CoCoCoNet: Conserved and Comparative Co-expression Across a Diverse Set of Species

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

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

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 which 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 which 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 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:

- 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 on autism disease genes. In addition, we provide substantially expanded detail in our supplement – providing details on

[*]To whom correspondence should be addressed. Email: jgillis@cshl.edu, [†]John Lee and Manthan Shah contributed equally.

network construction, resources used, quality control, and a complete walk-through of the webserver. We have made all methods and data available for use by other researchers, including the underlying network data and methods for assessing it.

MATERIAL 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 fewer than 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 (a.k.a. "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) (Sept 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 table S1 for more details.

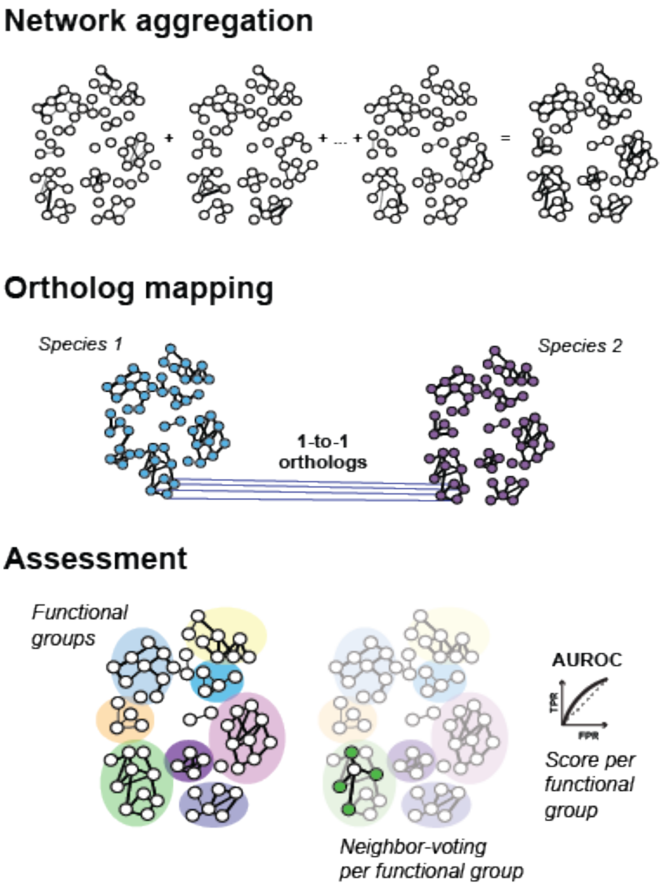


Figure 1. Schematic of underlying data. Co-expression networks are aggregated for each species, ortholog maps are generated for each pair of species, and data quality ~~of data~~ is assessed using a neighbor voting algorithm across all functional groups.

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 (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 (figures S3 and S4). If the average of the worst 10 correlation coefficients is less than 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. Figure 2 contains a detailed summary of the number of experiments and the total number of samples that went into the construction of the aggregate networks. Further detail on these datasets can be found in tables S2 and S4.

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 comprises of 39,517 samples across the 14 species, 34,729 of which utilize Illumina HiSeq 2000 or 2500 (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 (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, 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, 19, 20), we use Spearman's correlation. 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 more than 20 datasets.

Gene annotations and ortholog mapping

We use the Gene Ontology (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) (Jan 2020) and the Bioconductor package "biomaRt" (30) (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 an FDR correction.

Ortholog data is obtained from OrthoDB (31), allowing us to provide 1-to-1 ortholog maps for every pair of species included in CoCoCoNet (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.

