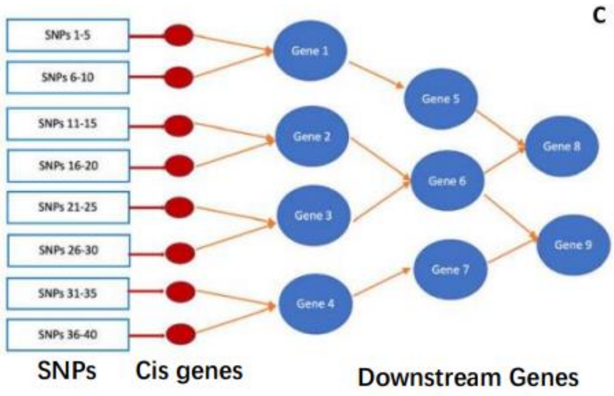
The causal network model includes variants, cis genes and downstream genes. Variants are differences in the DNA sequence among individuals within a species. SNPs are the most common type of genetic variation, where a single nucleotide—A, T, C, or G—in the genome differs between members of a species.



The left column shows groups of SNPs. Each group could represent a set of SNPs that are either closely located or are believed to be linked genetically. Cis Genes are located near the SNP groups and are directly regulated by them. Downstream Genes (trans) are the final targets in the network, whose expression is influenced by the cis genes and trans network structure. Then, the influence of cis genes to trans genes is trans effect times expression level of cis genes.

First, we need to generate genotype matrix G with n individuals and p variants by binomial distribution. Assume G_{n*p} is genotype matrix, $G_i \sim binomial(n, MAF_i)$, where $i = 1, 2, \dots, p$, MAF_i varies from 0.1 to 0.4.

Suppose $E_{cis}(j)$ is cis gene expression of jth cis gene, then

$$E_{cis}(j) = \sum_{i=1}^{p} G_i \cdot \beta_{i,j}^{cis} \cdot A_{i,j}^{cis} + \varepsilon^{cis}$$

where $\beta_{i,j}^{cis}$ is cis effect from ith variants to jth cis genes, $A_{i,j}^{cis}$ is association between ith variant and jth cis genes where $A_{i,j}^{cis} = 1$ when variant i has effect on cis gene j, otherwise $A_{i,j}^{cis} = 0$. $\varepsilon^{cis} \sim N(0, \sigma_{cis}^2)$. According to above plot, $A_{\square}^{cis}$ $(40 * 8)$ is

$$A_{\square}^{cis} = \begin{array}{r} 1-5 \\ 6-10 \\ 11-15 \\ 16-20 \\ 21-25 \\ 26-30 \\ 31-35 \\ 36-40 \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \square & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \square & \square & 1 & 0 & 0 & 0 & 0 & 0 \\ \square & \square & \square & 1 & 0 & 0 & 0 & 0 \\ \square & \square & \square & \square & 1 & 0 & 0 & 0 \\ \square & \square & \vdots & \square & \square & 1 & 0 & 0 \\ \square & \square & \square & \square & \square & \square & 1 & 0 \\ \square & \square & \square & \square & \square & \square & \square & 1 \end{bmatrix}$$

For those trans genes are directly affected by cis genes (such as gene 1-4 in above plot), then gene expression for trans-genes $E_{trans}(j)$ can be expressed by

$$E_{trans}(j) = \sum_{i=1}^{n.cis} E_{cis}(i) \cdot \beta_{i,j}^{cis-trans} \cdot A_{i,j}^{cis-trans} + \varepsilon^{cis-trans}$$

where $\beta_{i,j}^{cis-trans}$ is cis effect from ith cis gene to jth trans genes, $A_{i,j}^{cis-trans}$ is association between ith cis gene and jth trans gene where $A_{i,j}^{cis} = 1$ when cis gene i has effect on trans gene j, otherwise $A_{i,j}^{cis-trans} = 0$. $\varepsilon^{cis-trans} \sim N(0, \sigma_{cis-trans}^{2})$.

