Tutorials for the WGCNA package for R: WGCNA Background and glossary

Steve Horvath and Peter Langfelder

December 7, 2011

WGCNA begins with the understanding that the information captured by microarray experiments is far richer than a list of differentially expressed genes. Rather, microarray data are more completely represented by considering the relationships between measured transcripts, which can be assessed by pair-wise correlations between gene expression profiles. In most microarray data analyses, however, these relationships go essentially unexplored. WGCNA starts from the level of thousands of genes, identifies clinically interesting gene modules, and finally uses intramodular connectivity, gene significance (e.g. based on the correlation of a gene expression profile with a sample trait) to identify key genes in the disease pathways for further validation. WGCNA alleviates the multiple testing problem inherent in microarray data analysis. Instead of relating thousands of genes to a microarray sample trait, it focuses on the relationship between a few (typically less than 10) modules and the sample trait. Toward this end, it calculates the eigengene significance (correlation between sample trait and eigengene) and the corresponding p-value for each module. The module definition does not make use of a priori defined gene sets. Instead, modules are constructed from the expression data by using hierarchical clustering. Although it is advisable to relate the resulting modules to gene ontology information to assess their biological plausibility, it is not required. Because the modules may correspond to biological pathways, focusing the analysis on intramodular hub genes (or the module eigengenes) amounts to a biologically motivated data reduction scheme. Because the expression profiles of intramodular hub genes are highly correlated, typically dozens of candidate biomarkers result. Although these candidates are statistically equivalent, they may differ in terms of biological plausibility or clinical utility. Gene ontology information can be useful for further prioritizing intramodular hub genes. Examples of biological studies that show the importance of intramodular hub genes can be found reported in [4, 1, 2, 3, 5]. A flow chart of a typical network analysis is shown in Fig. 1. Below we present a short glossary of important network-related terms.

Term	Definition
Co-expression network	We define co-expression networks as undirected, weighted gene networks. The nodes of such a network correspond to gene expression profiles, and edges between genes are determined by the pairwise correlations between gene expressions. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft thresholding), the weighted gene co-expression network construction emphasizes high correlations at the expense of low correlations. Specifically, $a_{ij} = \lvert\mathrm{cor}(x_i, x_j)\rvert^{\beta}$ represents the adjacency of an unsigned network. Optionally, the user can also specify a signed co-expression network where the adjacency is defined as $a_{ij} = \lvert(1 + \mathrm{cor}(x_i, x_j))/2\rvert^{\beta}$.
Module	Modules are clusters of highly interconnected genes. In an unsigned co-expression network, modules correspond to clusters of genes with high absolute correlations. In a signed network, modules correspond to positively correlated genes.
Connectivity	For each gene, the connectivity (also known as degree) is defined as the sum of connection strengths with the other network genes: $k_i = \sum_{u \neq i} a_{ui}$. In co-expression networks, the connectivity measures how correlated a gene is with all other network genes.

Intramodular connectivity k_{IM}	Intramodular connectivity measures how connected, or co-expressed, a given gene is with respect to the genes of a particular module. The intramodular connectivity may be interpreted as a measure of module membership.
Module eigengene E	The module eigengene E is defined as the first principal component of a given module. It can be considered a representative of the gene expression profiles in a module.
Eigengene significance	When a microarray sample trait y is available (e.g. case control status or body weight), one can correlate the module eigengenes with this outcome. The correlation coefficient is referred to as eigengene significance.
Module Membership, also known as eigengene-based connectivity k_{ME}	For each gene, we define a "fuzzy" measure of module membership by correlating its gene expression profile with the module eigengene of a given module. For example, $MM^{blue}(i) = K^{blue}_{cor,i} = \text{cor}(x_i, E^{blue})$ measures how correlated gene i is to the blue module eigengene. $MM^{blue}(i)$ measures the membership of the i-th gene with respect to the blue module. If $MM^{blue}(i)$ is close to 0, the i-th gene is not part of the blue module. On the other hand, if $MM^{blue}(i)$ is close to 1 or -1, it is highly connected to the blue module genes. The sign of module membership encodes whether the gene has a positive or a negative relationship with the blue module eigengene. The module membership measure can be defined for all input genes (irrespective of their original module membership). It turns out that the module membership measure is highly related to the intramodular connectivity k_{IM}. Highly connected intramodular hub genes tend to have high module membership values to the respective module.
Hub gene	This loosely defined term is used as an abbreviation of "highly connected gene." By definition, genes inside co-expression modules tend to have high connectivity.
Gene significance GS	To incorporate external information into the co-expression network, we make use of gene significance measures. Abstractly speaking, the higher the absolute value of GS_i, the more biologically significant is the i-th gene. For example, GS_i could encode pathway membership (e.g. 1 if the gene is a known apoptosis gene and 0 otherwise), knockout essentiality, or the correlation with an external microarray sample trait. A gene significance measure could also be defined by minus log of a p-value. The only requirement is that gene significance of 0 indicates that the gene is not significant with regard to the biological question of interest. The gene significance can take on positive or negative values.
Module significance	Module significance is determined as the average absolute gene significance measure for all genes in a given module. When gene significance is defined as the correlation of gene expression profiles with an external trait y, this measure tends to be highly related to the correlation between the module eigengene and y.

