UC San Diego
UC San Diego Previously Published Works

Title
Mapping the common gene networks that underlie related diseases

Permalink
https://escholarship.org/uc/item/9zg6006t

Journal
Nature Protocols, 18(6)

ISSN
1754-2189

Authors
Rosenthal, Sara Brin
Wright, Sarah N
Liu, Sophie
et al.

Publication Date
2023-06-01

DOI
10.1038/s41596-022-00797-1

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Nat Protoc. Author manuscript; available in PMC 2023 June 11.

Published in final edited form as:
Nat Protoc. 2023 June ; 18(6): 1745–1759. doi:10.1038/s41596-022-00797-1.

Mapping the common gene networks that underlie related diseases

Sara Brin Rosenthal[*,1,2], **Sarah N. Wright**[2,3], **Sophie Liu**[2], **Christopher Churas**[2], **Daisy Chilin-Fuentes**[1,2], **Chi-Hua Chen**[4], **Kathleen M. Fisch**[1], **Dexter Pratt**[2], **Jason F. Kreisberg**[2], **Trey Ideker**[*,2,3]

[1]Center for Computational Biology & Bioinformatics, Department of Medicine, University of California, San Diego, La Jolla, California, USA.

[2]Department of Medicine, University of California San Diego, La Jolla, CA, USA

[3]Program in Bioinformatics and Systems Biology, University of California San Diego, La Jolla, CA, USA

[4]Center for Multimodal Imaging and Genetics, Department of Radiology, University of California San Diego, La Jolla, California, United States of America.

Abstract

A longstanding goal of biomedicine is to understand how alterations in molecular and cellular networks give rise to the spectrum of human diseases. For diseases with shared etiology, understanding the common causes allows for improved diagnosis of each disease, development of new therapies and more comprehensive identification of disease genes. Accordingly, this protocol describes how to evaluate the extent to which two diseases, each characterized by a set of mapped genes, are colocalized in a reference gene interaction network. This procedure uses network propagation to measure the network "distance" between gene sets. For colocalized diseases, the network can be further analyzed to extract common gene communities at progressive granularities. In particular, we show how to: (i) obtain input gene sets and a reference gene interaction network, (ii) identify common subnetworks of genes that encompass or are in close proximity to all gene sets, (iii) use multiscale community detection to identify systems and pathways represented by each common subnetwork to generate a network colocalized systems map, (iv) validate identified genes and systems using a mouse variant database, and (v) visualize and further investigate

[*]Correspondence to: tideker@ucsd.edu, sbrosenthal@health.ucsd.edu.

Author contributions

SBR co-wrote the manuscript, performed analysis, and supervised the software development. SNW co-wrote the manuscript and developed the software.SL, CC, and DCF developed the software. KF contributed to methods development and project conceptualization. CC contributed to methods development and manuscript revision. DP and JFK co-wrote the manuscript. TI conceptualized the project and co-wrote the manuscript.

Conflicts of Interest

T.I. is co-founder of Data4Cure, Inc., is on the Scientific Advisory Board, and has an equity interest. T.I. is on the Scientific Advisory Board of Ideaya BioSciences, Inc., has an equity interest, and receives sponsored research funding. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies.

Code Availability

The NetColoc software is freely available in public repositories, under the MIT license (DOI: 10.5281/zenodo.6654561). NetColoc code and example notebooks are available on a GitHub repository https://github.com/ucsd-ccbb/NetColoc. The NetColoc code is also available on PyPi https://pypi.org/project/netcoloc/.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

select genes, interactions and systems for relevance to phenotype(s) of interest. We demonstrate the utility of this approach by identifying shared biological mechanisms underlying autism and congenital heart disease. However, this protocol is general and can be applied to any gene sets attributed to diseases or other phenotypes with suspected joint association. A typical NetColoc run takes less than an hour. Software and documentation are available at https://github.com/ucsd-ccbb/NetColoc.

EDITORIAL SUMMARY

This protocol describes how to use NetColoc, a freely available tool to evaluate the extent to which two related diseases, each characterized by a set of mapped genes, are colocalized in a reference gene interaction network.

Introduction

Many biological studies result in groups of genes linked to phenotypes of interest. For example, genome wide association studies (GWAS) identify common variations in DNA, which may be mapped to relevant genes, that are associated with particular diseases or traits[1]. As whole-exome and whole-genome sequencing studies are increasingly performed, genes containing rare variants associated with disease have also been identified[2]. Similarly, studies of mRNA expression levels, first using microarrays and now using next-generation sequencing, often result in sets of genes whose expression levels are altered in a disease or in response to perturbations[3].

Interpreting these gene sets can be a complicated and time consuming process, often involving extensive literature review to contextualize results with known biology. Functional enrichment analysis[4], one method for interpretation, can provide insight into the biological pathways and processes underlying a phenotype by testing for known pathways that have more in common with the gene set of interest than would be expected by chance. Gene interaction networks add further context for interpreting gene sets, with tools such as network propagation enabling the identification of new disease genes and genetic modules[5,6]. Notably, networks can boost the signal of underpowered data, as marginally significant variants converge on localized regions of network space[7,8].

To date, network analyses have focused on approaches for studying single gene sets. Here, we describe a protocol to study the extent to which two gene sets are related to each other, even if the gene sets themselves share few common genes. Previously, we have used this approach to identify a set of shared pathways underlying autism and congenital heart disease[9], an analysis we illustrate and extend here. Additionally, this approach may be used in the future to connect genes and pathways in a cross-species GWAS. More broadly, such an analysis can help unravel the complex relationships between genotypes and phenotypes by pinpointing convergent pathways.

Applications of the method

This protocol is based on NetColoc[9], a tool for evaluating the extent to which two gene sets are colocalized in a gene interaction network and for identifying the functions that underlie

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

this colocalization. NetColoc relies on a dual network propagation approach to identify the region of network space that is significantly near two distinct input gene sets. This tool can be used to study any pair of gene sets, such as rare and common variants within the same disease, genes associated with two comorbid diseases, genetically correlated GWAS phenotypes, GWAS across two different species, or gene expression changes after treatment with two different drugs.

Comparison with other methods

Direct comparisons of gene sets are often used to determine the similarity between phenotypes or conditions, usually with some statistical test to assign significance. This approach can be problematic when gene sets are small, as the power to detect a significant overlap is limited. In addition, variants in different genes within the same pathway may result in similar phenotypes, a finding that would not be recognized by direct comparison. A related approach might be to first perform functional enrichment analysis[4,10] on the individual gene sets and then to assess the overlap between significantly enriched pathways. This approach takes advantage of prior biological knowledge in the form of manually curated gene pathways. Like direct comparisons though, this method is limited in the case of small gene sets with few significantly enriched pathways. In contrast, NetColoc works well with small-to-medium gene sets (*i.e.*, between 5 and 500 genes) as it can readily identify the network space significantly proximal across sets.

Several previous methods have employed network information to interconnect sets of genes[7,11–14]. Many of these methods are designed to analyze single gene sets or specific biological contexts, such as patient stratification based on cancer mutations, or association of causal variants to changes in gene expression[7,11]. In contrast, NetColoc provides a general statistical framework for evaluating and interpreting the relationships in gene network space of any two gene sets. Furthermore, NetColoc implements a degree-corrected propagation algorithm, as diffusion or propagation methods may be susceptible to over-representation of hub genes[15]. It also includes a statistical metric to assess the significance of network colocalization and integrates with clustering tools, pathway analysis, and a mouse variant database. These features enable facile interpretation and validation of network colocalized genes.