$$A_{\square}^{cis-trans} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore,

$$E_{trans}(1) = \sum_{i=1}^{2} E_{cis}(i) \cdot \beta_{i,1}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(2) = \sum_{i=3}^{4} E_{cis}(i) \cdot \beta_{i,2}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(3) = \sum_{i=5}^{6} E_{cis}(i) \cdot \beta_{i,3}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(4) = \sum_{i=7}^{8} E_{cis}(i) \cdot \beta_{i,4}^{cis-trans} + \varepsilon^{cis-trans}$$

For those trans genes are not directly affected by cis genes (such as gene 5-9 in above plot), then gene expression for trans-genes $E_{trans}(j)$ can be expressed by

$$E_{trans}(j) = \sum_{i=1}^{n.trans} E_{trans}(i) \cdot \beta_{i,j}^{trans} \cdot A_{i,j}^{trans} + \varepsilon^{trans}$$

where $A_{\square}^{trans}$ is association between ith trans gene and jth trans gene where $A_{i,j}^{cis} = 1$ when trans gene i has effect on trans gene j, otherwise $A_{i,j}^{cis-trans} = 0$.

$$A_{\square}^{trans} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore,

$$A^{variant-trans} = A_{\square}^{cis} \,\%*\%\, A_{\square}^{cis-trans} \,\%*\%\, A_{\square}^{trans}$$

Therefore,

$$E_{trans}(5) = E_{trans}(1) \cdot \beta_{1,5}^{trans} + \varepsilon^{trans}$$

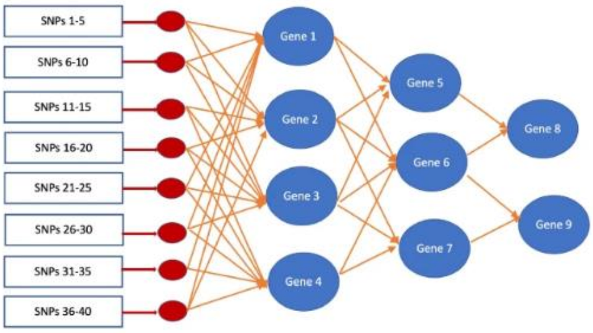
$$E_{trans}(6) = E_{trans}(2) \cdot \beta_{2,6}^{trans} + E_{trans}(3) \cdot \beta_{3,6}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(7) = E_{trans}(4) \cdot \beta_{4,7}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(8) = E_{trans}(5) \cdot \beta_{5,8}^{trans} + E_{trans}(6) \cdot \beta_{6,8}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(9) = E_{trans}(6) \cdot \beta_{6,9}^{trans} + E_{trans}(7) \cdot \beta_{7,9}^{trans} + \varepsilon^{trans}$$

Now, we consider a more complex model



For this model,

$$A_{\square}^{cis} = \begin{array}{r} 1-5 \\ 6-10 \\ 11-15 \\ 16-20 \\ 21-25 \\ 26-30 \\ 31-35 \\ 36-40 \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \square & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \square & \square & 1 & 0 & 0 & 0 & 0 & 0 \\ \square & \square & \square & 1 & 0 & 0 & 0 & 0 \\ \square & \square & \square & \square & 1 & 0 & 0 & 0 \\ \square & \square & \vdots & \square & \square & 1 & 0 & 0 \\ \square & \square & \square & \square & \square & \square & 1 & 0 \\ \square & \square & \square & \square & \square & \square & \square & 1 \end{bmatrix}$$

$$A_{\square}^{cis-trans} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_{\square}^{trans} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Therefore,

$$E_{trans}(1) = \sum_{i=1,2,5,6,7,8}^{\square} E_{cis}(i) \cdot \beta_{i,1}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(2) = \sum_{i=1,2,3,4,8}^{\square} E_{cis}(i) \cdot \beta_{i,2}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(3) = \sum_{i=1,2,3,4,5,6}^{\square} E_{cis}(i) \cdot \beta_{i,3}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(4) = \sum_{i=3,4,5,6,7,8}^{\square} E_{cis}(i) \cdot \beta_{i,4}^{cis-trans} + \varepsilon^{cis-trans}$$

$$E_{trans}(5) = \sum_{i=1,2,3}^{\square} E_{trans}(i) \cdot \beta_{i,5}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(6) = \sum_{i=1,2,3,4}^{\square} E_{trans}(i) \cdot \beta_{i,6}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(7) = \sum_{i=2,3,4}^{\square} E_{trans}(i) \cdot \beta_{i,7}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(8) = \sum_{i=5,6} E_{trans}(i) \cdot \beta_{i,8}^{trans} + \varepsilon^{trans}$$

$$E_{trans}(9) = \sum_{i=6,7} E_{trans}(i) \cdot \beta_{i,9}^{trans} + \varepsilon^{trans}$$

