**A Hybrid Approach to Assessing Gene Module Construction in Saline Versus Ethanol Treated Recombinant Inbred Mice**

**Background:** When attempting to identify the biological pathways associated with a response to alcohol (ethanol) in individuals, researchers encounter the problem of selecting which of the tens of thousands of genes might be a determining factor. Even if one can determine the differential co-expression between a control (saline) and case (ethanol) group, it is difficult to interpret the meaning of that difference with any confidence. Continued genetics/genomics studies are confirming the notion that genes operate in systems: very few traits are explained by a single gene. Building on the idea that genes operate like networks in telecommunications or social media, we will assess the gene network constructed in a saline-treated group and observe the preservation or disruption of that network in a corresponding ethanol-treated group of recombinant inbred mice.

Gene Set Enrichment Analysis (GSEA) analyzes genes associated with annotated biological pathways. While useful for interpretation, GSEA requires a priori knowledge of biological pathways to establish the relationship of multiple genes to a phenotypic profile instead of allowing the expression data to drive the formation of gene associations. (8)

Weighted Gene Co-expression Network Analysis (WGCNA) groups genes into modules by establishing “a connection weight to each gene pair” determined by their co-expression levels. WGCNA works well for relating gene co-expression to a phenotype but does not provide a clear mechanism to differentiate co-expression levels across two (or more) study groups. (Zhang and Horvath, 12)

The Discordant method (Siska et al.) uses a “binning” approach to calculate posterior probabilities of co-expression in one study group (control) juxtaposed against a second study group (case). For example, do two genes show a weak expression correlation in the control group but a strong correlation in the case group? Discordant provides a statistically powerful construct for evaluating differential co-expression in gene pairs but is not used to find modules of genes operating together. (6)

DiffCoEx introduced by Tesson et al. is a straightforward application of WGCNA that is able to quickly “identify gene co-expression differences between multiple conditions”. However, it is not clear how to interpret the results of its operation since it is not measuring genes operating as a system. (10)

Langfelder et al. construct a Zsummary score built on the patterns of high connectivity between genes to evaluate the preservation of network modules. Then, a Zsummary score greater than 10 provides “strong evidence of preservation”. Some argue that this cutoff is arbitrary and so, not very informative for drawing conclusions. (2)

We will utilize a hybrid approach that combines the best features of WGCNA and Discordant to produce a set of modules that can be assessed for differential co-expression. Since mice exhibit a similar ethanol response as do humans, we will apply our analysis technique to a study of two groups of recombinant inbred mice: one treated with saline and the other treated with ethanol.

**Study Aims:** We have two study aims for this research:

1. Find the biological pathways that are disrupted/activated in exposure to ethanol in recombinant inbred mice by analyzing differential co-expression levels using a hybrid approach of the Discordant and WGCNA methods.
2. Compare the hybrid approach to the WGCNA and Discordant methods individually for interpreting the meaning of the distinction between the saline and ethanol groups. Additionally, assess the resultant activated/disrupted modules relative to established biological pathway databases: KEGG, MsigDB, PPI, etc. (1)

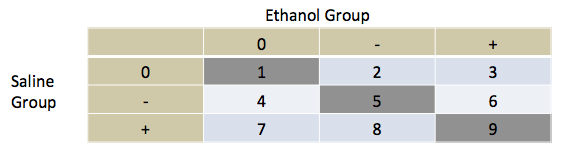
**Data:** This hybrid approach will be applied to a study of ethanol response in mice using a control group treated with saline and a case group treated with ethanol. The data are “messenger RNA (mRNA) sequencing data from a large RI mouse panel that have been bred from reciprocal crosses between the Inbred Long Sleep (ILS) and Inbred Short Sleep (ISS) strains, called the ILSXISS (LXS) panel. The resulting mRNA dataset contains 118 ethanol samples and 114 saline samples. Each group represents 40 LXS strains, 2 to 3 replicates each. Total RNA was extracted from whole brain tissue using RNeasy Mini Kits (Qiagen, Valencia, CA), quantity and quality were determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) and Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA). The libraries were prepared using Illumina scripts RNA-Seq Library Preparation Kit v2. Sequencing was performed on the Illumina HiSeq 2000 platform.” (3) The unwanted variation in the data was removed assuming a negative binomial probability distribution and the data were transformed to a log-normal scale using a variance stabilizing algorithm. (4) There are 14,184 genes in the original data set.

