identified set of genes forms a clear module (3). Once that is established, the specific topology within the network can be studied in detail to determine central nodes or to define critical co-expression relationships (4,5).

The utility of expression as a readout across biological systems has allowed co-expression network analysis to be applied very broadly: to group and classify genes in model organisms [e.g. *Arabidopsis* (6,7), mice (8) and yeast (9)], to find and characterize disease genes [e.g. in autism (10), Parkinson's disease (11) and heart disease (12)] and as an important contributor to sophisticated algorithms for inferring gene properties [e.g. miRNA targets (13), transcription factor regulation (14) and Gene Ontology (GO) annotations (15,16)]. Because evolution often works by rewiring existing gene–gene relationships, a particularly important area of co-expression analysis is cross-species comparison. Though it is well established that cross-species analyses can enrich for biologically relevant modules (17), even simple comparisons remain very challenging. With CoCoCoNet, we have aimed to systematize comparative co-expression, expanding the range of species covered in the field as a whole, improving the statistical rigor of network analysis within each species and enhancing the sophistication of integrative analyses across species.

CoCoCoNet allows users to access novel research areas by querying and comparing well-powered co-expression networks for 14 species. With a few clicks, researchers can input their gene or genes of interest, and look for co-expression relationships that may be conserved across large phylogenetic distances. While co-expression is a key component of other web servers and databases such as COXPRESdb (18), ATTED-II (19), GeneFriends (20), PlaNet (21), MouseNet (22) and GeneMANIA (23), few provide data beyond the standard model organisms (human, mouse, fly, roundworm, yeast and *Arabidopsis*). Those that do lack the ability to make cross-species comparisons. For example, PlaNet, COXPRESdb and ATTED-II provide co-expression data for several of the species we cover, but there are no convenient methods to directly compare the networks, nor do they perform any explicit analyses of co-expression strength. In contrast, CoCoCoNet provides users with convenient access to both data and methods for cross-species analyses. This opens up a range of potential research questions, such as the following:

*To whom correspondence should be addressed. Tel: +1 516 422 4041; Fax: +1 516 422 4109; Email: jgillis@cshl.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

- Which genes are related to my target gene, and do those relationships change across species?
- When has co-expression been conserved across large phylogenetic distances?
- Does my gene set subdivide into clusters that are maintained across species?
- Is my gene of interest co-expressed with other genes of interest in species I do not study?

In the following, we summarize the methods, data and operation of CoCoCoNet and walk through two use cases: one focused on highly co-expressed modules in yeast and another focused on autism disease genes. In addition, we provide substantially expanded detail in the Supplementary Data—providing details on network construction, resources used, quality control and a complete walk-through of the web server. We have made all methods and data available for use by other researchers, including the underlying network data and methods for assessing them.

MATERIALS AND METHODS

RNA-seq datasets

Because the quality of co-expression data is highly correlated with the total number of samples across all datasets (4), we aimed to collect as much data as possible for each species. To this end, we searched NCBI's Sequence Read Archive (SRA) database (24), using the R Bioconductor package 'SRAdb' (25) for bulk RNA-sequencing datasets (unique SRA Study IDs), excluding those with <10 samples. Cancer-related studies were also excluded since they are not likely to generalize well. To maximize the independence of co-expression measurements within individual datasets, we included only one replicate (also known as 'run_accession') per unique Biosample ID, choosing the replicate with the maximum amount of data by number of spots. Reference genomes and genome annotation files were downloaded from ENSEMBL (26) (September 2019). Sequence reads were downloaded directly from NCBI's ftp site (ftp://ftp.sra.ebi.ac.uk/vol1/fastq) and were aligned to the reference genome using STAR v2.6.0c (27). See Supplementary Table S1 for more details.

Datasets identified in SRAdb were included in our gold standard co-expression networks if they met two additional criteria: measurable expression of at least 50% of all genes (Supplementary Figure S2) and above-threshold similarity to an aggregate expression profile characterizing all datasets. Procedurally, this means that for every sample, we rank genes by expression level, then average these ranks across all samples within a dataset and finally average these dataset-level results to obtain a 'global average'. Next, we compute the Spearman's correlation between each sample in each dataset and this global average (Supplementary Figures S3 and S4). If the average of the worst 10 correlation coefficients is <0.3, we remove that dataset entirely.

