Aparna Nathan 04/27/2018

A network-based approach to functional interpretation of differential gene expression

Introduction

RNA-seq provides valuable insight into the expression profiles that drive cell-, tissue-, and phenotype-specific functions. A common approach to interpreting RNA-seq is to identify genes with different expression levels in different phenotypes. However, it is often challenging to make functional conclusions about the resulting significantly differentially-expressed genes (DEGs).[1] While the differential expression may be related to the observed phenotype, the method does not inherently define the connection and it excludes functionally-relevant transcripts that are not significantly differentially expressed.

Pathway gene set enrichment analysis calculates the overrepresentation of genes in the same functional pathway among the DEGs, but these results are limited by existing pathway or function annotations in databases. They also don't account for the functional context, comprised of the other genes and proteins expressed in the sample. To fully understand how differential expression of a particular gene contributes to a correlated phenotype, its functional data must also be considered. This can include a variety of data types, such as translation kinetics (since the active molecule is often the protein), regulation by transcription factors, and complex formation or other physical interactions.

In this proposal, I present a functionally-driven, network-based workflow to 1) identify modules of genes or proteins that correlate with the phenotype of interest, and 2) determine the functional role of the differentially-expressed modules. The goal of this method is to integrate gene expression data with other ancillary data (i.e., protein quantification, transcription factor binding motifs, and protein interaction networks) to expand the search space of candidate differentially-expressed genes beyond the traditional definition, and to offer a more context-specific functional role for these genes.

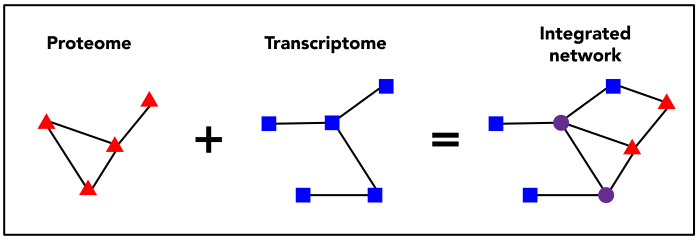
In accordance with these dual goals, the workflow has two stages. First, rather than identifying DEGs, I propose to identify differentially-expressed modules (DEMs) from co-expressed genes/proteins and from interacting proteins. The multilevel approach accounts for the functional potential of genes at the transcript, protein, and protein complex levels. Second, the functional role of the DEMs is determined through network topology analysis of the subnetworks representing the constituent genes' relationships at the two main levels of functional interactions: protein-DNA (approximated through transcription factor binding activity) and protein-protein.

This workflow can be applied in any experiment where varied data types are available, but I specifically propose an application in human brain tissue to understand the etiology of psychiatric disorders — namely, schizophrenia. This method is particularly well suited for human brain tissue because the brain has highly complex regulatory and functional programs. Prior RNA-seq-based studies of schizophrenia have pointed to a heterogeneous set of genes that haven't been functionally resolved, so a new method could be beneficial.[2,3,4]

Aim 1: Construct an integrated transcriptomic-proteomic co-expression network

Differential expression of both transcripts and proteins in schizophrenia can offer clues about the genetic basis of the disease. One common approach to identify modules of candidate genes is to construct co-expression network, with the assumption that co-expressed genes often act in shared pathways. Integrating transcript and protein expression into these networks will allow them to capture the way that a given gene's transcript and protein might be regulated differently and have distinct functions in the context of the disease.

Aim 1



1.1: Data generation via RNA-seq and mass spectrometry

Existing databases don't contain both RNA-seq and mass spectrometry measurements for controls and individuals with schizophrenia. There are resources with one of the data types, but in order to accurately compare transcript and protein expression, they should be measured in samples from the same subject. I propose carrying out RNA-seq and mass spectrometry on postmortem samples of the dorsolateral prefrontal cortex (DLPFC). This region of the brain has been previously shown to be dysregulated in schizophrenia.[5] Samples should be collected from comparable numbers of cases and controls (at least 20 biological replicates each).