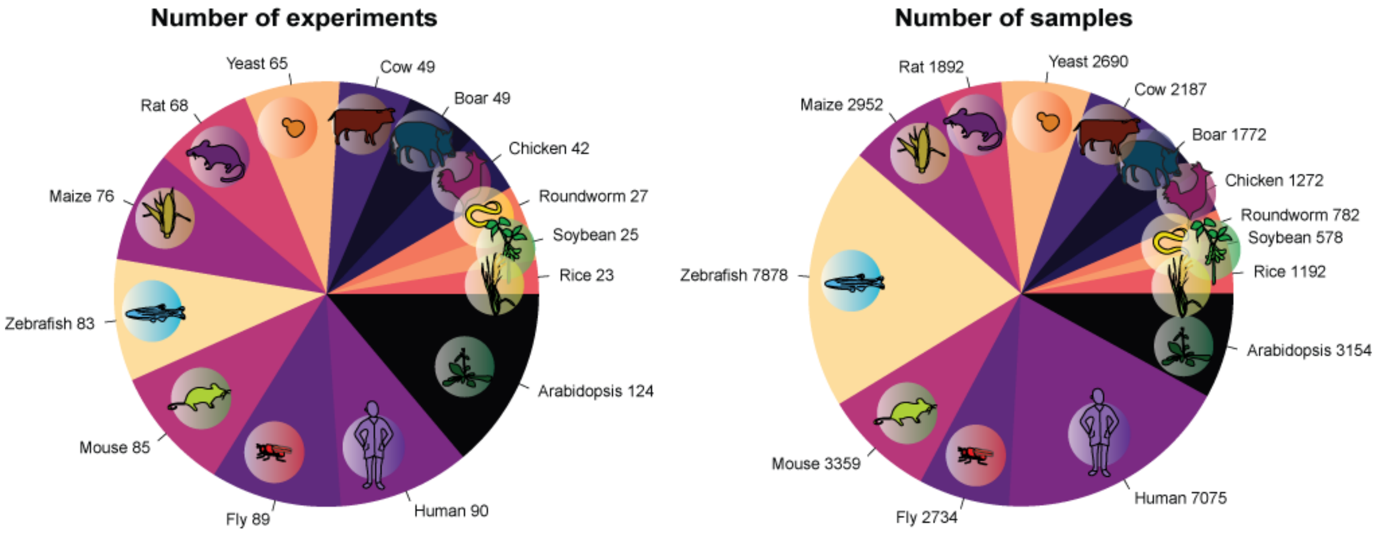
Implementation

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 above. 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 towards cluster centers and low degree nodes towards the cluster peripheries. Network data is stored in 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 about 200 genes will render a network within 30 seconds and implement EGAD for GO groups


Number of experiments

Number of samples

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

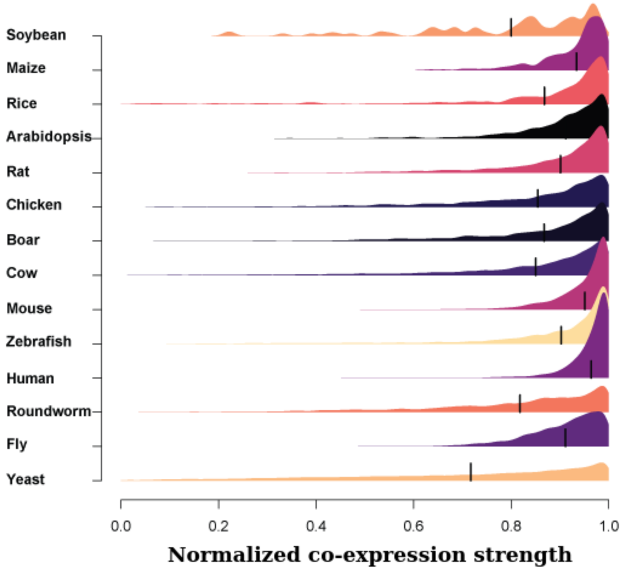


Figure 3. Distribution of co-expression values for ortholog mapped genes to the input of highly co-expressed yeast genes for each of the 13 other species.

within 10 seconds. An input of about 1000 genes will render a network within 5 minutes and implement EGAD for GO groups within a minute. Implementation of EGAD on the gene set takes between 30 seconds to 5 minutes 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 attached supplement(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 includes gene expression networks as HDF5 files, GO annotations, gene ID conversion tables, 1-to-1 ortholog mappings, the total degree of each gene, and example gene lists. During each step, the user is also able to download relevant data. In 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 supplement 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 above 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 figure S11 for a dendrogram and heatmap 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=0.9111, zebrafish=0.8770, fruitfly = 0.8320). A common feature of co-expression networks is hub genes which 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 (AUROCs $\approx$ 0.5). Together, these results indicate that yeast nucleolus genes form a functional module that is tightly conserved across distant species (figure 4).