In combination with our minimal sample requirements, these checks ensure that each dataset used in the aggregation of our co-expression networks is both well powered and likely to generalize. Further detail on these datasets can be found in Supplementary Tables S2 and S4.

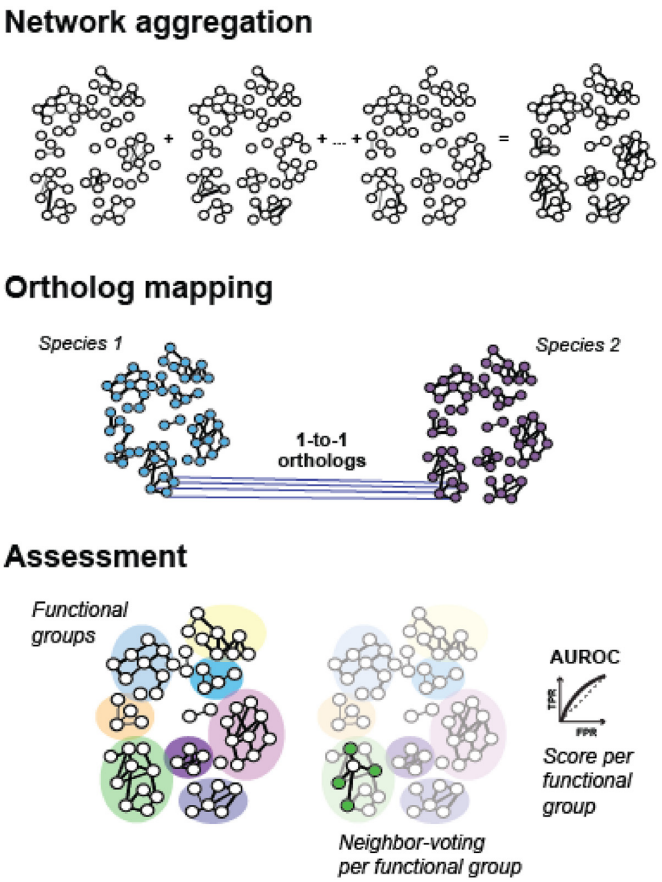

Figure 1. Schematic of underlying data. Co-expression networks are aggregated for each species, ortholog maps are generated for each pair of species and data quality is assessed using a neighbor voting algorithm across all functional groups.

We note that we did not limit our search to a single sequencing platform. In general, platform consistency is maintained within experiments, and co-expression networks are independently constructed and standardized; thus, the aggregation of these controlled networks is not affected by this class of variability. In total, our data comprise 39 517 samples across the 14 species, 34 729 of which utilize Illumina HiSeq 2000 or 2500 (Supplementary Table S4).

Co-expression network construction and aggregation

Co-expression networks for each dataset were constructed by computing Spearman's correlation between every pair of genes (Supplementary Figure S1). This generates a network that is then rank standardized and normalized by dividing through by the maximum rank (4). Genes that are not observed in a particular dataset naturally have no variance, making correlation computations impossible. We replace these NA values with the median value of the network. Networks obtained from individual datasets were then aggregated by adding all of the network adjacency matrices, and then rank standardizing and dividing by the maximum rank (Figure 1).

While other co-expression tools use Pearson's correlation as their primary metric (18–20), we use Spearman's corre-

lation. We have shown in (4) that there is marginal difference in performance using Pearson's correlation over Spearman's correlation. We utilize the non-parametric approach of Spearman's to ensure that outlier values do not have undue influence, allowing results to be driven by the power of larger data.

Within CoCoCoNet, users can choose to query aggregates built with almost all genes or those built with a smaller high-confidence set. Our minimal filter requires that genes be expressed at least once in at least half of the datasets. Genes that fail to meet this requirement are removed from the aggregate co-expression network, yielding the 'almost all genes' set. A more stringent filter allows for faster processing and provides greater confidence in retained links. To filter for genes that are well powered, we count the number of datasets where a gene has at least 10 reads in each of 10 or more samples. 'High-confidence genes' are those that meet these criteria in >20 datasets.

Gene annotations and ortholog mapping

We use the GO (28,29) to obtain gene function annotations. GO terms and gene associations were obtained by merging data from NCBI's website (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) (January 2020) and the Bioconductor package 'biomaRt' (30) (Supplementary Table S3). Terms were then propagated in the ontology tree using a transitive property and filtered to include terms annotating between 10 and 1000 genes. These are then used in enrichment analyses, performed using Fisher's exact test followed by a false discovery rate (FDR) correction.