Paired-end RNA-seq will be carried out for each sample with the Illumina HiSeq Ribo-Zero protocol. The reads will be processed with previously described RNA-seq quality control and analysis workflows, with the added intermediate step of surrogate variable analysis to correct for covariates (e.g., age, sex, ethnicity, batch effects).[6,7] Gene expression will be quantified in normalized transcripts per million.

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) will be used to quantify protein abundance under a standard tryptic digest protocol. The data will be searched against UniProt's annotated Swiss-Prot database for protein identification and analyzed with MaxQuant for label-free quantification and normalization.[8,9]

There are two main limitations to this experimental design. First, postmortem samples are subject to RNA degradation, and reflect the diversity of the individuals they came from. Second, brain tissue samples are heterogeneous and may obscure cell type-specific changes in expression.[10] An alternative is hiPSC-derived neural progenitor cells and neurons, but they are difficult to grow in large numbers and it is still unknown whether they recapitulate cellular phenotypes of schizophrenia.[11] To ensure the accuracy of the postmortem model, the expression profiles for control samples can be compared to known expression profiles for the DLPFC as a positive control.[12,13,14] Additionally, the cell-type composition across samples can be compared to ensure similarity by using cell-type deconvolution methods.

1.2: Merge transcript and protein quantification data in correlation network

While prior studies have constructed separate transcript and protein co-expression networks and looked for overlap between modules, I propose that the transcript and protein networks are integrated so that joint modules can be calculated.[15] Pearson coefficients are often used to correlate transcriptomic and proteomic expression data.[16] Therefore, I propose using weighted gene correlation network analysis (WGCNA), which uses a default Pearson correlation method.[17] The input will be a file containing the normalized expression level for each gene and protein in the sample. The output will be a weighted network with nodes for each gene and protein, and edges with weights corresponding to the power-transformed correlation in expression between the two nodes they connect. The network will only include nodes that exceed a minimum expression level and edges that exceed a minimum correlation level.

This network can be used to correlate the abundances of a particular transcript and its protein product based on the weight of the edge connecting them. Past studies have shown that, especially in the brain, the correlation between transcript expression and protein expression for a given gene is relatively low (around 0.5).[18] A gene's transcript and protein expression levels will be defined as highly correlated if their edge weight meets a threshold such that they share a set percentage of neighbors. For these genes, the transcript and protein nodes can be collapsed into one composite node representing the gene.

After this step, the network will have three types of nodes: transcript, protein, and composite (highly-correlated transcript + protein). Each type of edge represents a different hypothetical functional relationship: transcript to transcript nodes may be transcriptional regulation by ncRNAs, transcripts to protein nodes may be transcription factor-mediated regulation of expression (or conversely, translational regulation), and protein to protein nodes may be enzyme-substrate interactions.

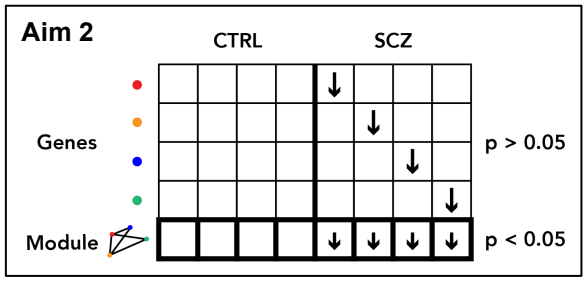
1.3: Define co-expressed modules and correlate with phenotype

Modules can be defined in a weighted gene correlation network using WGCNA's built-in hierarchical clustering algorithm. The accuracy and reproducibility of the modules can be ensured through various module preservation statistics that measure connectivity (e.g., correlation of intramodular connectivity) and density (e.g., cluster coefficient) of a module.[19] As a negative control, these metrics will be calculated for 1000 randomly permuted gene sets, and the Z score of the module's metric can be determined based on this null distribution. The genes comprising the module can also be tested for enrichment of Gene Ontology (GO) terms or cell type-specific signatures. This serves as a weak positive control since we expect to see some enrichment of broad functionality among co-expressed genes.