Development of the protocol

The NetColoc methodology was originally implemented in a previous work[9], where we analyzed the network overlap between two comorbid disorders: autism spectrum disorder (ASD) and congenital heart disease (CHD). We measured the extent to which high confidence variants from large-scale exome studies were colocalized in network space, identifying novel disease genes and the underlying biological systems. Here we describe and further develop the protocol underlying this approach, with significantly greater details and updated methods. Specifically, we use an improved community detection algorithm[24] to identify the systems and pathways underlying the network colocalization. We also updated the ASD gene set to use the most recent list of high confidence ASD genes to date[28] in our example. Because of these updates, the exact genes and pathways identified here are slightly different than previously reported[9], although the major findings remain consistent. We have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

also implemented a distinct validation step by integrating with a mouse variant database[16] to enable broad use beyond ASD and CHD.

Overview of the protocol

This protocol consists of five major stages (Figure 1) and is accompanied by an open-source codebase and Jupyter notebook available at https://github.com/ucsd-ccbb/NetColoc detailing each step needed to reproduce the example below.

Obtain input gene sets and gene network (Step 1–3)—NetColoc uses two sets of genes associated with phenotypes of interest and a gene interaction network as inputs. It is most useful when analyzing gene sets obtained from systematic genome-wide experiments, such as those mapped from GWAS loci, damaging variants from exome sequencing studies, or genes differentially expressed between experimental conditions. These genome-wide results are in contrast to manually curated, low-throughput gene sets, which are subject to bias by focusing on well-known genes. By using unbiased gene lists as inputs, poorly studied but physiologically important disease genes can be revealed based on their proximity to well-known disease genes, thereby facilitating the discovery of novel disease-related pathways.

In practice, we find that small to medium gene sets (between 5 and 500 genes) work best as inputs to NetColoc. If the input gene sets are very small, they may not characterize the phenotypes generally enough to yield meaningful results, i.e. characterizing the network proximal region to a single gene does not necessarily provide insight to the underlying phenotype. Alternatively, very large input gene sets result in the majority of the network being identified after network propagation, resulting in a lack of specificity. The sampling underlying the statistical framework also becomes an issue, due to the finite network (Supplementary Methods, Supplementary Fig. 1).

There exist a large number of gene interaction networks, some of which integrate numerous data types and databases[17–21]. Recent work has demonstrated that larger, more inclusive networks outperform smaller networks in disease gene discovery[22]. As such we recommend using a large and inclusive network for NetColoc analysis such as STRING[17] or PCNet[22].

Identification of a subnetwork of colocalized genes (Step 4–8)—NetColoc creates a network of colocalized genes, which includes some genes from both input sets alongside other genes identified by dual network propagation. Input genes identified in the colocalization network are generally genes that have strong evidence for association to at least one phenotype of interest but may be novel to the other. Genes found in both input gene sets are very likely to be included in the colocalization network. Notably, the colocalization network also includes genes that are identified by joint network proximity to the input gene sets but which are not themselves input genes. These genes represent candidate risk genes novel to both phenotypes.

Compute network colocalized systems map (Step 9–14)—The colocalization network generated by the previous step may be large, in which case it may be useful to separate this network into distinct communities using multiscale community detection.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Multiscale community detection identifies highly interconnected (modular) systems of genes, which can represent distinct protein complexes or biological pathways[23,24]. Identification of communities at multiple scales yields a hierarchical structure of discrete systems, with smaller, more specific systems contained within larger, more general ones[25]. The resulting NetColoc systems map provides a high-level view of the shared biological processes between phenotypes. Here, each system is essentially a hypothesis that those genes and interactions within it describe a pathway, process or complex that underlies the shared biology of both phenotypes. Because the resulting systems derive from the structure of the interaction network, they may be novel and context specific as they do not rely on a manual classification process.

Validate identified genes and systems (Step 15–24)—While the NetColoc systems map may recapitulate known biology – systems with known associations to both phenotypes – it may also present novel systems and system-phenotype associations. To prioritize such hypotheses, a key step is to integrate the systems map with independent datasets. For studies in mammals, one powerful independent resource is the Mouse Genome Informatics (MGI) database[16], a large catalog of genes that, when disrupted, cause specific phenotypes in mice. Such analysis serves to pinpoint conserved systems enriched for mouse phenotypes and to further nominate novel disease gene candidates within these systems for follow-up experiments. One limitation of such an approach is that not all systems are conserved across species. In such cases, a novel system would not be validated by integration with the MGI, and further investigation would be needed to verify the system was not a false positive.

Further exploration of select systems (Step 25–30)—Systems of interest and the genes and interactions contained therein can be further investigated by importing the NetColoc systems map into Cytoscape[26]. These investigations may proceed in a number of directions. For example, systems without annotations may represent novel pathways, processes or complexes. Genes within systems of interest that are also associated with a particular phenotype in mice may be good candidates for further studies. Further, the interactions between novel disease gene candidates and known disease genes may be examined for further insights into functions related to the disease.

Limitations:

Limitations of NetColoc include the requirement that the input gene sets be of moderate size (approximately 5–500 genes, see above). Additionally, it can be computationally challenging and even practically prohibitive to perform network propagation on very dense gene interaction networks (>3 million edges). Currently, NetColoc is designed to operate on a single pair of gene sets, representing a pair of phenotypes. In future work we may allow for three or more input gene sets.

Materials

Hardware

A computer or server with 32 GB RAM, running Python 3 (may run with less memory, depending on size of network). The workflow has been tested on MacOS 10.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Software

- Python packages: click, matplotlib, ndex2, network, numpy, seaborn, tqdm, mygene, scipy, statsmodels, gprofiler-official, ipywidgets, ipycytoscape, ddot, cdapsutil.
- Cytoscape, version 3.9 or later (https://cytoscape.org/)

Example data:

- Text files containing input gene lists: CHD_HC.tsv, and Satterstrom--Top-102-ASD-genes--May2019.csv. Included in the NetColoc GitHub repository
- Text file containing input gene list for scored gene list example: E-MTAB-6863-query-results.tsv. Included in the NetColoc GitHub repository

CRITICAL Input data should be text files, with one column containing the names of the input genes (there may be other columns which are not used). Input gene lists should be between 5 and 500 genes. There may optionally be a column for a per-gene score (a p-value or log fold change, for example), used in some optional parts of the workflow.