Ortholog data are obtained from OrthoDB (31), allowing us to provide one-to-one ortholog maps for every pair of species included in CoCoCoNet (Supplementary Table S5). This is accomplished by searching for the most recent phylogenetic split between the two query species and obtaining inferred orthology groups for all genes descended from the common ancestor. Genes are then filtered to the corresponding input species and mapped to each other (Figure 1).

Network assessment

Guilt-by-association-based methods are used to ascertain the quality of co-expression networks (32) and can also be used to determine the connectivity of a gene set. To accomplish this, CoCoCoNet implements functions from the Bioconductor package 'EGAD' (33) on the gene set provided by the user, along with the orthologs from the second species selected, and GO annotations (Figure 1). EGAD measures the performance of a network and a gene set through the neighbor voting algorithm and reports an area under the receiver operating curve (AUROC) or the area under the precision recall curve (AUPRC). These performances can also be compared to predictions based solely on node degree (34). AUCs close to 0.5 indicate poor performance, 0.7 being quite good and 1 being perfect. If the AUCs from both species are high, the tested gene sets and their co-expression modules can be said to be conserved, particularly if the node degree bias is low.

Implementation

This web server is implemented using the open source R Shiny Server (35). In our networks, nodes are genes and edges are normalized average correlation statistics across all underlying datasets, as detailed earlier. Visual clustering of each network is implemented using the physical properties of the network and the 'visNetwork' R package (36). We assign each node a mass proportional to its total node degree, where the larger the mass, the more repulsive the node. A Barnes–Hut *n*-body simulation (37) is applied, forcing high-degree nodes toward cluster centers and low-degree nodes toward the cluster peripheries. Network data are stored in the HDF5 format, which allows for rapid search of specified data. Histograms and scatter plots are generated using the R package 'ggplot2' (38) and made interactive using the R package 'plotly' (39).

Web server description

CoCoCoNet is designed to be simple to use and as intuitive as possible. User interactions are divided into three subsequent phases. The first step simply requests input genes and a species. The second step requires the input of a secondary species, and the final step asks the user what metric to use in characterizing the output subnetworks. Visualizations of the network and the distribution of co-expression values are reported after running the first two steps. In addition, gene set enrichment is applied, and genes with over-represented GO terms can be visualized directly in the subnetwork. In the final step, we characterize the connectivity of the gene set as well as any subnetworks related to GO terms within the gene set. We typically report this as an AUROC, which specifies the degree to which the network topology allows reconstruction of the set of genes used as training, if some fraction of them are hidden (i.e. cross-validation).

Overall, an input of ∼200 genes will render a network within 30 s and implement EGAD for GO groups within 10 s. An input of ∼1000 genes will render a network within 5 min and implement EGAD for GO groups within 1 min. Implementation of EGAD on the gene set takes between 30 s and 5 min depending on the selected species and the gene set to compare. In the interest of user experience, we impose an upper limit of 1000 genes since larger queries may interfere with processes of other users. For larger scale inquiries, we recommend downloading the relevant data and using *CoCoCoNetLite*, available at ftp://milton.cshl.edu/data/scripts/cococonetLite.R. We refer readers to the Supplementary Data (Supplementary Figures S5–S10) for a detailed tutorial and usage guide of CoCoCoNet.

Downloadable

All data and R scripts used to generate results are available at ftp://milton.cshl.edu/data. Data include gene expression networks as HDF5 files, GO annotations, gene ID conversion tables, one-to-one ortholog mappings, the total degree of each gene and example gene lists. Figure 2 contains a detailed break down of the 895 datasets and the 39 517 samples that went into the construction of the aggregate networks, with details available for download. During each step, the user is also able to download relevant data. In