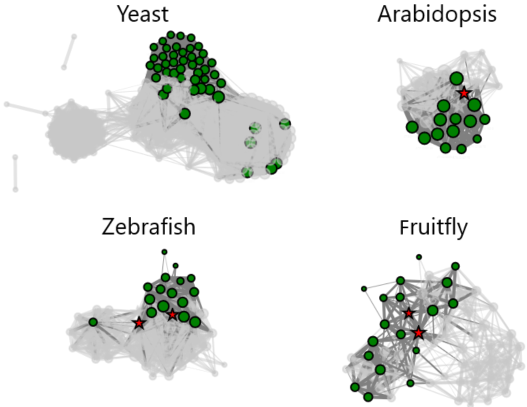


Figure 4. Highly co-expressed yeast (S. cerevisiae) genes are mapped to orthologous genes in Arabidopsis (A. thaliana), zebrafish (D. rerio), and fruitfly (D. melanogaster). Genes annotated with the nucleolus (GO:0005730) are highlighted, and the top 1% of connections are shown. Red stars denote highly 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, 44, 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 1-to-1 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 (AUROCs 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 (AUROCs of 0.880 in human and 0.859 in mouse), suggesting a shared mechanism that is conserved between human and mouse (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 in 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.

FUNDING

This work was supported by the National Institute of Health [R01 LM012736, R01 MH113005 to JL, MS, SB and JG and K99 MH120050 to MC].

Conflict of interest statement. None declared.

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.

REFERENCES

1. B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 4:Article17, 2005.
2. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A., 95(25):14863–14868, Dec 1998.
3. D. J. Allocco, I. S. Kohane, and A. J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics, 5:18, Feb 2004.
4. S. Ballouz, W. Verleyen, and J. Gillis. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics, 31(13):2123–2130, Jul 2015.
5. M. Crow, A. Paul, S. Ballouz, Z. J. Huang, and J. Gillis. Exploiting single-cell expression to characterize co-expression replicability. Genome Biol., 17:101, May 2016.
6. L. Mao, J. L. Van Hemert, S. Dash, and J. A. Dickerson. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics, 10:346, Oct 2009.
7. W. Liu, L. Lin, Z. Zhang, S. Liu, K. Gao, Y. Lv, H. Tao, and H. He. Gene co-expression network analysis identifies trait-related modules in Arabidopsis thaliana. Planta, 249(5):1487–1501, May 2019.
8. G. Monaco, S. van Dam, J. L. Casal Novo Ribeiro, A. Larbi, and J. P. de Magalhaes. 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, Nov 2015.
9. M. R. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics, 7:40, Mar 2006.
10. I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature, 474(7351):380–384, May 2011.
11. G. George, S. Singh, S. B. Lokappa, and J. Varkey. Gene co-expression network analysis for identifying genetic markers in Parkinson's disease - a three-way comparative approach. Genomics, 111(4):819–830, 07 2019.
12. F. E. Dewey, M. V. Perez, M. T. Wheeler, C. Watt, J. Spin, P. Langfelder, S. Horvath, S. Hannenhalli, T. P. Cappola, and E. A. Ashley. Gene coexpression network topology of cardiac development, hypertrophy, and failure. Circ Cardiovasc Genet, 4(1):26–35, Feb 2011.
13. A. A. Ammah, D. N. Do, N. Bissonnette, N. Gavry, and E. M. Ibeagha-Awemu. Co-Expression Network Analysis Identifies miRNA - mRNA Networks Potentially Regulating Milk Traits and Blood Metabolites. Int J Mol Sci, 19(9), Aug 2018.
14. Y. Liu, P. Lu, Y. Wang, B. E. Morrow, B. Zhou, and D. Zheng. Spatiotemporal Gene Coexpression and Regulation in Mouse Cardiomyocytes of Early Cardiac Morphogenesis. J Am Heart Assoc, 8(15):e012941, Aug 2019.
15. L. Pena-Castillo, M. Tasan, C. L. Myers, H. Lee, and et. al. Joshi, T. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol., 9 Suppl 1:S2, 2008.
16. Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, and et. al D'Andrea, D. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol., 17(1):184, 09 2016.
17. C. Ruprecht, N. Vaid, S. Proost, S. Persson, and M. Mutwil. Beyond Genomics: Studying Evolution with Gene Coexpression Networks. Trends Plant Sci., 22(4):298–307, 04 2017.
18. T. Obayashi, Y. Kagaya, Y. Aoki, S. Tadaka, and K. Kinoshita. 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(D1):D55–D62, Jan 2019.
19. T. Obayashi, Y. Aoki, S. Tadaka, Y. Kagaya, and K. Kinoshita. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of

the Statistical Property of the Mutual Rank Index. Plant Cell Physiol., 59(2):440, 02 2018.