Software setup

- Python 3 installation (https://www.python.org/)
- Jupyter notebook installation (https://jupyter.org/install)
- Cytoscape installation (https://cytoscape.org/)
- NetColoc installation and dependencies (click, matplotlib, ndex2, network, numpy, seaborn, tqdm, mygene, scipy, statsmodels, gprofiler-official, ipywidgets, ipycytoscape, DDOT, cdapsutil). NetColoc and all dependencies except DDOT and cdapsutil will be automatically installed with `pip install netcoloc`. Cdapsutil can be installed with `pip install cdapsutil`. DDOT can be installed by cloning the repository, using the following commands

```
git clone --branch python3 https://github.com/idekerlab/ddot.git
cd ddot
python setup.py bdist_wheel
pip install dist/ddot*py3*whl
```

Procedure

CRITICAL The python code required for Steps 1–24 is included in the supplementary procedure to improve readability. The remaining steps 25–30 should be carried out in NDEx and Cytoscape. Additionally, steps 1–24 are demonstrated in an example notebook: https://github.com/ucsd-ccbb/NetColoc/blob/main/example_notebooks/ASD_CHD_NetColoc_analysis.ipynb.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Obtain input gene sets and gene network (<5 minutes)

1. Load required packages into Python.

   ?Troubleshooting

2. Select two gene sets of interest. Load gene sets from text files into Python. These gene sets should contain between 5 and 500 genes and come from experimental data, rather than manual curation, to avoid bias.

   In some use cases, the gene sets of interest may accompany a score (such as p-value, or log fold change, in an RNA-Seq differential expression experiment). For these use cases, we provide an optional step to aid the researcher in finding an optimal choice of threshold by sweeping over a range of filtering criteria to maximize the observed divided by expected network intersection size. The genes which meet these criteria should be retained for use in the following steps. This process is illustrated in an example notebook https://github.com/ucsd-ccbb/NetColoc/blob/main/example_notebooks/Evalute_scored_input_gene_lists.ipynb.

3. Select a gene interaction network to use for the analysis. Identify the network UUID (Universally Unique IDentifier) on NDEx[27] and use this to import to a Jupyter notebook. We recommend using PCNet[22] as a starting point, but a user may want to switch to "STRING high confidence" if using a machine with low memory (< 8GB RAM). CRITICAL Verify that nomenclature for the input genes matches the nomenclature for the interaction network (e.g., both are from the same species, and both use Entrez ID or HGNC symbol)

   ?Troubleshooting

Identify subnetworks of colocalized genes (20 minutes)

4. Precalculate matrices needed for network propagation, using the functions `netprop.get_normalized_adjacency_matrix` and `netprop.get_individual_heats_matrix`, referred to as w' and w" in the following. This step will take a few minutes (more for denser networks). A benchmarking analysis demonstrates that the runtime required scales with the number of edges (w') and the number of nodes (w") (Supplementary Fig. 2a,b). If the researcher plans on running multiple analyses they may find it useful to save these matrices as numpy binary files. We include instructions for saving and reloading. We caution that, because these matrices are not sparse, saving and reloading can take a few minutes, and the saved file can be a few GB, so for many networks it may be faster to recompute the matrices each time. The diffusion parameter, which controls the rate of propagation through the network, may be set in this step. In practice, we have found that results are not dependent on the choice of this parameter (Supplementary Fig. 3), and recommend using the default value of 0.5.

   ?Troubleshooting

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Subset input genes sets to genes found in the selected network. Only genes contained in the interaction network will be retained as "seed" genes for downstream analysis.

   ?Troubleshooting

6. Compute network proximity scores from both seed gene sets, z_1 and z_2, independently, using the function `netprop_zscore.calculate_heat_zscores`. The network proximity scores include a correction for the degree distribution of the input gene sets (Supplementary Fig. 4). The runtime required for computing the network proximity scores increases linearly with the number of nodes in the underlying interaction network and with the size of the input gene list (Supplementary Fig. 2c).

7. Build the NetColoc subnetwork and evaluate it for significant network colocalization. To build the NetColoc subnetwork, we take the product of the two proximity vectors as follows:

$$z_{coloc} = z_1 * z_2$$

   We then select genes with z_{coloc} greater than a threshold ($z_{coloc} > 3$ default, but can be set by the user), and network proximity scores individually larger than a nominal threshold ($z_1 > 1.5$, and $z_2 > 1.5$ default, but can be set by the user). The genes meeting these criteria and associated interactions make up the network colocalization subnetwork. We have found that the default threshold values work well in practice to find the set of genes that is proximal to both seed gene sets. Tuning the thresholds higher will lead to fewer false positives but more false negatives. Similarly, tuning them lower will lead to more false positives but fewer false negatives. Either may be warranted given the specifics of an experiment. The researcher may conduct a sensitivity analysis of these thresholds to find a balance between a higher NetColoc enrichment score, but smaller network, and a lower NetColoc enrichment score, but larger network (Supplementary Fig. 5). The function `network_colocalization.calculate_network_enrichment` is provided to enable such a sensitivity analysis. In this function, the network colocalization score is computed for the gene set pair, based on the observed network overlap and expected network overlap from a null distribution, over a range of z-score thresholds. We recommend using the default thresholds unless the use case calls for higher or lower stringency. Choosing the thresholds which optimize the network colocalization score risks leaving out important phenotype-related genes. If gene sets are significantly colocalized, proceed with the analysis. Gene sets that are not significantly colocalized in the network have no evidence for shared underlying pathways, and thus proceeding with an analysis of the network intersection in this case is not likely to return meaningful results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. (OPTIONAL) Transform NetColoc subnetwork edges to cosine similarities with the function `network_colocalization.transform_edges`. The cosine similarity score between two genes represents the extent to which those genes have similar interactors. In practice, the cosine similarity transformed score helps to visually reveal the underlying clustering structure present in a network.

Compute network colocalized systems map (5 minutes)

9. Convert network colocalization subnetwork from network graph format to ndex graph format, for compatibility with community detection module.

10. Run community detection on the NetColoc subnetwork to identify highly interacting subsystems. We recommend using the HiDef clustering algorithm[24] which is included in the NetColoc dependency cdapsutil which performs community detection, along with other commonly used clustering algorithms.

11. Convert the NetColoc hierarchy to networkx format, and write out features of the hierarchy to a pandas dataframe, for easier manipulation in Python.

12. (OPTIONAL) Systems that do not contain any seed genes may be removed in order to focus on systems in which perturbations are known to have an effect.

    ?Troubleshooting

13. (OPTIONAL) Examine the structure of the NetColoc hierarchy with an interactive sneak peak within the Jupyter notebook. Full annotation and visualization are conducted later in the analysis pipeline, but the researcher may find it helpful to get a sense of the size and structure of the NetColoc hierarchy.

    ?Troubleshooting

14. Annotate systems with gprofiler[10], a functional enrichment tool. Annotate moderately sized systems (between 50 to 1000 genes per system) if the systems are significantly enriched for a Gene Ontology (GO) biological process. To increase the stringency of the annotation, require that the GO term is enriched with $p<1\times10^{-5}$ and shares at least 3 genes with the system. Label the system using the GO term that meets these criteria and has the highest sum of precision and recall. Systems without a GO term meeting these criteria are labeled with their unique system ID.

Validate identified genes and systems (15 minutes)

15. Load and parse mouse variant database, using functionality in the validation module included with NetColoc.

16. Identify phenotype(s) of interest. We recommend including a negative control, a phenotype that is not expected to overlap with the two phenotypes of interest.

17. Compute the enrichment of selected phenotype(s) in the NetColoc subnetwork as a whole to identify the phenotypes with the strongest association with the full NetColoc subnetwork. By computing the enrichment in the entire NetColoc

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

subnetwork, we identify the phenotypes with the strongest association with the entire set of genes identified to be related to both input sets.