Assume E_{n*g} is gene expression matrix, then

$$E = [E_{trans}(1), E_{trans}(2), \dots, E_{trans}(n.trans)]$$

In this project, we use WGCNA to construct gene co-expression modules (trans genes), where genes are connected through correlations among their residualized expression levels. In our project, we use OED dataset. We used first 1500 genes to generate 20 gene modules. For every 500 genes, we apply WGCNA to them and get the gene modules. Any gene module having more than 200 genes will be eliminated since the correlation between genes in these gene modules are weak.

A network is fully specified by its adjacency matrix a_{ij}, a symmetric n*n matrix with entries in [0, 1]. Assume s_{ij} is gene co-expression similarity measure of a pair of genes i and j. Many co-expression studies use the absolute value of the correlation as an unsigned co-expression similarity measure

$$s_{ij}^{unsigned} = |cor(x_i, x_j)|$$

However, no distinction is made between gene repression and activation if we use unsigned co-expression similarity measure. We can use a simple transformation of the correlation if we want similarity between genes reflects the sign of the correlation

$$s_{ij}^{signed} = 0.5 + 0.5 * cor(x_i, x_j)$$

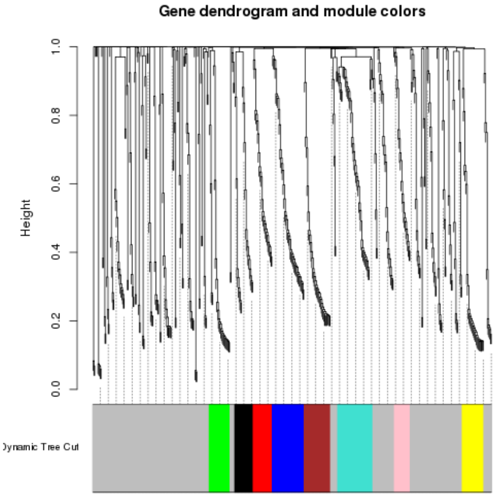
Then, adjacency matrix can be defined by raising the co-expression similarity to a power β

$$a_{ij} = s_{ij}^{\beta}$$

A major step in the module centric analysis is to cluster genes into network modules using a network proximity measure. Typically, WGCNA uses the topological overlap measure (TOM) as proximity.

$$O_{ij} = \frac{\sum_{k=1,k\neq i,j}^{n} a_{ik} * a_{kj} + a_{ij}}{\min\left\{\sum_{k=1,k\neq i}^{n} a_{ik}, \sum_{k=1,k\neq j}^{n} a_{jk}\right\} + 1 - a_{ij}}$$

Once the network has been constructed, we will perform module detection. The default method for module detection is hierarchical clustering. The result of proximity is used as input of hierarchical clustering. Modules are defined as branches of the resulting cluster tree, and branches of the hierarchical clustering tree can be identified by dynamic branch cut method.



Now we have genotype matrix and gene expression matrix, we can calculate Σ_{GE}. Σ_{GE} is the cross-correlation matrix obtained using the *Z* values from the standard trans-eQTL mapping across all pairs of variants and gene-expressions. It is important to adjust for the dependence within the variants and gene expression levels. We can obtain Z matrix of Σ_{GE} by univariate linear regression. transCCA seeks to estimate sparse linear combinations of variants (u) and genes (v) such that the correlation between Gu and Ev is maximized cor(Gu, Ev).

$$(u, v) = \underset{u,v}{\arg max}\{(Ev)^T Gu\}$$

Assume $\Sigma_{GE}(Z)$ is Z matrix of Σ_{GE}. For each column of E, we perform the univariate linear regression separately for each column of G

$$E[,j] \sim \beta_{ij} \cdot G[,i]$$

$$\Sigma_{GE}(Z)[i,j] = \beta_{ij}/se(\beta_{ij})$$

Then we can perform transCCA function using $\Sigma_{GE}(Z)$, p and K.