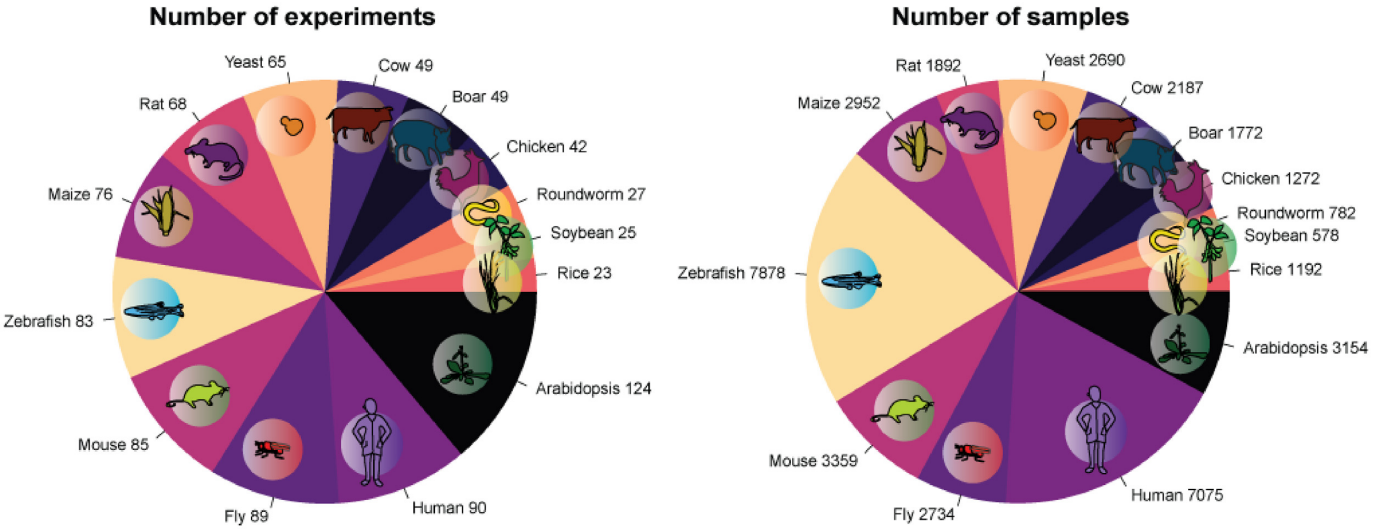

Figure 2. Left: Counts of experiments expressing at least half of all genes. Right: Counts of samples with a correlation with the global average >0.3.

the first two steps, co-expression networks and functional enrichment results can be downloaded, with subnetworks in coordinate format. In the final section, the user is able to download the AUROC (or AUPRC) scores of each GO term for each species.

CASE STUDIES

Highly co-expressed yeast genes

Co-expression was first exploited as a global tool for characterization of gene function by Eisen *et al.* in a study of yeast (2), so for our first use case we returned to this original benchmark gene set to walk through a simple validational use of the main feature of CoCoCoNet. To define an interesting gene set to explore, we first pruned the Eisen list by filtering for genes with very high co-expression with at least one other gene (see the Supplementary Data for details). Then, mapping this set of genes to every other species in CoCoCoNet, we see that most orthologs remain very highly co-expressed with one another, with average co-expression link strengths >0.8 (i.e. in the top 20%; see Figure 3). Beyond the individual network links, the overall topology exhibits strikingly well-defined modules. The first cluster contains primarily ribosomal protein and translation-related genes, in good agreement with group **I** in the 1998 Eisen *et al.* paper. Another cluster contains predominately proteasome-related genes, analogous to group **C**, while the largest cluster contains genes with functions relating to the nucleolus and translation regulation, among others. See Supplementary Figure S11 for a dendrogram and heat map of these 231 genes.

Using the ortholog mapping feature of CoCoCoNet, and restricting our attention to the nucleolus (GO:0005730), we can evaluate the co-expression of the input yeast genes in other species. As expected for a structure that is common to all eukaryotes, we find that this function is highly conserved even at extreme phylogenetic distances (e.g. yeast AUROC = 0.9070, *Arabidopsis* AUROC = 0.9111, zebrafish AUROC = 0.8770, fruit fly AUROC = 0.8320). A common feature of co-expression networks is hub genes that are strongly connected to many others (i.e. they have high node degree). Supporting the specificity of the nucleolus gene–gene connections, we find that our control test, which uses node degree alone to predict module connectivity, has almost no performance (AUROC ≈ 0.5). Together, these results indicate that yeast nucleolus genes form a functional module that is tightly conserved across distant species (Figure 4).