18. Compute the enrichment of phenotype(s) in NetColoc subsystems. Some phenotypes may have stronger associations with NetColoc subsystems than with the full subnetwork. In this step, we calculate the enrichment of selected phenotypes in each NetColoc subsystem.

19. Annotate the NetColoc systems map with mouse variant data, input genes and enriched GO terms.

20. Export the NetColoc systems map to NDEx with the default style. Default style maps the fraction of seed genes from input set 1 (red) and input set 2 (blue) to node pie charts in NDEx. The remaining white fraction indicates the fraction of genes in each system that are not in either input set but that are implicated by network propagation (Figure 2a,b).

21. Apply another template style to the NetColoc Systems Map for mouse variant view and export to NDEx. Select the property to be mapped to system node colors (should be one of the mouse variant phenotypes previously identified). In this style, the log odds ratio is mapped to the system node color. Systems that are not significantly enriched for the phenotype are white ($p<0.05$; Figure 2a,c).

?Troubleshooting

22. Add genes associated with mouse variant phenotypes to the NetColoc subnetwork and export to NDEx.

?Troubleshooting

23. Upload the cosine-similarity transformed NetColoc subnetwork to NDEx.

24. Add the four networks from above to a network set on NDEx, using ‘add_networks_to_networkset’ function from the NetColoc dependency ndex2, with the UUIDs defined for each individual network

Further exploration of select systems (10 minutes for automated steps, but manual investigation piece is more time consuming- a researcher may spend days fine-tuning visualization and researching genes and systems in the networks)

25. Import the four networks from the network set on NDEx (step 24) to Cytoscape. Navigate to the network set on the NDEx account page (Figure 2a). Open each network in a new tab and click “open in Cytoscape” (Figure 2b,c,d).

26. Apply ‘yfiles organic’ layout to NetColoc subnetwork with network edges, and NetColoc subnetwork with cosine similarity edges.

?Troubleshooting

27. Apply ‘yfiles tree’ layout to the NetColoc systems map. Apply copycat layout to NetColoc systems with the mouse variant view from step 21, to ensure both systems maps have identical layouts.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

?Troubleshooting

28. (OPTIONAL) Fine-tune the layout and visualization. Some options include: a) manually adjusting positions of genes/systems so labels are legible, b) modifying color schemes, c) setting non-seed gene label transparency to 0 for large networks to improve legibility.

?Troubleshooting

29. If not installed already, install the 'Community Detection' app on the Cytoscape app store. Analyze systems of interest. Right click on a system of interest from the NetColoc systems map. Scroll to "Apps," then "Community Detection," then click "View interactions for selected node." This will bring up a prompt to select a network for which to view the interactions between genes in the selected system. Select either the NetColoc subnetwork with network edges or the NetColoc subnetwork with cosine similarity edges. A new network will be created consisting of the genes and interactions in the selected system.

30. Further analyze a system of interest by selecting genes causing a phenotype of interest when knocked out in mice. These genes are available in the node table view.

Timing

A typical run through the NetColoc workflow (steps 1–25) takes 10–60 minutes, to run on a 32 GB-RAM machine with an i7 processor, depending on the number of nodes and edges of the selected network, and depending on size of input gene sets. The workflow has been tested on networks of up to ~40 million edges. Runtime for larger or denser networks may be prohibitive. Timing for exploration and interpretation of results (steps 26–31) depends on the researcher and the nature of the scientific questions.

Troubleshooting

Troubleshooting advice can be found in Table 1.

Anticipated Results

Use case: Network colocalization of two comorbid disorders.

An apt demonstration of the operation and utility of the NetColoc workflow is its recent use in a published analysis of gene sets associated with two comorbid diseases, ASD and CHD[9]. NetColoc shows that these two disorders have a significant shared component, defined as the size of the observed colocalization subnetwork divided by the expected size of such a subnetwork given randomly selected genes. Such a shared component demonstrates that the two diseases impact common pathways, despite having largely distinct gene sets.

Here we recapitulate and expand the main results from the published paper, using the most recent list of high confidence ASD genes[28] and an updated workflow. In particular, we performed a revised NetColoc analysis using 102 genes associated with ASD variants[28], 66

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genes associated with CHD variants[29], and the PCNet human gene interaction network[22]. We identified 773 genes in the common subnetwork, compared to 257 expected from the null model ($p < 1\times10^{-100}$ by permutation test; Figure 3a,b). A sensitivity analysis of a range of z-score thresholds revealed that more stringent z-score thresholds resulted in a higher NetColoc enrichment score, but a lower number of genes identified (Supplementary Fig. 5). Here we accepted a slightly lower NetColoc enrichment score in favor of a larger pool of possible new disease genes.

Application of the HiDef multiscale community detection algorithm[23,24] revealed 81 highly interconnected gene systems, with 42 containing one or more seed genes associated with ASD or CHD (Figure 3c). These systems were then functionally annotated with known biological pathways and processes. Prominent annotations included pathways related to ion channels, chromatin modification, and histone modification. These pathways have known relevance to both ASD and CHD[30–33].

In the above procedure, we have demonstrated how to investigate the genes and interactions in a subsystem of interest (steps 29 and 31, see Figure 4). The researcher can examine how the input genes are connected to other input genes and to other network-implicated genes in the subsystem. In the case of ASD and CHD, such network-implicated genes may represent new disease risk genes.

Gene systems validated in mouse variant models.

Mapping additional data onto the NetColoc systems map can provide useful evidence to support (or counter) hypotheses from particular gene systems. The premise of this analysis is that the genes in the same network neighborhood are more likely to have roles in the same phenotype of interest, and that other data types may provide complementary evidence for those roles.

Accordingly, we integrated the ASD-CHD NetColoc systems map with a database of mouse gene disruptions linked to resulting specific phenotypes[16]. This analysis indicated that the genes in the ASD-CHD NetColoc subnetwork were significantly enriched for genes that, when disrupted in mice, lead to both abnormal brain and heart phenotypes (Figure 4a). A negative control phenotype – abnormal innate immunity – was not similarly enriched. The enrichment was even more pronounced within individual gene systems, with some systems having 12-fold enrichment of abnormal neuron morphology genes (Figure 4b). Furthermore, many genes that were identified by network proximity (*i.e.*, not in the input sets) were important for normal neuronal morphology in mice. For example, in a system annotated for L1CAM interactions, ankyrin and sodium channel genes ANK2, SCN1A and SCN2A were ASD seed genes, and ANK3 was a CHD seed gene. Of these, individual disruptions of ANK2, SCN1A, and ANK3 resulted in abnormal neuron morphology in mice. Notably, of the other network-implicated genes in this L1CAM interactions system, seven genes – SCN1B, SPTBN2, NFASC, SPTBN4, NRCAM, CLCN2 and CHRNB1 – also demonstrated abnormal neuronal morphology when disrupted in mice, suggesting that the pathway as a whole plays an important role in the nervous system (Figure 4c,d).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Benchmarking on Gene Ontology (GO)

We benchmarked NetColoc on two branches of GO [34], to establish a baseline for network colocalization scores in a controlled setting. Here we expect gene set pairs to be closely related if they are nearby in the ontology and less related if they are distant in the ontology. Specifically, we examined the lipid metabolic process branch (GO:0006629), containing 1,413 genes and 950 terms, and the cell development branch (GO:0048468), containing 2,085 genes and 867 terms. We selected pairs of gene sets expected to be highly related but distinct (*i.e.*, non-overlapping gene sets within the same term), pairs of gene sets expected to be somewhat related (*i.e.*, genes selected from different terms in the same branch) and pairs of gene sets expected to be unrelated (*i.e.*, genes selected from different terms from different branches). We subjected each of these gene set pairs to our network colocalization procedure using the PCNet network[22] and measured the network colocalization scores (Figure 5a). As expected, the within-term gene sets had the highest network colocalization, and the between-term gene sets from the same branch had intermediate network colocalization values. Gene sets chosen from different branches had network colocalization values indistinguishable from the baseline.