In order to reduce the amount of data processed by the Discordant package, to avoid skews in the correlations and to reduce potential false positives that arise in the algorithms used in the package, the genes will be analyzed for outliers across the strains. We will use the grubbs.test within the R package: “outliers” due to the fact that the original count data has been transformed to follow a Gaussian distribution. This function will be applied to the dataset that includes both the saline and ethanol groups. The application of this test to the original dataset reduces the dataset to 9879 genes to be analyzed. From this resultant dataset, we will calculate heritability scores based on the R2 values of ANOVA tests for each gene to determine the genes whose variability of expression is mostly due to genetics. From this step, we will retain the genes with high heritability (greater than 0.4?)

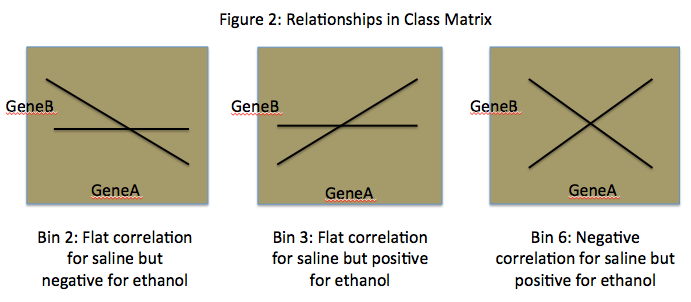
**Plan for analysis: First Discordant, Then WGCNA**

The differential co-expression for the reduced set of genes in the saline versus ethanol groups will be analyzed first using the Discordant method to estimate posterior probabilities which will be used as the input to the WGCNA method. RStudio Version 1.1.383 using R Version 3.4.1, the discordant method version 1.2.0 and WGCNA Version 1.61 will be used to conduct the analysis.

**Discordant:** Since the input data is located in a single dataset for both case and control, this type of dataset is referenced by the Discordant documentation as “within –omics”. (9) The first step in the Discordant method is to create ‘within’ correlation vectors using the default Spearman method. The Discordant function: createVectors is used to create these two vectors, one of which contains the expression correlations for each gene pair for the saline group while the other contains the same for the ethanol group. All possible pair-wise combinations of the gene expression correlations are contained in these vectors that are the primary input for the Discordant method itself. The Discordant method “estimates a posterior probability for each possible paired correlation scenario to achieve binning” in a class matrix as depicted in Table 1. (6) There are 6 bins in this matrix (2, 3, 4, 6, 7, 8) that represent differential co-expression between the two groups while the remaining 3 bins contain the probabilities in which there is no difference between the 2 groups so not of interest for this analysis (1, 5, 9).

Table 1: Class Matrix for the posterior probabilities for each gene

To interpret the meaning of the bins in the class matrix, Figure 2 demonstrates the relationships exhibited by the bins. For example, Bin 2 holds the posterior probability that Gene B has no correlation to Gene A in the saline group but has a negative correlation to Gene A in the ethanol group. Correspondingly, Bin 3 shows the posterior probability that Gene B has no correlation to Gene A in the saline group in the saline group but has a positive correlation to Gene A in the ethanol group. Finally, Bin 6 shows the posterior probability that Gene B has a negative correlation to Gene A in the saline group but has a positive correlation to Gene A in the ethanol group.



For analysis, we will group the 6 bins of differential co-expression. That is, the posterior probability values for bins 2 and 3 will be added together to assess the differential co-expression (positive or negative) when the saline group is flat. Correspondingly, bins 4 and 7 will be added together to assess the differential co-expression when the ethanol group is flat. Finally, bins 6 and 8 will be added together to assess the differential co-expression when the saline relationships travel in the opposite direction from ethanol relationships. These three datasets will provide the input to the WGCNA method, and for convenience, we will refer to these as classes of FlatSaline (Bins 2 and 3), FlatEthanol (Bins 4 and 7) and Crossed (Bins 6 and 8) in the rest of this discussion. FlatSaline, FlatEthanol, and Crossed will each be treated to their own WGCNA analysis. That is, we will draw conclusions about each of these classes separately and observe the similarities and differences between the classes.

**WGCNA:** The first step in the WGCNA method requires a measure of co-expression similarity and typically uses a correlation metric as its input. We believe that the posterior probability values generated in the previous Discordant step offer a legitimate, statistically sound measure of similarity for these gene pairs. The typical second step of WGCNA is to create an adjacency matrix by raising the correlation measures to a “soft threshold”. There is some question whether this second step is appropriate given the source of the similarity measures. It is possible that the posterior probabilities themselves will serve as the adjacency matrix. This researcher will work through hypothetical scenarios of binning to determine whether the initial matrix serves as the similarity matrix or the adjacency matrix. Regardless of the chosen path, this step produces an adjacency matrix that contains values in the interval (0,1).