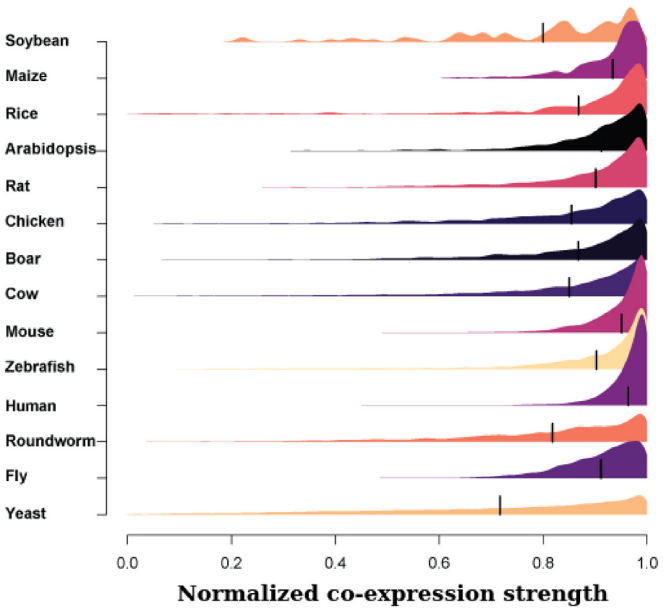

Figure 3. Distribution of co-expression values for ortholog mapped genes to the input of highly co-expressed yeast genes for each of the 13 other species.

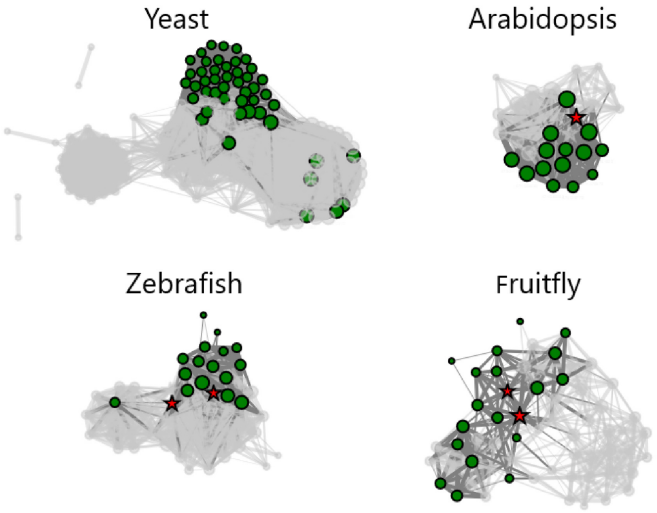

Figure 4. Highly co-expressed yeast (*Saccharomyces cerevisiae*) genes are mapped to orthologous genes in *Arabidopsis* (*Arabidopsis thaliana*), zebrafish (*Danio rerio*) and fruit fly (*Drosophila melanogaster*). Genes annotated with the nucleolus (GO:0005730) are highlighted, and the top 1% of connections are shown. Red stars denote highly connected genes as measured by their node degree.

Autism spectrum disorder-associated genes

The success of translational disease research relies on the conservation of gene function between model organisms and humans. However, in many cases, it remains unclear whether disease mechanisms are sufficiently similar (40,41). Failures of translation have been particularly notable within the neurosciences (42).

Autism spectrum disorder (ASD) is a syndrome with known phenotypic and genetic heterogeneity (43–45). Past analyses have found that ASD genes fall into two major functional categories: those involved in gene expression regulation (GER) and those involved in neuronal communication (NC) (46,47). This suggests that cases may be subtyped based on the gene networks that are affected by rare inherited or *de novo* variants. Here, we consider the co-expression of a set of 102 genes associated with ASD identified by the Autism Sequencing Consortium in (46) along with the corresponding one-to-one orthologs in mouse, where functional translation is likely to be key. These genes were used as input to CoCoCoNet with default parameters.

Enrichment analyses of the 102 gene subnetworks in both mouse and human indicate that GER and NC terms are over-represented, as expected. CoCoCoNet also permits direct comparison of the GER and NC modules within and across species, suggesting which gene relationships can be meaningfully assessed in the mouse as a model system. Inputting the GER and NC gene sets into CoCoCoNet one at a time, we can consider the modularity of each gene set independently using the 'Compute the gene set score' feature. We find that the 58 GER genes have high co-expression edge strengths with one another (average of 0.81), but they are not preferentially connected with one another at all (AUROC of 0.415 in human and 0.556 in mouse). This suggests that while gene regulation is obviously an important function and the strong co-expression edges of the genes reflect this, they also possess equally strong relationships with other genes, making targeted translation between species difficult. In contrast, the 24 NC genes have relatively weak edge strengths (average of 0.53), but are very preferentially connected with one another (AUROC of 0.880 in human and 0.859 in mouse), suggesting a shared mechanism that is conserved between human and mouse (Supplementary Figure S12).