We also examined how the network colocalization degrades with increasingly noisy input data. We selected disjoint gene set pairs from within the same term and measured the network colocalization as a selected fraction of the genes in each set were replaced with randomly selected genes. As expected, the network colocalization decreased with increasing noise (Figure 5b), eventually reaching a baseline value when all genes from the input sets had been replaced with randomly selected genes. Notably, although the network colocalization decreased significantly, the procedure could still detect a significant network colocalization even with 80% random genes, illustrating the resilience of NetColoc to noisy input gene sets.

Conclusion

The network colocalization protocol presented here provides a quantifiable and reproducible workflow for probing the extent to which two gene sets impact similar biological processes and pathways. It presents a roadmap for prioritizing genes and pathways at the intersection of two diseases or phenotypes, and for the discovery of disease genes that may be missed by sequencing studies. As an example, consider GWAS, which detects common variants of low effect size, and exome sequencing studies, which detect rare variants with high effect sizes. Moderately rare variants with moderate effect sizes may then be missed by both of these efforts but could be identified by NetColoc. Furthermore, cohorts relying on the detection of *de novo* variants – variants present in a child but not in either parent – would miss alleles with recessive patterns of inheritance. These recessive variants may be identified with NetColoc, as the approach imposes no restrictions on variant type.

A natural extension will be to generalize the workflow to more than two input gene sets. For example, many neurological disorders have significant comorbidities, where a patient with one disorder is more likely to develop another[35,36]. By systematic application of a generalized network colocalization workflow, we may be able to stratify the features shared among many comorbid disorders, as well as those that are specific to one disorder. Such

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a generalization could be used to disentangle the complex relationships between genotypes and phenotypes more broadly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the following grants from the National Institutes of Health: U24 CA184427 to D.P., R50 CA243885 to J.F.K. and U01 MH115747, R01 HG009979, P50 DA037844 and P41 GM103504 to T.I. This research was partially supported by the Altman Clinical & Translational Research Institute (ACTRI) at the University of California, San Diego. The ACTRI is funded from awards issued by the National Center for Advancing Translational Sciences, NIH UL1TR001442.

Data Availability

The input gene lists used for illustration of the protocol may be found in the supplementary materials of two papers. The ASD input gene lists were acquired from Satterstrom et al 2020[28]. The CHD input gene lists were acquired from Jin et al. 2017 [29]. The differential expression data used for illustration of the scored input gene list alternate step were acquired from the EBI expression atlas (https://www.ebi.ac.uk/gxa/home), from Ramnath et al. 2018 [37]. The molecular interaction networks used in this workflow were acquired from the network data exchange (ndexbio.org); PCNet[22] UUID 4de852d9-9908-11e9-bcaf-0ac135e8bacf, STRING[17] UUID 275bd84e-3d18-11e8-a935-0ac135e8bacf.

References

1. Tam V et al. Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467–484 (2019). [PubMed: 31068683]
2. Cirulli ET & Goldstein DB Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11, 415–425 (2010). [PubMed: 20479773]
3. Stark R, Grzelak M & Hadfield J RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656 (2019). [PubMed: 31341269]
4. Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550 (2005). [PubMed: 16199517]
5. Cowen L, Ideker T, Raphael BJ & Sharan R Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. 18, 551–562 (2017). [PubMed: 28607512]
6. Vanunu O, Magger O, Ruppin E, Shlomi T & Sharan R Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6, e1000641 (2010). [PubMed: 20090828]
7. Hofree M, Shen JP, Carter H, Gross A & Ideker T Network-based stratification of tumor mutations. Nat. Methods 10, 1108–1115 (2013). [PubMed: 24037242]
8. Leiserson MDM et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat. Genet. 47, 106–114 (2015). [PubMed: 25501392]
9. Rosenthal SB et al. A convergent molecular network underlying autism and congenital heart disease. Cell Systems (2021) doi:10.1016/j.cels.2021.07.009.
10. Raudvere U et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 47, W191–W198 (2019). [PubMed: 31066453]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

11. Paull EO et al. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics 29, 2757–2764 (2013). [PubMed: 23986566]
12. Jia P & Zhao Z VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. PLoS Comput. Biol. 10, e1003460 (2014). [PubMed: 24516372]
13. Ruffalo M, Koyutürk M & Sharan R Network-Based Integration of Disparate Omic Data To Identify 'Silent Players' in Cancer. PLOS Computational Biology vol. 11 e1004595 (2015). [PubMed: 26683094]
14. Tuncbag N et al. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLOS Computational Biology vol. 12 e1004879 (2016). [PubMed: 27096930]
15. Erten S, Bebek G, Ewing RM & Koyutürk M DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. BioData Min. 4, 19 (2011). [PubMed: 21699738]
16. Eppig JT et al. Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. Methods Mol. Biol. 1488, 47–73 (2017). [PubMed: 27933520]
17. Szklarczyk D et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2018).
18. Breitkreutz B-J et al. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 36, D637–40 (2008). [PubMed: 18000002]
19. Blom UM, Wang PI, Shim JE & Marcotte EM Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome (2011).
20. Greene CS et al. Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. 47, 569–576 (2015). [PubMed: 25915600]
21. Hermjakob H IntAct: an open source molecular interaction database. Nucleic Acids Research vol. 32 452D–455 (2004).
22. Huang JK et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. Cell Syst 6, 484–495.e5 (2018). [PubMed: 29605183]
23. Singhal A et al. Multiscale community detection in Cytoscape. PLoS Comput. Biol. 16, e1008239 (2020). [PubMed: 33095781]
24. Zheng F et al. HiDeF: identifying persistent structures in multiscale 'omics data. Genome Biology vol. 22 (2021).
25. Simon HA The Architecture of Complexity. Proc. Am. Philos. Soc. 106, 467–482 (1962).
26. Shannon P et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003). [PubMed: 14597658]
27. Pratt D et al. NDEx, the Network Data Exchange. Cell Syst 1, 302–305 (2015). [PubMed: 26594663]
28. Satterstrom FK et al. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180, 568–584.e23 (2020). [PubMed: 31981491]
29. Jin SC et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat. Genet. 49, 1593–1601 (2017). [PubMed: 28991257]
30. Zaidi S & Brueckner M Genetics and Genomics of Congenital Heart Disease. Circ. Res. 120, 923–940 (2017). [PubMed: 28302740]
31. Lasalle JM Autism genes keep turning up chromatin. OA Autism 1, 14 (2013). [PubMed: 24404383]
32. Ackerman MJ The long QT syndrome: ion channel diseases of the heart. Mayo Clin. Proc. 73, 250–269 (1998). [PubMed: 9511785]
33. Colbert CM & Pan E Ion channel properties underlying axonal action potential initiation in pyramidal neurons. Nat. Neurosci. 5, 533–538 (2002). [PubMed: 11992119]
34. Ashburner M et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000). [PubMed: 10802651]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