From this adjacency matrix, a topological overlay matrix will be constructed to take into account the neighbors of the genes. This step allows us to utilize pathways shared by gene pairs instead of just examining the direct correlation between a gene pair. A dissimilarity matrix will be constructed from that topological overlay. Using the dissimilarity matrix, a tree structure will be constructed with the help of a dynamic tree-cutting algorithm to generate unique gene modules for each of the classes (FlatSaline, FlatEthanol, and Crossed). To interpret the results of this analysis, we will explore the gene modules that are identified uniquely in each of the classes. For example, in the FlatSaline class, we will recognize gene modules that are negatively or positively co-expressed in the ethanol group but have no co-expression in the saline group. Alternatively, in the FlatEthanol class, we will distinguish gene modules that are activated in the saline group but are disrupted in the ethanol group. Additionally, in the Crossed class, we will spot gene modules that arise from a negative co-expression relationship in the saline group but a positive co-expression in the ethanol group (or vice versa). We can then describe the differential co-expression for each scenario to draw conclusions about our confidence level of the impact of ethanol treatment versus saline treatment in mice.

**Evaluation:** To determine the benefit of applying the hybrid approach to the ethanol response problem, we will also analyze the data using Discordant by itself and WGCNA by itself.

In the case in which Discordant is used, we will reach the end of the analysis when producing the posterior probability distributions in the steps outlined above. From these probability distributions, we continue to the class phase (FlatSaline, FlatEthanol and Crossed). For each of these classes, we will have a set of probability distributions for each gene pair and we will determine a cutoff threshold for which the relationship between that gene pair is ‘interesting’. Since we believe that much of the probability distribution for any given gene pair lies in Bins 1,5 and 7 (that is, no differential co-expression between saline and ethanol), we think the threshold should be somewhere around 0.20. Based on that threshold, we will determine those gene pairs that exceed the threshold in each of the classes and will call those differentially co-expressed under the Discordant method.

In the case of using WGCNA by itself, we will examine the saline group and the ethanol group separately as the source for our analysis. We will use the default Pearson correlation value as the input to each correlation matrix and then proceed with the normal steps in WGCNA: “soft” thresholding, adjacency matrix, topological overlay, dissimilarity, tree-cutting and gene module generation. Each of these analyses will produce a set of gene modules and we will examine these modules to determine which are activated in the saline group but disrupted in the ethanol. Similarly, we will observe those modules identified in the ethanol group but disrupted for saline. In this scenario, it will be useful to compare these results to the preservation score (Zsummary) outlined by Langfelder et al. (2)

Additionally, we will examine publicly available databases (KEGG, MsigDB, PPI, etc.) that map established biological pathways to determine if our analysis sheds some biological insight.

Cited Resources

1.Kegg database: <http://www.genome.jp/kegg/pathway.html#mapping>

2.Langfelder P, Luo R, Oldham MC, Horvath S. Is My Network Module Preserved and Reproducible? Bourne PE, ed. *PLoS Computational Biology*. 2011;7(1):e1001057. doi:10.1371/journal.pcbi.1001057.

3. Rudra P, Shi WJ, Vestal B, et al. Model based heritability scores for high-throughput sequencing data. BMC Bioinformatics. 2017;18:143. doi:10.1186/s12859-017-1539-6

4. (WJ Shi, personal communication, December 12, 2017)

5. Siska C and Kechris K (2016). discordant: The Discordant Method: A Novel Approach for Differential Correlation. R package version 1.2.0, <https://github.com/siskac/discordant>.

6. Siska, C, R. Bowler, and K. Kechris, “The discordant method: a novel approach for differential correlation,” Bioinformatics, vol. 32, no. 5, pp. 690–696, 2015

7. Siska, C. and K. Kechris, “Differential correlation for sequencing data.” Submitted, 2016.

8. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102.

9. Siska: Discordant R package: <https://www.bioconductor.org/packages/3.7/bioc/vignettes/discordant/inst/doc/Discordant_vignette.pdf>

10.Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics*. 2010;11:497. doi:10.1186/1471-2105-11-497.

11. Weiss JN, Karma A, MacLellan WR, et al. “Good Enough Solutions” and the Genetics of Complex Diseases. *Circulation research*. 2012;111(4):493-504. doi:10.1161/CIRCRESAHA.112.269084.

12. Zhang, Bin and Horvath, Steve (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis," *Statistical Applications in Genetics and Molecular Biology*: Vol.4: Iss. 1, Article 17.