DISCUSSION AND OUTLOOK

Co-expression networks are useful tools for investigating gene function, but they require large-scale data aggregation to be powered, and this has limited their broader use. We have carefully curated and generated aggregate co-expression networks for 14 species, chosen because they have sufficient RNA-sequencing data as well as GO annotations. We share them via the CoCoCoNet web server to aid researchers in their comparative analyses.

CoCoCoNet provides fast enrichment and conservation scores, displayed in a user-friendly manner. Here, we have walked through two applications of CoCoCoNet, but there are many other possibilities. We make it easy to reproduce the analyses done on the web server by providing code alongside visual outputs and quantitative results. In addition, we strongly encourage users to download networks and explore them with their own biological questions in mind. We expect that future releases will encompass data from a wider variety of organisms as new research emerges.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Stephan Fischer, Risa Kawaguchi, Hamsini Suresh, Sukalp Muzumdar, Shaina Lu, Benjamin Harris, Jonathan Werner, Nathan Fox and John Hover for testing the server and offering valuable comments. We thank all researchers who make their data available.

FUNDING

National Institutes of Health [R01 LM012736 and R01 MH113005 to J.L., M.S., S.B. and J.G.; K99 MH120050 to M.C.]. Funding for open access charge: National Institutes of Health [R01 LM012736].
Conflict of interest statement. None declared.

REFERENCES

1. Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, 17.
2. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 14863–14868.
3. Allocco,D.J., Kohane,I.S. and Butte,A.J. (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, **5**, 18.
4. Ballouz,S., Verleyen,W. and Gillis,J. (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, **31**, 2123–2130.

