**Thesis Proposal**

Lauren Vanderlinden

Spring 2011

**Overview/Background**

**Alcoholism**

Alcoholism is a broad diagnosis and includes patients who are either psychologically or physically addicted to alcohol. The DSM-IV breaks down alcoholism into two definitions: alcohol abuse and alcohol dependence. Alcohol dependence is defined by DSM-IV as a person who meets 3 of the following criteria over a year: tolerance, alcohol withdrawal signs/symptoms, drinking more than intended, unsuccessful attempts to cut down on use, excessive time related to alcohol (obtaining, hangover), impaired social or work activities due to alcohol, or use despite physical or psychological consequences. Alcohol abuse is defined by DSM-IV as a person who meets 1 of the following criteria over a year: role impairment (e.g. failed work), hazardous use (e.g. driving while drinking), legal problems related to alcohol use, or social/interpersonal problems due to alcohol. Alcoholism is a terrible disease which disables its victims and costs society billions of dollars every year. Society has suspected alcoholism is an inheritable for a long time and now those suspicions are supported by research. Numerous studies have found evidence alcoholism is genetically controlled12-13,15-16, but is also influenced by environmental factors. These studies show heritability by using twin sibling and adoption studies16, but have not found a specific alcoholic gene.

Along with two definitions for alcoholism, there are also numerous co-morbidities, including an extremely high co-morbidity rate with depressive disorders1. The National Institute on Alcohol Abuse and Alcoholism (NIAAA) reported 37% of persons with alcohol abuse disorders also have psychiatric disorders and some of these disorders with high rates of alcohol abuse/dependence include attention deficit hyperactivity disorder (ADHD), post-traumatic stress disorder (PTSD), anxiety, major depressive disorder (MDD), and drug use disorders11. Co-morbidities present concerns in treatment due to lack of knowledge about what disease is primary or if the diseases are co-occurring with no interrelationship11. All these co-morbidities and the broad definition of alcoholism make for a heterogeneous population of patients. Some have tried to differentiate between types of alcoholics in hopes to find neurobiological pathways that are unique to each one. Lesch addressed co-morbidities by establishing a classification system, known as the Lesch Alcoholism Typology (LAT), which categorizes alcoholics into 1 of 4 categories based on other psychiatric phenotypes. Lesch has shown these different sub-types of alcoholics respond better to different treatments. An overarching goal in the field is to accurately diagnose a patient into sub-types of alcoholics for optimal treatment. We are specifically interested in the co-morbidity between alcoholism and MDD.