References

[1] Marc R.J. Carlson, Bin Zhang, Zixing Fang, Steve Horvath, Paul S. Mishel, and Stanley F. Nelson. Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. *BMC Genomics*, 7(40), 2006.

[2] Peter S. Gargalovic, Minori Imura, Bin Zhang, Nima M. Gharavi, Michael J. Clark, Joanne Pagnon, Wen-Pin Yang, Aiqing He, Amy Truong, Shilpa Patel, Stanley F. Nelson, Steve Horvath, Judith A. Berliner, Todd G. Kirchgessner, and Aldons J. Lusis. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *PNAS*, 103(34):12741–12746, 2006.

[3] A. Ghazalpour, S. Doss, B. Zhang, C. Plaisier, S. Wang, E.E. Schadt, A. Thomas, T.A. Drake, A.J. Lusis, and S. Horvath. Integrating genetics and network analysis to characterize genes related to mouse weight. *PloS Genetics*, 2(8):e130, 2006.

[4] S. Horvath, B. Zhang, M. Carlson, K.V. Lu, S. Zhu, R.M. Felciano, M.F. Laurance, W. Zhao, Q. Shu, Y. Lee, A.C. Scheck, L.M. Liau, H. Wu, D.H. Geschwind, P.G. Febbo, H.I. Kornblum, T.F. Cloughesy, S.F. Nelson, and P.S. Mischel. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a novel molecular target. *Proc. Natl. Acad. Sci. USA*, 103(46):17402–17407, 2006.

[5] Jeremy A. Miller, Michael C. Oldham, and Daniel H. Geschwind. A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging. *J. Neurosci.*, 28(6):1410–1420, 2008.

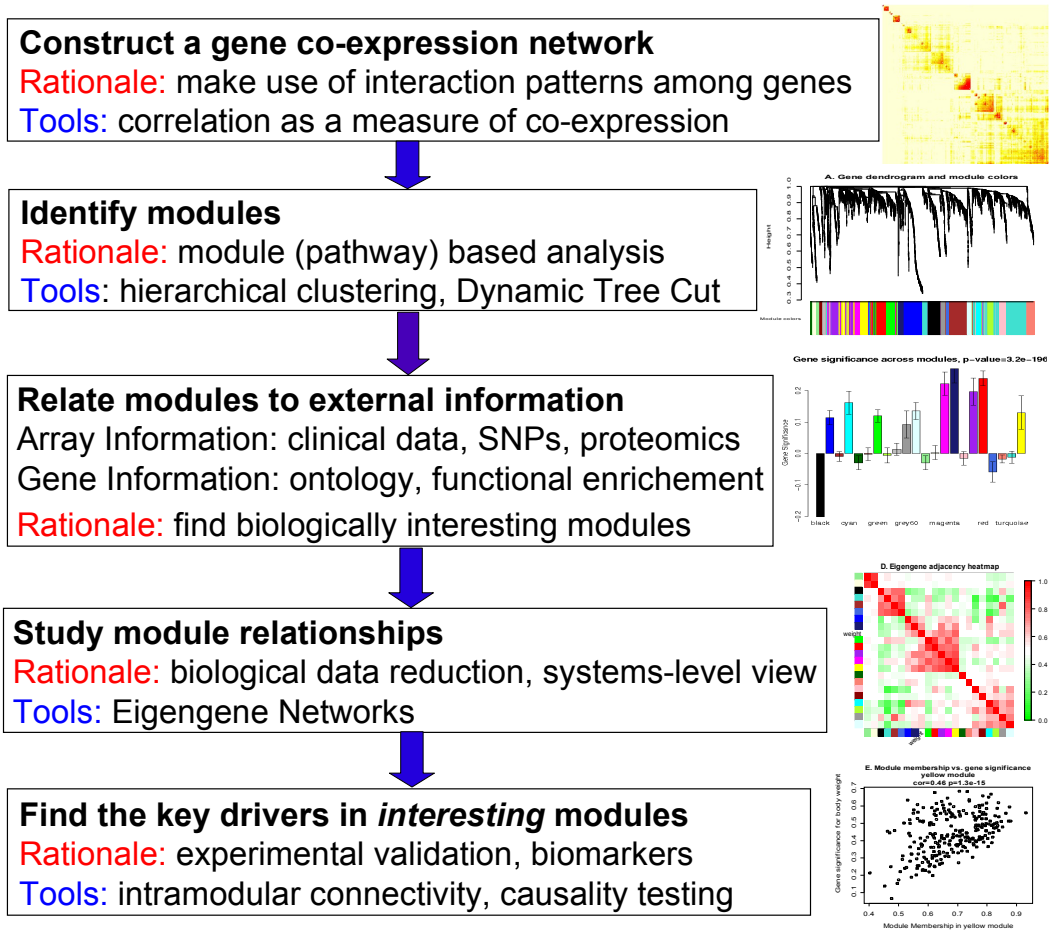

Figure 1: Overview of a typical WGCNA analysis.