Each module in the integrated network is defined by an eigengene — the first principal component of the module expression matrix — that serves as an expression profile for the constituent genes. The next step is to correlate the module eigengenes with the phenotypes of interest. Because the eigengenes in the integrated network incorporate both genomic and proteomic data, they may offer improved insight into modules with different activity in cases and controls. To determine the significance of the correlation between each eigengene and the two phenotypes, I will use a t-test to compare the distribution of correlations of the eigengene with samples of each phenotype. (For more phenotypes, an ANOVA test would be appropriate.) As a negative control, I will permute the sample labels and measure the correlation of each eigengene with the now-arbitrarily labeled phenotypes.

Aim 2: Identify differentially-expressed interaction modules

One shortcoming of co-expression networks is that they assume functionally-similar genes involved in a phenotype are correlated to each other and to the phenotype. However, this overlooks genes that might not be significantly correlated by themselves, but instead are part of a multimeric interaction that is correlated. The interaction can then be perturbed by differential expression of any of its subunits.[20]



2.1: Define an expression-constrained protein-protein interaction network

First, we need to identify the proteins that could be interacting in the DLPFC. This information can be extracted from protein-protein interaction (PPI) databases. Even if we limit our search to databases that exclusively contain observed interactions (rather than predicted interactions), there are still many publically available databases, and I will use STRING as a starting network of all possible human PPIs.[21] From this basis, I will prune the network to only include proteins that were identified at an abundance exceeding a set threshold in any of the samples. I will also remove any protein with expression level that is uncorrelated with its transcript expression level because the goal is to gain insight into the effects of differential gene expression. The resulting PPI network will be specific to the expression constraints of the samples in this study. This is a relatively conservative approach that will eliminate many nodes and edges from the network (including nodes for which a transcript was detected, but no protein), and although it may introduce some bias against proteins that are difficult to quantify by LC-MS/MS, it is a high-confidence way to eliminate any interactions that are not tissue-specific.

2.2: Identify genes in differentially-expressed interaction modules

The goal of integrating the PPI data is to identify genes that play an important functional role in the phenotype, but aren't significantly differentially expressed between cases and controls. One method to identify these genes is to adjust differential expression p-values to

account for interactions.[22] This approach relies on two main constraints: 1) the change in a gene's differential expression p-value should be minimized, and 2) genes with strongly interacting protein products should have similar significances of differential expression. Considering these parameters, the algorithm minimizes an energy function that models the distribution of the significance of differential expression across the interaction network. I propose using this algorithm to recalculate p-values for every protein in the co-expression pathway.

The result is a network in which the significant differential expression converges on modules of interconnected proteins that initially had non-significant differential expression. To test the correlation of these modules with the phenotype, I propose to take the geometric mean of the normalized TPM count for these genes for each sample, and test this module measurement for differential expression between the cases and controls. The geometric mean introduces some bias for the module's expression to be driven lower by the underexpression of any one component, but this may be acceptable for our biological question because in a multimeric interaction, decreased abundance of one monomer is stoichiometrically more likely to change the multimer's function than an increase in the abundance of one monomer.

The algorithm's authors also acknowledge a risk for inflating the p-values of highly connected nodes in the interaction network. Even though they show that highly connected subnetworks that are unrelated to a phenotype are not enriched for significance in that phenotype, this concern can be addressed in the current application of the algorithm with the negative control of permuting the phenotype labels on the samples and re-calculating the differential expression of the interaction modules between the now-arbitrarily-labeled cases and controls. I expect to see non-significant differential expression in these tests.

Aim 3: Infer transcription factor-mediated regulation driving phenotypic correlation

With this expanded set of candidate DEMs that may be implicated in schizophrenia (based on co-expression and PPIs), the next step is to identify the functional role that they might play in the disease etiology. One potential explanation of some of the co-expression relationships that are correlated with the phenotype is that they reflect transcription factors regulating expression of their disease-associated targets (specifically via the edges between protein and gene nodes).