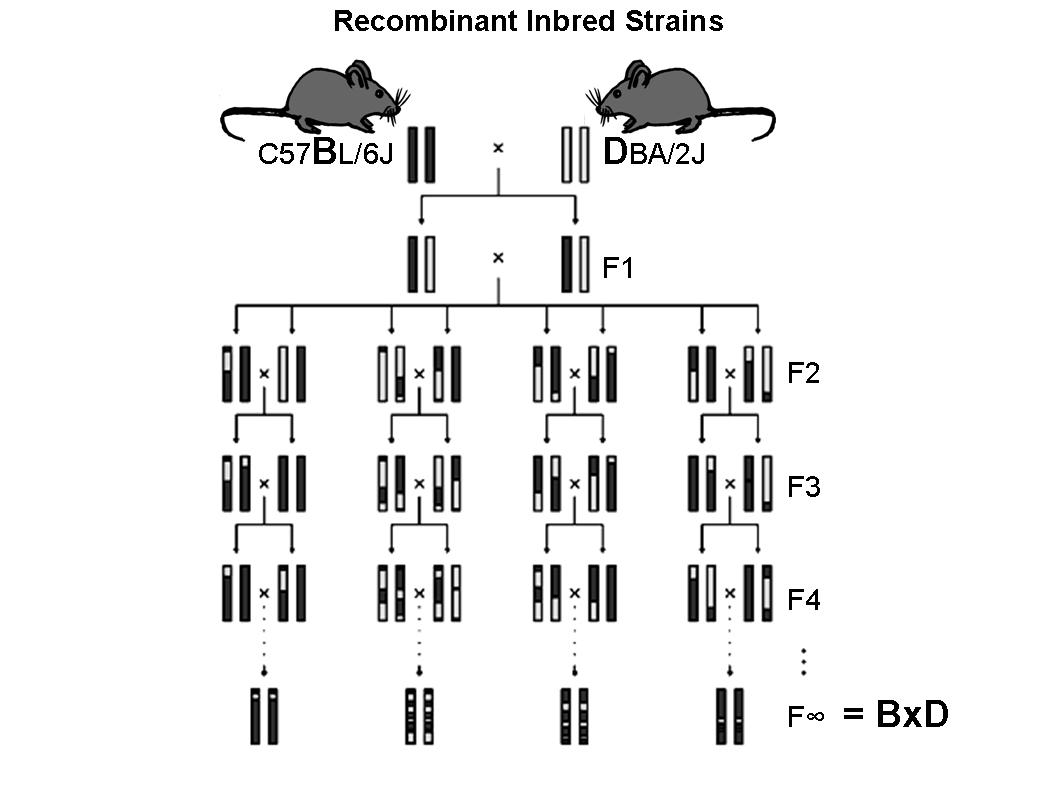
**Mouse Models**

Due to the high heterogeneity of alcoholics there are numerous benefits in studying alcoholism in animal models. We can control the environment the animals, so we are sure to be examining the genetic control alone. We also know a complete drinking history of each animal whereas humans can often under/over report this information. Alcoholism and other psychiatric diseases are conceptualized as disorders of brain circuits and neurological pathways. Therefore, it is ideal to look at brain tissue expression levels for our study. However, a brain biopsy on humans is an extremely dangerous procedure and could potential damage cognitive function in the biopsied area. A biopsy can only look at a certain area of the brain (the part biopsied) while in an animal study we can look at the whole brain. A mouse model is ideal for this study because 90% of the mouse and human genomes are regions conserved and can be mapped to a corresponding region of the other18. There are no animal models which we can label “alcoholic”, but we can look at endophenotype’s to model alcoholic behavior. Acute and chronic toxicity from alcohol, quantity of alcohol consumed, tolerance and physical dependence are all different responses to ethanol from the central nervous system (CNS)17,18. Using endophenotypes that beak down the different responses of alcohol gives us a more stable behavior which we can measure accurately. We will be using ethanol consumption as an endophenotype for alcoholism for this will tell us about preference and consumption.

In many mouse studies for complex traits, conclusions are drawn based on mice from an F2 generation. Inbred parents are bred to form a F1 generation (one known allele from each parent) and the F1 are crossed together to form an F2 generation (mosaic of alleles from both parents due to recombination). In the F2 generation, each mouse is genetically unique and the population is not replaceable. Because the genetic make-up of mice within the F2 population cannot be replicated, all genetic and phenotypic measurements must be done on the same mouse. As an alternative, we are using a recombinant inbred (RI) panel. RI strains are formed as described above, but instead of examining the F2 generation they are continually bred together for more than 20 generations and thus generating multiple strains of inbred F2 mice. Figure 1 shows a depiction of this process. A considerable advantage of using a RI panel is that they are a renewal resource; because all mice within an RI strain are genetically identical. All genetic and phenotype information gathered is cumulative across generations and researchers. Each RI strain is unique and has roughly equal amounts of genetic contribution from each parent and their genetics are independently distributed among strains. Another advantage of using RI strains is the wide range of phenotypes they express. Instead of having two parental mice that are extremely different in many traits, these mice offer a continuum of traits.