5. Crow,M., Paul,A., Ballouz,S., Huang,Z.J. and Gillis,J. (2016) Exploiting single-cell expression to characterize co-expression replicability. *Genome Biol.*, **17**, 101.
6. Mao,L., Van Hemert,J.L., Dash,S. and Dickerson,J.A. (2009) Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, **10**, 346.
7. Liu,W., Lin,L., Zhang,Z., Liu,S., Gao,K., Lv,Y., Tao,H. and He,H. (2019) Gene co-expression network analysis identifies trait-related modules in *Arabidopsis thaliana*. *Planta*, **249**, 1487–1501.
8. Monaco,G., van Dam,S., Casal Novo Ribeiro,J.L., Larbi,A. and de Magalhaes,J.P. (2015) A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels. *BMC Evol. Biol.*, **15**, 259.
9. Carlson,M.R., Zhang,B., Fang,Z., Mischel,P.S., Horvath,S. and Nelson,S.F. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, **7**, 40.
10. Voineagu,I., Wang,X., Johnston,P., Lowe,J.K., Tian,Y., Horvath,S., Mill,J., Cantor,R.M., Blencowe,B.J. and Geschwind,D.H. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380–384.
11. George,G., Singh,S., Lokappa,S.B. and Varkey,J. (2019) Gene co-expression network analysis for identifying genetic markers in Parkinson's disease: a three-way comparative approach. *Genomics*, **111**, 819–830.
12. Dewey,F.E., Perez,M.V., Wheeler,M.T., Watt,C., Spin,J., Langfelder,P., Horvath,S., Hannenhalli,S., Cappola,T.P. and Ashley,E.A. (2011) Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circ. Cardiovasc. Genet.*, **4**, 26–35.
13. Ammah,A.A., Do,D.N., Bissonnette,N., Gavry,N. and Ibeagha-Awemu,E.M. (2018) Co-expression network analysis identifies miRNA–mRNA networks potentially regulating milk traits and blood metabolites. *Int. J. Mol. Sci.*, **19**, 2500.
14. Liu,Y., Lu,P., Wang,Y., Morrow,B.E., Zhou,B. and Zheng,D. (2019) Spatiotemporal gene coexpression and regulation in mouse cardiomyocytes of early cardiac morphogenesis. *J. Am. Heart Assoc.*, **8**, e012941.
15. Pena-Castillo,L., Tasan,M., Myers,C.L., Lee,H., Joshi,T., Zhang,C., Guan,Y., Leone,M., Pagnani,A., Kim,W.K. *et al.* (2008) A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.*, **9**, S2.
16. Jiang,Y., Oron,T.R., Clark,W.T., Bankapur,A.R., D'Andrea,D., Lepore,R., Funk,C.S., Kahanda,I., Verspoor,K.M., Ben-Hur,A. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.
17. Ruprecht,C., Vaid,N., Proost,S., Persson,S. and Mutwil,M. (2017) Beyond genomics: studying evolution with gene coexpression networks. *Trends Plant Sci.*, **22**, 298–307.
18. Obayashi,T., Kagaya,Y., Aoki,Y., Tadaka,S. and Kinoshita,K. (2019) COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.*, **47**, D55–D62.
19. Obayashi,T., Aoki,Y., Tadaka,S., Kagaya,Y. and Kinoshita,K. (2018) ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.*, **59**, 440.
20. van Dam,S., Cordeiro,R., Craig,T., van Dam,J., Wood,S.H. and de Magalhães,J.P. (2012) GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. *BMC Genomics*, **13**, 535.
21. Proost,S. and Mutwil,M. (2017) PlaNet: comparative co-expression network analyses for plants. *Methods Mol. Biol.*, **1533**, 213–227.
22. Kim,E., Hwang,S., Kim,H., Shim,H., Kang,B., Yang,S., Shim,J.H., Shin,S.Y., Marcotte,E.M. and Lee,I. (2016) MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. *Nucleic Acids Res.*, **44**, D848–D854.
23. 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.
24. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The Sequence Read Archive. *Nucleic Acids Res.*, **39**, 19–21.
25. Zhu,Y., Stephens,R.M., Meltzer,P.S. and Davis,S.R. (2013) SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, **14**, 19.
26. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.
27. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
28. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
29. The,Gene Ontology Consortium (2019) The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
30. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
31. Kriventseva,E.V., Kuznetsov,D., Tegenfeldt,F., Manni,M., Dias,R., Simao,F.A. and Zdobnov,E.M. (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **47**, D807–D811.
32. Oliver,S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.
33. Ballouz,S., Weber,M., Pavlidis,P. and Gillis,J. (2017) EGAD: ultra-fast functional analysis of gene networks. *Bioinformatics*, **33**, 612–614.
34. Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on 'guilt by association' analysis. *PLoS One*, **6**, e17258.
35. Chang,W., Cheng,J., Allaire,J., Xie,Y. and McPherson,J. (2019) shiny: web application framework for R. R package version 1.4.0.2.
36. Almende,B.V., Thieurmel,B. and Robert,T. (2019) visNetwork: network visualization using 'vis.js' library. R package version 2.0.9.
37. Barnes,J. and Hut,P. (1986) A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature*, **324**, 446–449.
38. Wickham,H. (2016) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer, NY.
39. Sievert,C. (2018) plotly for R.
40. Seok,J., Warren,H.S., Cuenca,A.G., Mindrinos,M.N., Baker,H.V., Xu,W., Richards,D.R., McDonald-Smith,G.P., Gao,H., Hennessy,L. *et al.* (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 3507–3512.
41. Takao,K. and Miyakawa,T. (2015) Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc. Natl Acad. Sci. U.S.A.*, **112**, 1167–1172.
42. Kola,I. and Landis,J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.*, **3**, 711–715.
43. Bruining,H., de Sonneville,L., Swaab,H., de Jonge,M., Kas,M., van Engeland,H. and Vorstman,J. (2010) Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. *PLoS One*, **5**, e10887.
44. An,J.Y. and Claudianos,C. (2016) Genetic heterogeneity in autism: from single gene to a pathway perspective. *Neurosci. Biobehav. Rev*, **68**, 442–453.
45. Jeste,S.S. and Geschwind,D.H. (2014) Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat. Rev. Neurol.*, **10**, 74–81.
46. Satterstrom,F.K., Kosmicki,J.A., Wang,J, Breen,M. S., De Rubeis,S., An,J., Peng,M., Collins,R., Grove,J., Klei,L. *et al.*(2020) Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, **180**, 568–584.
47. De Rubeis,S., He,X., Goldberg,A.P., Poultney,C.S., Samocha,K., Cicek,A.E., Kou,Y., Liu,L., Fromer,M., Walker,S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**, 209–215.