20. S. van Dam, R. Cordeiro, T. Craig, J. van Dam, S. H. Wood, and J. P. de Magalhães. GeneFriends: an online co-expression analysis tool to identify novel gene targets for aging and complex diseases. BMC Genomics, 13:535, Oct 2012.
21. S. Proost and M. Mutwil. PlaNet: Comparative Co-Expression Network Analyses for Plants. Methods Mol. Biol., 1533:213–227, 2017.
22. E. Kim, S. Hwang, H. Kim, H. Shim, B. Kang, S. Yang, J. H. Shim, S. Y. Shin, E. M. Marcotte, and I. Lee. MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. Nucleic Acids Res., 44(D1):D848–854, Jan 2016.
23. M. Franz, H. Rodriguez, C. Lopes, K. Zuberi, J. Montojo, G. D. Bader, and Q. Morris. GeneMANIA update 2018. Nucleic Acids Res., 46(W1):W60–W64, 07 2018.
24. R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. Nucleic Acids Res., 39(Database issue):19–21, Jan 2011.
25. Y. Zhu, R. M. Stephens, P. S. Meltzer, and S. R. Davis. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics, 14:19, Jan 2013.
26. F. Cunningham, P. Achuthan, W. Akanni, J. Allen, and et. al. Amode, M. R. Ensembl 2019. Nucleic Acids Res., 47(D1):D745–D751, Jan 2019.
27. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1):15–21, Jan 2013.
28. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25(1):25–29, May 2000.
29. No authors listed. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res., 47(D1):D330–D338, Jan 2019.
30. S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16):3439–3440, Aug 2005.
31. E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simao, and E. M. Zdobnov. 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(D1):D807–D811, Jan 2019.
32. S. Oliver. Guilt-by-association goes global. Nature, 403(6770):601–603, Feb 2000.
33. S. Ballouz, M. Weber, P. Pavlidis, and J. Gillis. EGAD: ultra-fast functional analysis of gene networks. Bioinformatics, 33(4):612–614, 02 2017.
34. J. Gillis and P. Pavlidis. The impact of multifunctional genes on “guilt by association” analysis. PLoS ONE, 6(2):e17258, Feb 2011.
35. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. shiny: Web Application Framework for R, 2019. R package version 1.4.0.
36. Almende B.V., Benoit Thieurmel, and Titouan Robert. visNetwork: Network Visualization using 'vis.js' Library, 2019. R package version 2.0.9.
37. Josh Barnes and Piet Hut. A hierarchical O(N log N) force-calculation algorithm. Nature, 324(6096):446–449, Dec 1986.
38. Hadley Wickham. ggplot2: Elegant Gra phics for Data Analysis. Springer-Verlag New York, 2016.
39. Carson Sievert. plotly for R, 2018.
40. J. Seok, H. S. Warren, A. G. Cuenca, and et. al. Mindrinos, M. N. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc. Natl. Acad. Sci. U.S.A., 110(9):3507–3512, Feb 2013.
41. K. Takao and T. Miyakawa. Genomic responses in mouse models greatly mimic human inflammatory diseases. Proc. Natl. Acad. Sci. U.S.A., 112(4):1167–1172, Jan 2015.
42. I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov, 3(8):711–715, 08 2004.
43. H. Bruining, L. de Sonneville, H. Swaab, M. de Jonge, M. Kas, H. van Engeland, and J. Vorstman. Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. PLoS ONE, 5(5):e10887, May 2010.
44. J. Y. An and C. Claudianos. Genetic heterogeneity in autism: From single gene to a pathway perspective. Neurosci Biobehav Rev, 68:442–453, Sep 2016.
45. S. S. Jeste and D. H. Geschwind. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nat Rev Neurol, 10(2):74–81, Feb 2014.
46. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell, 180(3):568–584, Feb 2020.
47. S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, and et. al. Samocha, K. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature, 515(7526):209–215, Nov 2014.

FIGURE CAPTIONS

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.

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

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 (S. cerevisiae) genes are mapped to orthologous genes in Arabidopsis (A. thaliana), zebrafish (D. rerio), and fruitfly (D. 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.