BXD RI mice are a cross between a C57BL/6J female mouse and a DBA/2J male mouse. These two inbred mice differ on a wide range of behavior and physiological measures. It has been shown DBA mice show an extreme aversion to alcohol and morphine while C57 mice have a preference for alcohol and morphine. There is a high amount of genetic variation between the RI strains and we will be able to get a continuum of traits across the strains. These strains express a wide variety of endophenotypes that are related to alcoholism, such as alcohol preference and acute functional tolerance but also have a wide variety of differing traits not related to alcoholism, like coat color. An advantage of studying BXD RI mice is there are numerous resources available beyond our brain expression data: complete DNA sequence information on both of the parents (Sanger Institute), an extensive marker set (Wellcome Trust), and large phenotype databases (GeneNetwork.org). Not only can we look into endophenotypes for alcoholism, but we can also look into endophenotypes for other diseases which alcoholism has a high co-morbidity, such as MDD.

Figure 1: BXD RI strains (http://www.nin.knaw.nl/Portals/0/Department/Levelt/20080516\_NIN\_BXD.jpg)



Identifying genes that predispose an individual to alcoholism or MDD would aid in choosing an effective treatment. Traditionally, a gene-centric approach was used in hope to find the single gene that determines an individual’s risk factor for a disease13. However, genes and gene products interact at many different levels and rarely work in isolation within the cell. The gene-centric approach does not take these epistatic effects into account and is biologically oversimplified10

**Expression Arrays**

In this study, we will be using expression arrays to measure the transcription level (amount of RNA available for translation to protein) of genes. The central dogma of biology states DNA is transcribed to RNA which is translated to functional gene products. These products are most often proteins, but there are also some non-protein coding genes as well. In most cases, transcript expression is related to the amount of these products. Although technology is not currently available to measure global levels of all proteins simultaneously, we can measure transcript levels of thousands of genes at once using a microarray. Gene transcript expression measures how active the gene is or how much gene product is available for synthesis. For example, human serum albumin is the most abundant blood protein and therefore the expression for the albumin gene (ALB) will be higher than all other genes in a blood sample.

The process of generating microarray data involves several steps. The first of which mRNA is extracted from the sample and in our case this is mouse brain tissue. The mRNA goes through a process called reverse transcription, resulting in cDNA, and is labeled with a fluorescent dye The array contains thousands of short nucleic acid polymers consisting of specific DNA sequences, called probes. These probes are used to bind to or hybridize the labeled cDNA to the array and there are multiple probes for a single gene referred to as a probe-set and there can also be multiple probe-sets for a single gene. There can be multiple isoform’s for a single gene (e.g. RNA can have alternative splicing that results in different gene products). Some probe-sets can identify a specific gene isoform, while others may be generic to all isoforms. The array used in this study is based on the 3’ end of the RNA and probes are placed at different lengths of this 3’ tail. The array is then scanned and quantitation of gene expression is made based off of the fluorescent intensity of a probe-set and will refer to a probe-set as a transcript from here on.