35. Hesdorffer DC Comorbidity between neurological illness and psychiatric disorders. CNS Spectr. 21, 230–238 (2016). [PubMed: 26898322]

36. Willsey AJ et al. The Psychiatric Cell Map Initiative: A Convergent Systems Biological Approach to Illuminating Key Molecular Pathways in Neuropsychiatric Disorders. Cell 174, 505–520 (2018). [PubMed: 30053424]

37. Ramnath D et al. Hepatic expression profiling identifies steatosis-independent and steatosis-driven advanced fibrosis genes. JCI Insight vol. 3 (2018).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

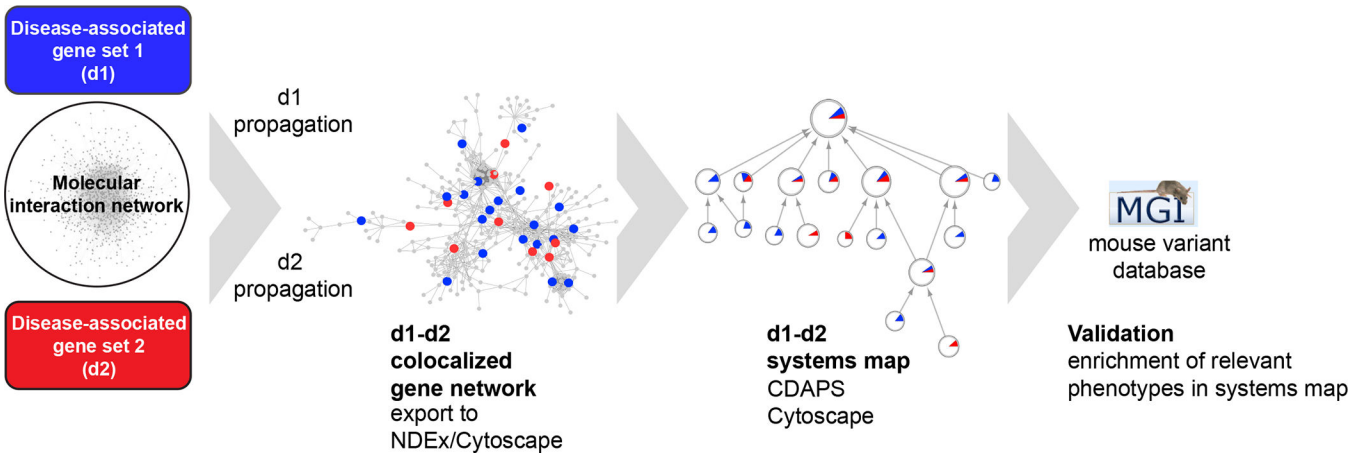

Figure 1: Workflow of the protocol.

Two disease-associated gene sets (d1 and d2), along with a selected molecular interaction network, are the inputs to the workflow. Network propagation is conducted from each gene set individually on the selected interaction network. The resulting propagation scores are combined to create the d1-d2 colocalized gene network. Application of a hierarchical cluster algorithm results in the d1-d2 systems map. The genes and systems in the d1-d2 systems map are then interrogated with the Mouse Genome Informatics (MGI) database, for enrichment of relevant phenotypes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

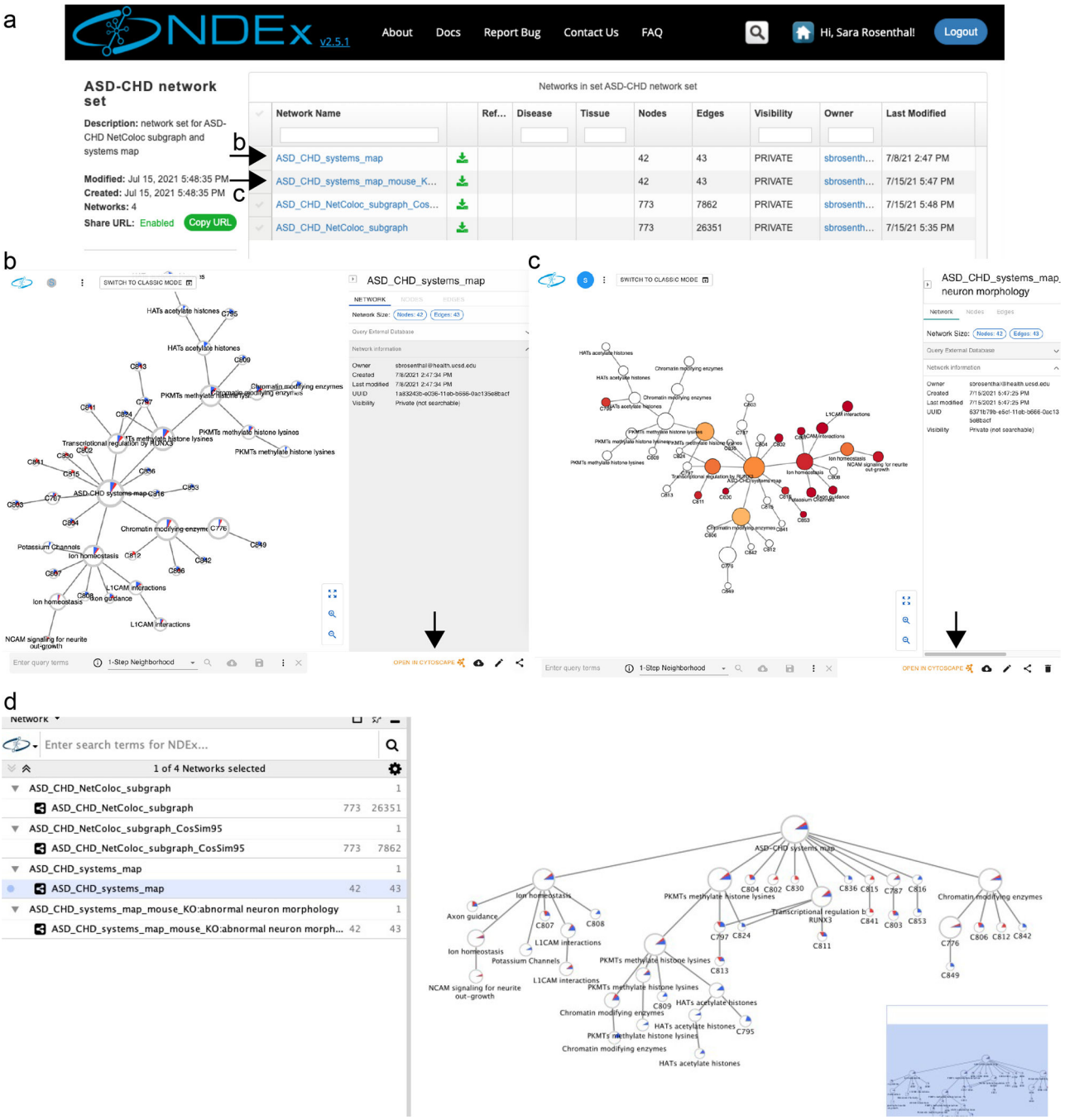

Figure 2: Exploration of NetColoc systems map.