3.1: Predict regulatory networks based on expressed TFs and sequence motifs

In order to construct a putative network that accurately represents the specific regulatory state of the samples, I propose using an algorithm that integrates the different types of data (gene expression, protein expression, protein interaction) to inform the regulatory network. This can be accomplished with the PANDA algorithm, which employs a message-passing approach to integrate networks and predict condition-specific regulatory activity.[23] The inputs will be a list of TFs expressed in any of these samples, target motifs for these TFs (from JASPAR), gene expression matrix, and PPI network.[24] The output will be a directed, weighted network with TFs and genes as nodes, and edges going from TF node to regulated gene node.

It may be beneficial to validate a random subset of the TFs for which regulatory activity is predicted. In this experimental system (postmortem tissue), I will only be able to validate the binding of the TFs, but not the effect on gene expression. If primary samples from the controls and individuals with schizophrenia are available, this predicted regulatory network can be verified with ChIP-seq. This will require access to antibodies for all TFs predicted to be involved in the network, and following a standard ChIP-seq protocol, I will identify the genomic loci where each TF is bound. I may be able to validate the effect on gene expression with a neuronal cell line in which the gene is regulated by the same TF. By knocking down TF expression with a siRNA and using RT-qPCR to measure expression of the specific target gene, I can confirm the predicted regulatory relationship. (Conditions with a scrambled/non-specific siRNA and an RNAi-resistant rescue plasmid serve as the negative and positive controls, respectively.)

3.2: Identify overlap between regulatory network and DEMs

This regulatory network will contain subnetworks of co-regulated genes; to determine whether any of these modules are explain any of the co-expressed DEMs, I propose identifying modules that overlap with DEMs. Since I don't necessarily expect complete overlap between modules in the two networks, I will count the number of shared edges between pairs of DEMs and regulatory modules and rank the pairs by percent overlap. Even though the regulatory network is directed, it will be treated as undirected. The significance of the overlap will be calculated based on the overlap between each regulatory module and random gene sets of the same size as the DEMs. If there are regulatory modules that significantly overlap with the DEMs, it may be useful for someone else to take on the task of validating them (since it is outside the scope of this project). Identifying this overlap allows us to annotate the nature of some of the observed co-expression relationships.

Aim 4: Characterize effects of perturbing functional PPIs

The final goal of this proposal is to use interactions from the PPI network to infer the way that differential expression of its components would disrupt functional relationships. This can be accomplished by measuring the impact of DEMs on flow through the PPI network.

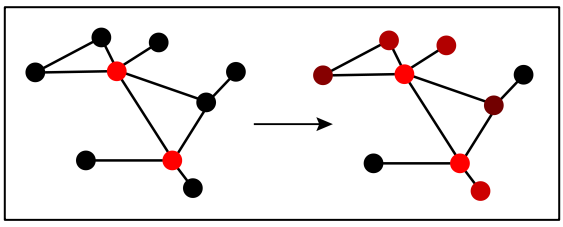
4.1: Map DEMs to functional clusters in PPI network

For each DEM, the constituent genes' products can be located in the PPI network. The PPI network can then be clustered with an algorithm like clusterONE, which is specifically tailored to identify overlapping clusters in PPI data.[25] This will assign each constituent gene product to one or more clusters, and the enrichment of DEGs can be measured in each cluster. As described previously, the significance of the enrichment can be determined by calculating the enrichment of random gene sets of the same size as the DEMs and generating a permutation p-value. Then, a simple first-pass assignment of function can be done with GSEA for GO terms or pathway gene sets. This will highlight any DEMs or subsets of DEMs that are enriched for interactors. Because some of our DEMs were originally derived from the PPI network, we expect that those DEMs will be enriched in the individual clusters (although it is not necessarily true for all the interaction-derived DEMs, because they were defined based on the p-value adjustment algorithm, rather than a cluster definition algorithm).

4.2: Calculate perturbation of information flow in PPI network by DEMs

Studies have proposed that random walks are the best way to model flow in a PPI network.[26,27] Therefore, I propose using a popular network propagation algorithm to identify clusters of protein interactions that will be affected by the differential expression of an individual protein.[28] As its input, this topology-based algorithm requires a weighted network and a subset of nodes to assign non-zero scores based on their association with the phenotype in question. In this case, the subset of nodes will be the proteins in one of the DEMs and from these starting points, the algorithm will propagate the non-zero schizophrenia association of these proteins into the neighboring nodes (proteins that the implicated proteins interact with) based on the weight of the edge (representing the strength and significance of the interaction).