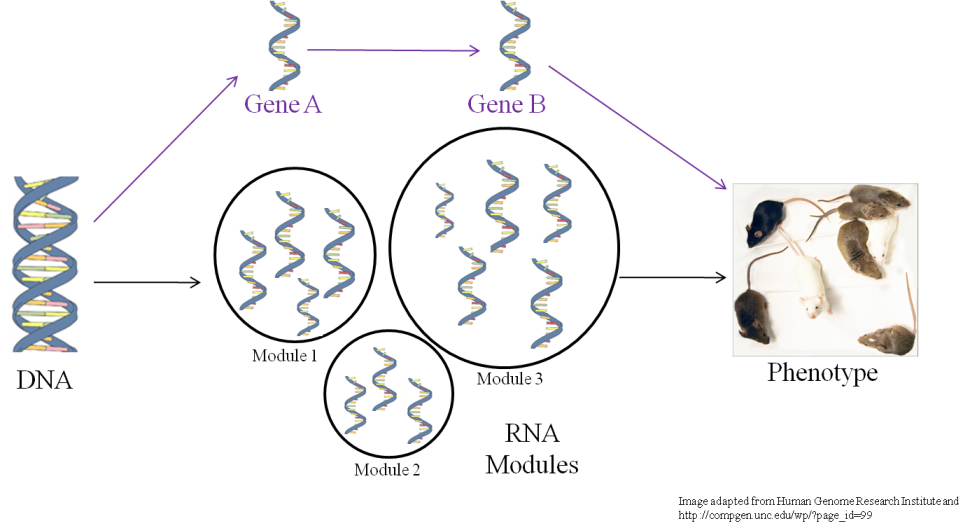
**Network Analyses**

Understanding the genetic interactions involved in the etiology of these diseases could lead to novel therapeutic targets or biomarkers. Network analysis tools are increasingly being used to describe genetic interactions as they allow for a systems biology approach opposed to studying each gene independently10,14 . A weighted-gene co-expression network analysis (WGCNA)3 is a popular network analysis method of interest to us because it focuses on modules of genes instead of individual genes. Because it focuses on a gene systems (e.g. a certain pathway or biological function) rather than looking at a particular gene it more accurately reflects biological truth. We also gain statistical power using this method. Instead of testing association of every gene individually with the phenotype, we will only test association of a summary measure of each module with the phenotype. This reduces the number of tests executed and reduced the multiple-testing burden. WGCNA also gives us insight to how genes interact with one another. Not only are gene transcripts with similar expression profiles identified as modules, but this technique allows us to statistically describe how these expression levels are related. For example, if one gene transcript is tightly connected to all gene transcripts in a module it is highly probable that this gene transcript is driving the co-expression of the other gene transcripts within that module.

WGCNA first estimates the correlation among genes across samples, then identifies clusters of highly correlated genes, and summarizes them by “hub” genes or an eigengene (1st principle component). We can then examine the relationships between clusters and phenotypic (e.g. behavioral) traits in addition to observing the interactions between clusters. Oldham and colleagues used this technique to determine different transcription modules related to certain cell types5. Based on the cell types, they were able to distinguish what area of the brain certain transcriptional processes happened without purifying out the specific cell types, which can be both time and financially consuming5.

Using this technique, we can gain insight into how the transcription process is controlled. We can connect modules to genetic variants in the DNA sequence that influences the observed expression patterns. Furthermore, we can see how the transcription differences influence the observable characteristics, such as endophenotypes related to alcoholism and depression. Figure 2 is a schematic example of what we can learn from this analysis. For example, using a gene-centric approach we could only determine how a single gene is associated with phenotype. In WGCNA we have the ability to determine if transcript A influences transcript B and if that process has a significant association with phenotype (see figure 2).

Figure 2: WGCNA schematic



With BXD RI mice and WGCNA we will be able to achieve our goal of better understanding the complex relationship between gene expression and endophenotypic behavior related to alcoholism and depression. We will do this by identifying and characterizing expression modules associated with endophenotypes of alcoholism/depression. The BXD expression data is publically available through the PhenoGen website (<http://phenogen.ucdenver.edu>) and the statistical methods included in WGCNA are available as an R package (<http://127.0.0.1:27287/library/WGCNA/html/WGCNA-package.html>).

**Methods**

**Expression Arrays**

Gene expression data was previously generated for the whole brain tissue from adult male mice using the Affymetrix mouse whole genome oligonucleotide array (MOE430 v2, Affymetrix, Santa Clara, CA) to measure RNA expression in 30 BXD RI strains plus the 2 parental strains from the Jackson Laboratory. Four to seven mice were sacrificed per strain and RNA from each mouse was hybridized to a separate array. The methods used are the same as those described in Tabakoff et al 20086.

**Cleaning & Filtering Data**

Prior to normalization, individual probes will be removed if their nucleotide sequence did not uniquely map to a region in the mouse genome or contained a known single nucleotide polymorphism (SNP). SNPs cause problems in the hybridization process. If a sample contains a SNP it results in a different sequence than that of the probe and the sample will not bind as well to the probe and the resulting expression is not reliable. Entire probe sets will be removed if less than four of the original eleven remain after filtering. The samples will be normalized using robust multichip analysis (RMA)4. MAS5, an algorithm made specifically for Affymetrix arrays, will be used to evaluate if expression level measurements by probe sets are above background noise (present, absent or marginal). If a probe set does not have at least one present call throughout all the samples, the probe set will be dropped from the RMA normalized data set. The RMA normalization method gives one value per probe-set and is highly utilized in the field. The RMA normalized and filtered data set will be used for further analysis. In our initial data set 29,468 probe sets out of the original 41,380 remained after filtering.