a) Network set view of 4 output networks from steps 1–25. b) NDEx view of the NetColoc systems map with default view. Arrow indicates a button to open in Cytoscape. c) NDEx view of the NetColoc systems map with node colors and sizes mapped according to the mouse variant view, where natural log of the odds ratio is indicated by node fill color and systems not significantly enriched (p>0.05) are indicated with white nodes. Arrow indicates a button to open in Cytoscape. d) Cytoscape view of 4 output networks (left), and the NetColoc systems map, after applying the y-files tree layout algorithm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

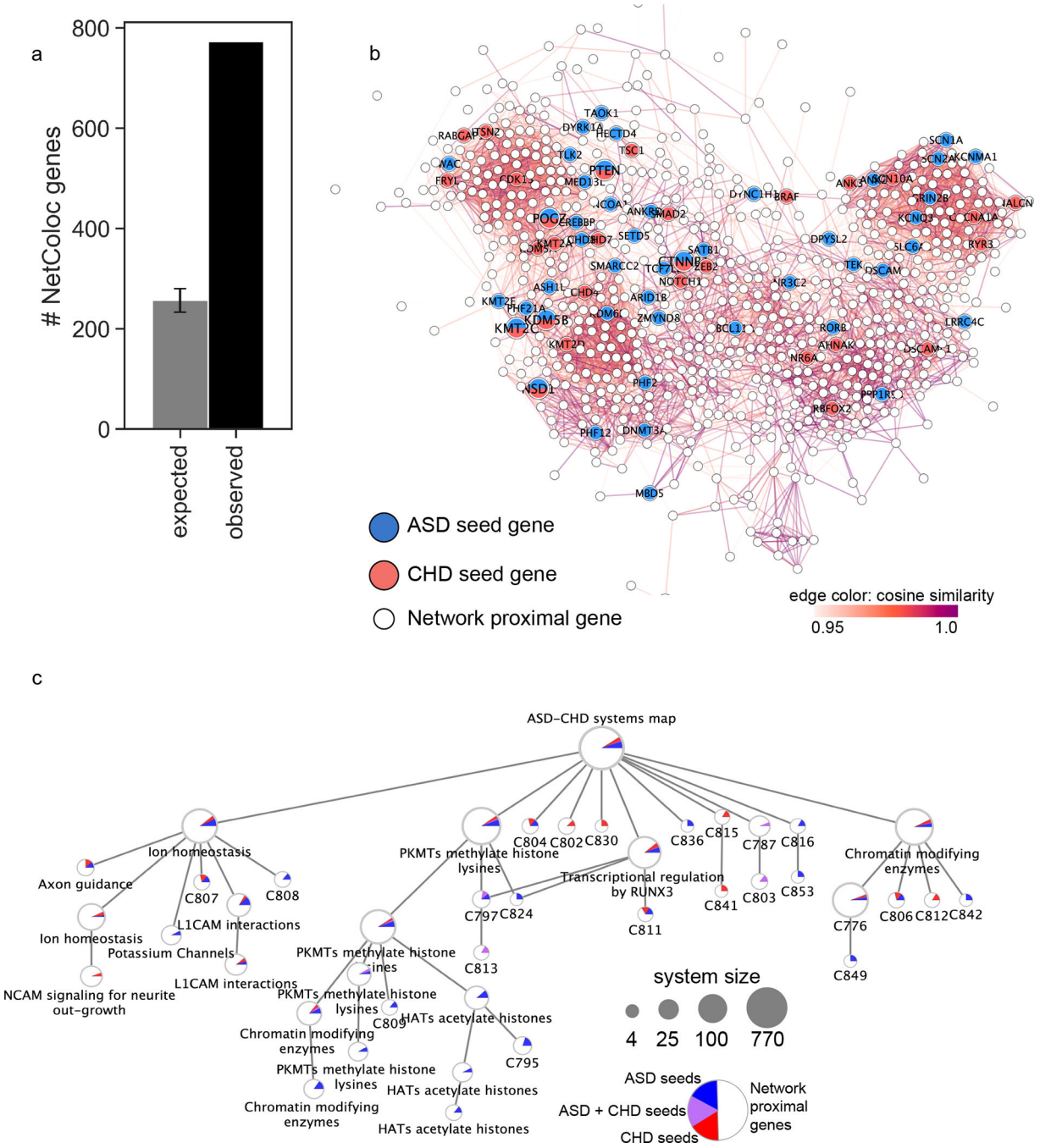

Figure 3: Network colocalization of ASD and CHD.

a) NetColoc enrichment measured between ASD and CHD. Black bar shows the number of genes observed in the NetColoc subnetwork. Gray bar shows the number of genes expected in the null model. Error bars show 95% confidence intervals. b) Visualization of ASD-CHD NetColoc subnetwork. ASD seed (input) genes indicated as blue nodes, and CHD seed (input) genes indicated as red nodes. Genes that are seeds for both ASD and CHD are indicated with half red and half blue nodes. Genes in the NetColoc subnetwork implicated by network proximity are indicated with smaller white nodes. Gene positions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are determined by a spring-embedded layout in Cytoscape. Edges shown here are cosine similarities. c) Hierarchical map of systems in the ASD-CHD NetColoc subnetwork. Child systems are contained within parent systems. Pie charts indicate the fraction of genes per system belonging to ASD seeds (blue), CHD seeds (red), or to both ASD and CHD seeds (purple), versus the remaining genes implicated by network proximity (white). Systems were labeled with significantly enriched biological pathways where possible.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

