SUPPORTING INFORMATION

Workflow for Rapidly Extracting Biological Insights from Complex, Multicondition Proteomics Experiments with WGCNA and PloGO2

Jemma X. Wu[1], Dana Pascovici[1], Yunqi Wu[1], Adam Walker[3,4], Mehdi Mirzaei [1,2,5]*

[1] Australian Proteome Analysis Facility, Macquarie University, Sydney, NSW, Australia

[2] Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia

[3] Neurodegeneration Pathobiology Laboratory, Queensland Brain Institute, The University of Queensland, QLD, Australia

[4] Centre for Motor Neuron Disease Research, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Macquarie University, NSW, Australia

[5] Department of Clinical Sciences, Macquarie University, Sydney, NSW, Australia

* Correspondence: jemma.wu@mq.edu.au.

Table of contents

A short tutorial on using WGCNA for proteomics co-expression analysis

Background

Weighted Gene Correlation Network Analysis (WGCNA) is a method proposed for gene co-expression network modelling and clustering(1, 2). Instead of using the raw pairwise correlation to measure the relationships between genes, it uses a weighted correlation which is modelled as the pairwise correlation raised to a power. For clustering, it uses the topology overlap similarity (TOM) as the distance measurement in the hierarchical clustering algorithm and uses a dynamic tree cutting algorithm to generate an optimal set of clusters. Though it has been widely used in genomics data co-expression analysis, its application in proteomics area is still relatively new. In this short tutorial, we will describe briefly our in-house workflow for using WGCNA for proteomics co-expression analysis and demonstrate it using the rice-leave TMT data published (3). For a detailed WGCNA tutorial, readers can refer to a comprehensive WGCNA online tutorial (4).

Data pre-processing

Before the data goes into the WGCNA workflow, it usually needs to be pre-processed. There are many different techniques used for quantitative proteomics, each of which have their own recommended pre-processing; for instance, in our lab we routinely use label free data with spectral counting quantitation via NSAF (5), label free SWATH data (6), and labelled data such as TMT (7) and iTRAQ (8) – each have slightly different pre-processing and handling, and some have different normalisation approaches that are specific to each. Even for the same technique the data pre-processing procedure can be different under various contexts. For example, for SWATH-MS, if the peptide spectra library used for data extraction is big and noisy, then additional FDR-based peptide filtering is recommended and can greatly improve the data quality (9). As an aside, some recent public datasets make available side by side proteomics and transcriptomics datasets from the same

patients, which can be used to compare the numerical characteristics of protein versus transcript data (10).

Usually the raw proteomics data will be searched and extracted using vendor-specific commercial software, for example, PeakView (Sciex) for SWATH data and ProteomeDiscover (Thermo) for TMT data. Such software often has data filtering features which can be used to remove low abundance and low confidence data and perform FDR control. The resulting data will be combined into a matrix of quantitation with proteins (rows) and samples (columns). Proteins with missing values can be removed or imputed at pre-processing. WGCNA can also work with data with missing values. For both datasets we used in this paper, proteins with missing values were removed.

Data normalisation

The pre-processed quantitative proteomics data needs to be properly normalised before executing WGNCA analysis. There are many different types of data normalisation, and the aim is to remove technical difference as much as possible while preserving the real biological variability. Naturally each normalisation method in effect changes the data, so special consideration has to be given to this process.

Normalisation for data within a single batch aims to minimize the difference such as arising from slightly different column amount or other mass spectrometer-related difference. The common types normalisation in this case include total area normalisation and median normalisation. In total area normalisation, the value of each individual protein is divided by the total protein amount for the sample, in order to account for perhaps unequal loading. In median normalisation, the value of each individual protein is divided by the median protein amount for the sample, and where thus the assumption is that most protein should not differ in abundance across samples. In both cases, we prefer to keep the data on the same scale as the raw data, hence scale by Total/max(Total) or Median/max(Median), respectively.

If the data comes from multiple batches (such as several TMT runs or SWATH batches), additional normalisation steps may be needed to align abundance across batches for each individual protein. Batch normalisation methods such as IRS (11) can be used, for instance for data coming from different TMT runs. Usually data quality checking (such as boxplots, PCA, correlation and CV if technical replicates are present) is performed before and after the data normalisation to make sure the data quality is suitable for analysis.

For the label-free TMT rice leaf ratio dataset, we applied median normalisation for each sample; for the SWATH plasma dataset, we first aggregated the peptide peak areas to proteins and then applied total area normalisation for each sample.

After the normalisation, the protein expression was log-transformed before the analysis was performed.