Extensive quality control will be implemented prior to analysis. In addition, data will be thoroughly examined for batch effects related to processing. Due to the large scale of this study, the microarrays where run in batches over a year and a half period, resulting in 15 batches determined by the month the array was scanned. ComBat analysis is one way to adjust for any batch effects9. The time and technician that ran the arrays can influence the outcome of the expression values, and therefore, we will correct for this bias.

**Data Analysis Proposal**

**WGCNA Overview**

Weighted gene co-expression network analysis (WGCNA) will be performed to determine co-expression patterns among the transcripts across RI strains. Using this analysis, highly correlated genes make up clusters/modules and using eigengene network methodology we can determine how these clusters interact with one another and characterize individual modules. We are specifically looking for highly correlated genes to determine if they are biologically linked in either function or location. For each strain, we will be using strain mean expression values for all correlation measures.

**Constructing Gene-Coexpression Networks**

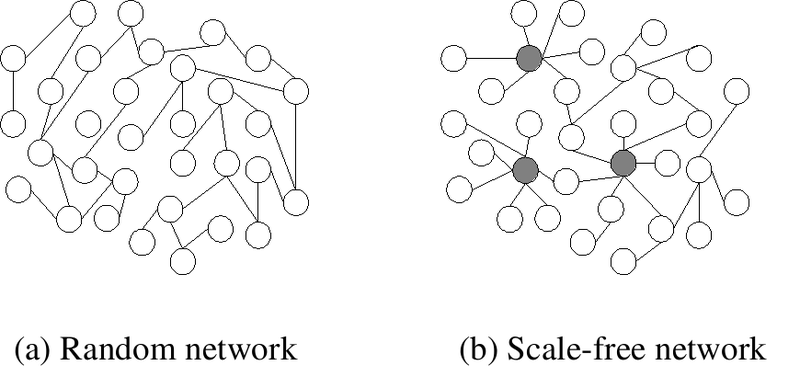
The first step in creating this network is to measure the magnitude of correlation of gene expression between each pair of genes, *xi, xj*, with a Spearman correlation coefficient. A weighted adjacency matrix, *aij*, will be constructed of all the resulting gene correlations by using a power function. The equation below shows this power adjacency function:

*aij = |cor(xi,xj)|β*

The adjacency matrix is what defines the co-expression network and we want the adjacency matrix to be scale-free. A scale-free network asymptotically follows a power law, *P(m) ~ m-γ*. Where *P(m)* is the probability of having *m* connections and γ is a parameter set normally between 2 and 3. In a scale-free network, the probability of a transcript being connected to many other transcripts is low, meaning that hub genes that are connected to many other genes are a rare occurrence. Most observed biological networks have been identified as “scale-free”,3 ­so it is reasonable to believe that our genetic networks should be as well. Figure 3 shows an illustration of a scale-free network vs. a random network. Therefore, we will choose a β parameter such that it is scale-free based on the proposed model-fitting index from Zang and Horvath et al 20053. The index is the coefficient of determination (R2) for the linear model log(P(*k*)) vs. log(*k*), where *k* is connectivity (defined below) and P(*k*) is the probability of *k*. An R2 ≥ 0.8 is considered scale-free. Connectivity (*k*) is defined below:

*ki* =

Figure 3: taken from <http://en.wikipedia.org/wiki/Scale-free_network>. The image below shows a random network (a) and a scale-free network (b). In the (b), the larger hubs are shown in grey.