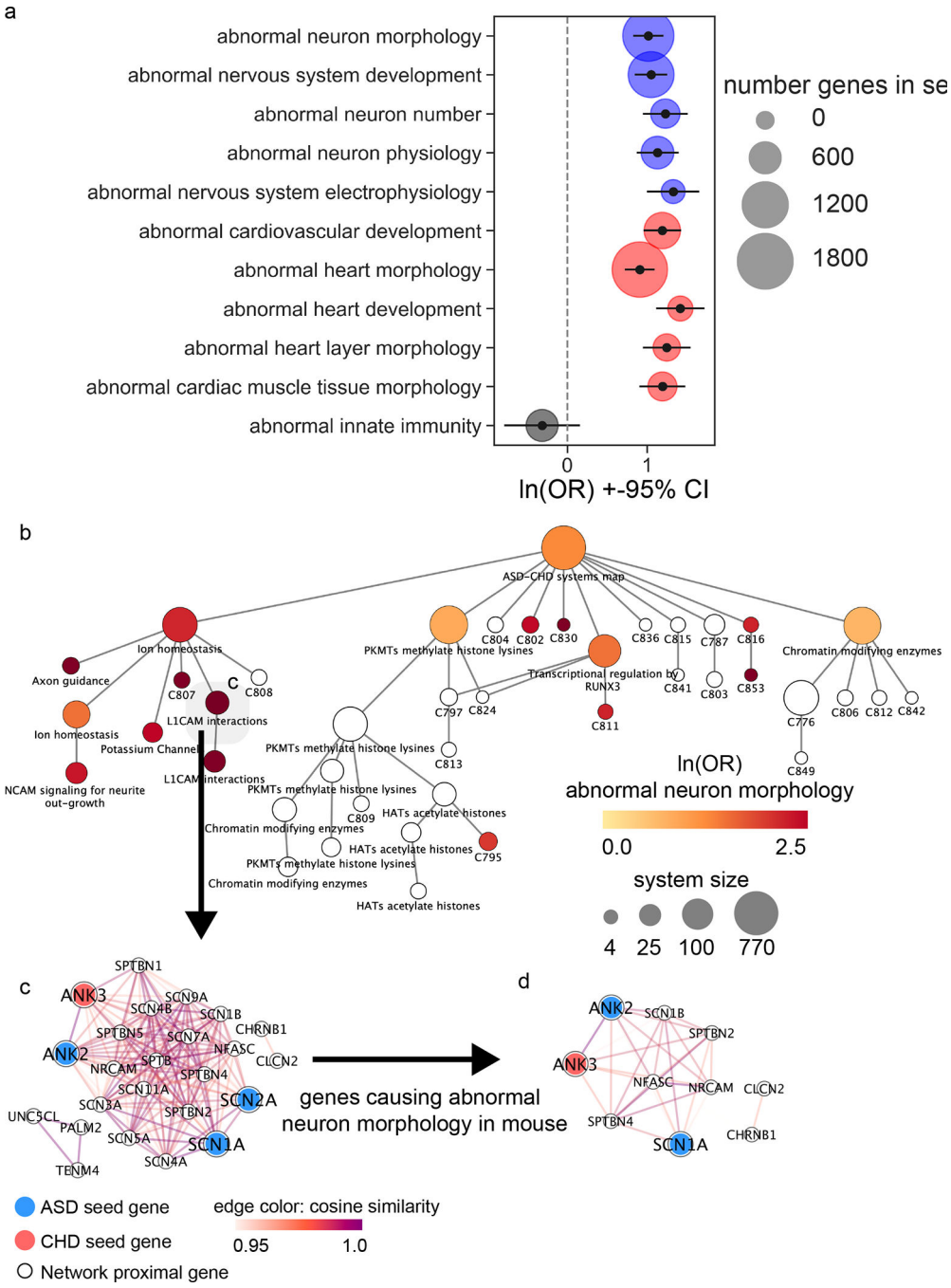

Figure 4: Validation of ASD-CHD systems map.

a) Scatterplot showing odds ratio (OR) of rnrichment of genes from the ASD-CHD NetColoc subnetwork that cause brain-related phenotypes (red), heart-related phenotypes (blue), or a negative control phenotype (gray) when mutated in mice. Error bars show a 95% confidence interval around the log odds ratio (OR). Circle size indicates the number of genes in the selected phenotype. b) Enrichment of genes causing abnormal neuron morphology in the ASD-CHD systems map, with natural log of the odds ratio indicated by node fill color. Systems not significantly enriched (p>0.05) are indicated with white nodes. c) Genes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and interactions contained within the L1CAM interactions system indicated in (a). d) Genes and interactions within the L1CAM interactions system which result in abnormal neuron morphology when knocked out in mice. Edge colors in c) and d) represent cosine similarity.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

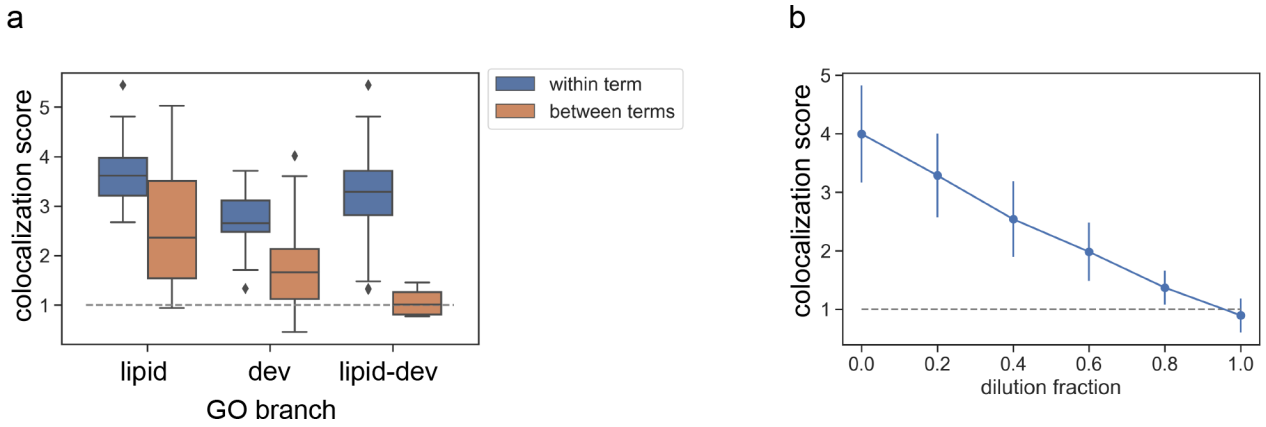

Figure 5: Benchmarking on Gene Ontology (GO).

a) Boxplots showing network colocalization score, defined as the size of the observed network colocalization subnetwork, divided by the expected size given randomly selected genes, measured for gene set pairs selected from within the same GO terms (blue), and between different GO terms (orange). Within-term and between-term network colocalization is measured in the lipid metabolic process branch (GO:0006629, lipid), in the cell development branch (GO:0048468, dev), and between the lipid metabolic process branch and the cell development branch (lipid-dev). b) Network colocalization is measured for gene set pairs selected from within the same GO terms, and diluted by a specified fraction of randomly selected genes. Error bars show 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Troubleshooting table.

Step	Problem	Possible Reason	Solution
1	ImportError ddot package	DDOT not installed	Install DDOT package. See https://github.com/ucsd-ccbb/NetColoc#dependencies for instructions
3, 21,22	Operation hangs or there is an error	Slow or no internet connection, NDEx server down	Verify internet connection, retry, report issue with NDEx https://www.ndexbio.org
4	Memory error calculating w_double_prime	Hardware is underpowered	Try loading a smaller network in step 3 (recommend STRING high confidence: UUID: 275bd84e-3d18-11e8-a935-0ac135e8bacf)
5	No input genes overlap with network	Gene nomenclature may be incompatible	Verify that input genes and network nomenclature are the same (e.g., both are mouse or both are Entrez ID)
12	Operation hangs or there is an error	Slow or no internet, community detection service (CDAPS) down, such that Community Detection algorithm is unable to generate a result. CDAPS is accessed from NetColoc dependency cdapsutil.	Verify internet connection and retry. If still failing try running locally via Docker: https://cdapsutil.readthedocs.io/en /latest/auicktutorial.html#step-2-choose-where-to-run. If all else fails report issue at: https://cdapsutil.readthedocs.io/en /latest/contributing.html#report-bugs
13	Blank cell instead of network visualization	Running in Jupyter Labs instead of Jupyter Notebooks	Alternate installation instructions required for use of ipycytoscape in Jupyter Labs https://github.com/cytoscape/ipycy toscape
26	Open in Cytoscape button grayed out	Cytoscape not open, or version out of date	Open Cytoscape, or update Cytoscape version
27, 28	yFiles layouts are not	yFiles Layout	Install yFiles Layout Algorithms
	available	Algorithms not installed in Cytoscape	through Cytoscape app manager