Aim 4



Together, this series of network-based analyses will provide a more sensitive approach to identify and understand the function of DEGs in disease. It can be expanded to include other types of data (e.g., ncRNA binding targets, using the approach from Aim 3) and experimental set-ups (e.g., single-cell sequencing with REAP-seq for quantifying transcripts and proteins).[29]

[1] Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*(12), 550.
[2] Wu, J.Q., Wang, X., Beveridge, N.J., Tooney, P.A., Scott, R.J., Carr, V.J., and Cairns, M.J. (2012). Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. PLoS One *7*(4), e36351.
[3] Ramaker, R.C., et al. (2017). Post-mortem molecular profiling of three psychiatric disorders. Genome Med *9*, 72.
[4] Gandal, M.J., et al. (2018). Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. Science *359*(6376), 693-697.
[5] Fillman, S.G., Cloonan, N., Catts, V.S., Miller, L.C., Wong, J., McCrossin, T., Cairns, M., and Weickert, C.S. (2013). Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. Mol Psychiatry *18*(2), 206-214.
[6] Yang, I.S., and Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. Genomics Inform *13*(4), 119-125.
[7] Leek, J.T., and Storey, J.D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet *3*(9), 1724-1735.
[8] The Uniprot Consortium. (2018). UniProt: the universal protein knowledgebase. Nucleic Acids Res *46*(5), 2699.
[9] Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc *11*(12), 2301-2319.
[10] Hosp, F., and Mann, M. (2017). A Primer on Concepts and Applications of Proteomics in Neuroscience. Neuron *96*(3), 558-571
[11] Hoffman, G.E., Hartley, B.J., Flaherty, E., Ladran, I., Gochman, P., Ruderfer, D.M., Stahl, E.A., Rapoport, J., Sklar, P., and Brennand, K.J. (2017). Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. Nat Commun *8*(1), 2225.
[12] Hawrylycz, M.J., et al. (2012). An anatomically comprehensive atlas of the adult human brain transcriptome. Nature *489*(7416), 391-399.
[13] Colantuoni, C., et al. (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. Nature *478*(7370), 519-523.
[14] Uhlén, M., et al. (2015). Tissue-based map of the human proteome. Science *347*(6220), 1260419.
[15] Seyfried, N.T., et al. (2016). A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. Cell Syst *4*(1), 60-72.
[16] Kuo, T.-C., Tian, T.-F., and Tseng, Y.J. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol *7*,84.
[17] Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.
[18] Sharma, K., et al. (2015). Cell type- and brain region-resolved mouse brain proteome. Nat Neurosci *18*(12), 1819-31.
[19] Langfelder, P., Luo, R., Oldham, M.C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? PLoS Comput Bio *7*(1), e1001057.
[20] Matalon, O., Horovitz, A., and Levy, E.D. (2014). Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. Curr Opin Struct Biol *26*, 113-120.
[21] Sklarczyk, D., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.

[22] Poirel, C.L., Rahman, A., Rodrigues, R.R., Krishnan, A., Addesa, J., and Murali, T.M. (2013). Reconciling differential gene expression data with molecular interaction networks. Bioinformatics *29*(5), 622-9.
[23] Glass, K., Huttenhower, C., Quackenbush, J., and Yuan, G.-C. (2013). Passing Messages between Biological Networks to Refine Predicted Interactions. PLoS One *8*(5), e64832.
[24] Khan, A., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res *46*, D260-D266.
[25] Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods *9*(5), 471-472.
[26] Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. Bioinformatics *26*(8), 1057-1063.
[27] Gulbahce, N., et al. (2012). Viral Perturbations of Host Networks Reflect Disease Etiology. PLoS Comput Biol *8*(6), e1002531.
[28] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol *6*(1), e1000641.
[29] Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol *35*(10), 936-939.