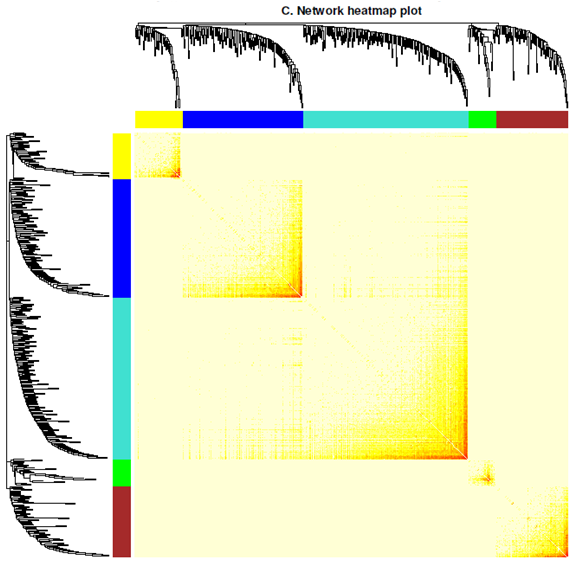
[](http://upload.wikimedia.org/wikipedia/commons/7/77/Scale-free_network_sample.png)

To identify highly co-regulated genes, the adjacency matrix will be transformed into a topological overlap matrix (TOM). This matrix shows interactions between two genes (*i* and *j*) and their indirect interactions by their relationships with all the other genes in the network. The TOM is created by using the following equation:

TOMij =

We define the distances between two genes as DistTOMij = 1 – TOM­ij. Module detection will be made by TOM-based similarity measure coupled with average linkage hierarchical clustering. We will choose a distance criterion to distinguish individual modules. This approach leads to more network findings than compared to 1 – correlation of every pair of genes. We not only determine a gene-to-gene interaction, but also how those two genes relate to the other genes. We can visualize the inter- and intra-connectivity using a network heat map. Figure 3 show an example of this from Langfelder and Horvath as described in their paper detailing the steps involved in performing WGCNA in R7.

Figure 3: Network Heat Map Example from Langfelder and Horvath7



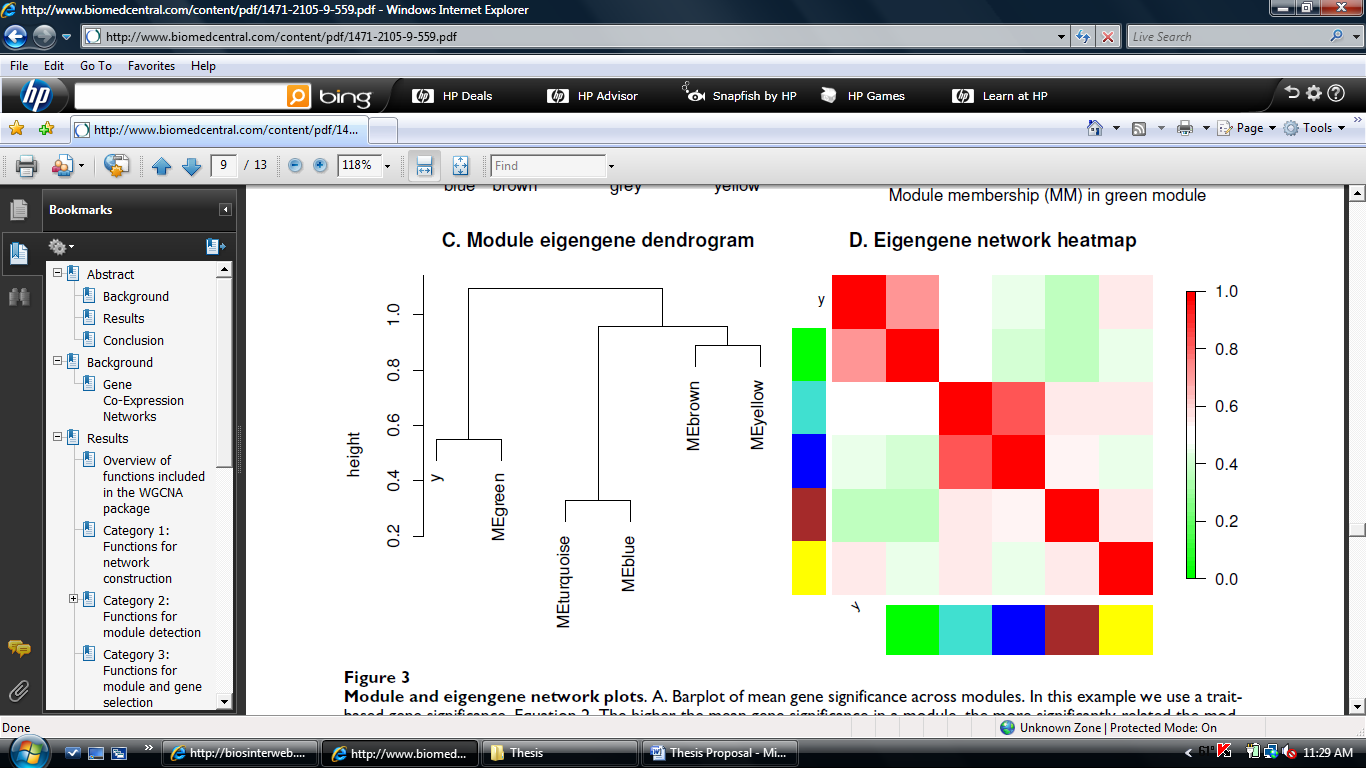
Each branch of the dendrogram and each row/column of the heatmap represents a gene. There are 5 modules in the example above, which are color coded below the dendrogram. The darker yellow/orange color corresponds to a greater interconnectivity. For each module, the “hub” genes with the most intraconnectivity are shown at the tips of the branches because they show the most connectivity with genes within the defined module7. We can also assign a continuous measure of module membership, know in the field as fussy membership. Instead of defining a gene as included or excluded in a module, this measurement can show genes that are near the boundaries or an intermediate between two modules7. Module membership between gene *i* and module *q* is defined as:

*= cor(xi, E(q))*

Where *xi* is defined as the profile for gene *i* and *E(q)*is the eigengene for module *q*7.

After defining modules and eigengenes, we can visualize networks of genes related through coexpression. An example heatmap for eigengenes only is shown below7:

Figure 4: Example Eigengene Network Heatmap7



This heatmap is from Langerfelder and Horvarth and is the resulting eigengene heatmap from the figure 3. We can see the data has been considerably reduced. This will help us with the multiple testing issue that is hard to overcome in microarray studies. We will not get a single gene that is solely responsible for alcoholism/depression endophenotypes (which we do not expect), but we will hopefully get a module/network of genes that are associated with these endophenotypes and explore these genes/pathways further.

Once the eigengenes are identified, we will correlate these “genes” with alcohol/depression endophenotypes using a Pearson correlation. The higher the absolute value of correlation between eigengenes and phenotype, the more evidence that those modules are involved in the etiology of alcoholism and/or depression and thus with further validation provide a genetic roadmap for novel therapeutic treatments and personalized medicine.

To further understand the biological significance of these modules, we can correlate our eigengenes with a single nucleotide polymorphism (SNP) dataset available to us via Wellcome Trust. If there is significant association, we could potentially include certain SNP(s) in our genetic pathway to out endophenotypes by gaining insight on the genetic control of transcription. We will also perform enrichment analyses and can identify gene ontology (GO) terms (which include biological processes, cellular components and molecular functions) and KEGG (Kyoto Encyclopedia of Genes and Genomes)pathways that are over-represented within a module. If this is the case, we can gain insight to biological relevance of the co-expression of these genes.

To conclude, a WGCNA will give us needed power and biological relevance for our study. We also have the ability to determine if significant modules are related to a certain biological system and can gain insight into the big biological picture of disease.

**Constructing a Gene-Coexpression Network Outline**

* Step 1:
  + Correlation matrix
  + Create distance
  + Use topology to declare clusters
* Step 2:
  + Identify eigengenes/“hub” genes
* Step 3:
  + Identify modules associated with alcohol/depression by correlating endophenotypes with eigengenes
* Step 4:
  + Characterize relationships among transcripts within a module using network topological properties such as connectivity
  + Correlate eigengenes with SNP dataset to gain insight into genetic control of transcription
  + Look for over-representation of Gene Ontology terms/KEGG pathways within modules to gain insight into biological relevance

**References**

1. Muhonen L.H., Lahti J., Sinclair D., Lönnqvist J., Alho H. *Substance Abuse Treatment, Prevention, and Policy* 2008, **3**:20
2. <http://www.nature.com/nature/journal/v452/n7186/pdf/nature06758.pdf>
3. Zhang B, Horvath S. A general framework for weighted gene co-expression

network analysis. Stat Appl Genet Mol Biol. 2005

1. [Shakya K](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Shakya%20K%22%5BAuthor%5D), [Ruskin HJ](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Ruskin%20HJ%22%5BAuthor%5D), [Kerr G](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Kerr%20G%22%5BAuthor%5D), [Crane M](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Crane%20M%22%5BAuthor%5D), [Becker J](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Becker%20J%22%5BAuthor%5D). Comparison of microarray preprocessing methods [Adv Exp Med Biol.](javascript:AL_get(this,%20'jour',%20'Adv%20Exp%20Med%20Biol.');) 2010;680:139-47.

# Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind DH. Functional organization of the transcriptome in human brain.

1. Tabakoff B, Saba L, Kechris K, Hu W, Bhave SV, Finn DA, Grahame NJ, Hoffman PL. The genomic determinants of alcohol preference in mice. *Mamm Genome* (2008) 19:352–365
2. Langfelder P, Horvath S.WGCNA: an R package for weighted correlation network analysis**.** *BMC Bioinformatics* 2008, **9**:559
3. Barabási, Albert-László ["Scale-Free Networks"](http://www.nd.edu/~networks/Publication%20Categories/01%20Review%20Articles/ScaleFree_Scientific%20Ameri%20288,%2060-69%20(2003).pdf). *Scientific American*, 288:60-69, May 2003.
4. Johnson WE, Rabinovic A, and Li C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics*
5. Jain S, Heutink P. From Single Genes to Gene Network High-Throughput-High-Content Screening for Neurological Disease. *Neuron Perspective* (2010) 68:207-217.
6. <http://www.niaaa.nih.gov/Pages/default.aspx>
7. Tabakoff B, Saba L, Kechris K, Hu W, Bhave SV, Finn DA, Grahame NJ, Hoffman PL: **The genomic determinants of alcohol preferencein mice.** *Mamm Genome* 2008, **19(5):**352-365.
8. **Evidence of Common and Specific Genetic Effects: Association of the Muscarinic Acetylcholine Receptor M2 (CHRM2) Gene with Alcohol Dependence and Major Depressive Syndrome.** Jen C. Wang et al. in *Human Molecular Genetics,* Vol. 13, No. 17, pages 1903–1911;June 30, 2004.
9. Schadt, E.E. Molecular Networks as Sensors and Drivers of Common Human Diseases. *Nature* 2009, **461**, 218-223.
10. Kalivas PW, O'Brien C: **Drug addiction as a pathology of staged neuroplasticity.** *Neuropsychopharmacology* 2008, **33(1):**166-180.
11. Merikangas, K.R., Leckman, J.F., Prusoff, B.A., Pauls, D.L. and Weissman, M.M. (1985) Familial transmission of depression and alcoholism. Arch. Gen. Psychiat., 42, 367–372.
12. Tabakoff, B., Cornell, N., Hoffman, P.L., 1986. Alcohol tolerance. Ann Emerg Med 15, 1005-1012.
13. Ehlers, C.L., Walter, N.A., Dick, D.M., Buck, K.J., Crabbe, J.C., 2010. A comparison of selected quantitative trait loci associated with alcohol use phenotypes in humans and mouse models. Addict Biol 15, 185-199.

Power point reference: <http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/>