WGCNA steps and parameters

The WGCNA workflow starts after the data quality checking, normalisation and log-transforming of protein abundance or ratio matrix. The dataset can include all proteins from the experiments or only a subset such as the differentially expressed proteins. The data can include missing values and WGCNA can check (function *goodSamplesGenes*) and remove proteins or samples with too many missing values. Figure S1 shows our generic workflow for WGCNA analysis.

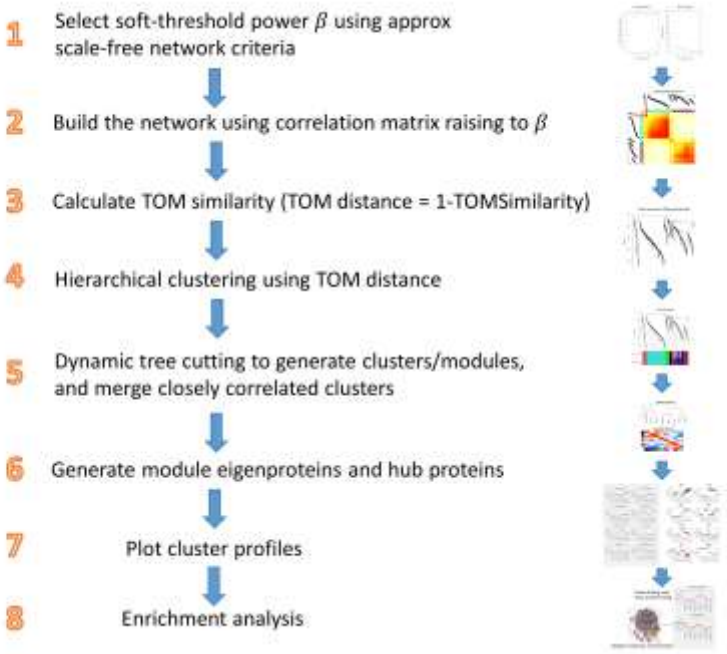

Figure S1. WGCNA workflow

Firstly, a soft threshold is selected for constructing the weighted correlation matrix by using the approximate scale-free network criteria. A scale free network is one where the topology is dominated by a few highly connected nodes, which link the rest of the less connected nodes to the system(2); it is assumed that most biologically relevant networks should satisfy this property. By raising the correlation to the selected soft threshold, the correlation network becomes scale-free. A parameter *RsquaredCut,* which indicates the cut-off value for the correlation R square*,* can be used to adjust the soft threshold selection. The default value is 0.85 in the WGCNA R package(1). WGCNA can build signed or unsigned network adjacency matrix and the default is *unsigned* which transforms all correlations to positive values. However, we recommend that *signed* network adjacency matrix should be used for proteomics datasets so that proteins with different regulation trends will be clustered separately.

Then the TOM distance is calculated from the network adjacency; the aim of the TOM distance is to calculate clusters which are tightly correlated to each other. Hierarchical clustering is performed based on the TOM distance by using average linkage, and then a set of clusters are obtained by using

the dynamic tree cutting method. The number of clusters can be adjusted by changing the parameter value of *minClusterSize*, which controls the minimum number of proteins in each cluster. The optimal value for *minClusterSize* depends on the number of proteins in the dataset. Based on our experience, for small to medium proteomics datasets (for example few hundreds to a thousand), a value between 20 to 30 could generate a good number of well separated clusters. For relatively large dataset (say a few thousand proteins), the value might need to be increased to around 50. For the analysis of both our datasets, the value of *minClusterSize* was set as 20. An automatic cluster merging function is invoked to merge the closely correlated clusters from the dynamic tree cutting. A parameter *cutHeight* can be used to adjust the merging criterion. For both our datasets, we set it as 0.1.

An eigenprotein is generated for each cluster, as the first principle component by using the singular value decomposition. The eigenprotein can be viewed as a representative protein for a cluster, though it is not one of the actual proteins but a "virtual protein", a linear combination of the proteins in the cluster with certain coefficients.

An kME value is then computed for each protein and each cluster, as the correlation between each protein and the eigenprotein from its respective cluster. The kME value can be seen as a measurement for the intra-module connectivity(1); the higher the kME the more connectivity the protein has to other proteins in its cluster. The top few proteins (for example 6 in (12), 10 in other publications) with the highest kME in each cluster can be regarded as hub proteins for the cluster. A hub protein (or gene) is a "loosely defined" term to describe the proteins (genes) that are highly connected(1). Due to this loose definition, a strict cut-off value for selecting hub proteins is not recommended. However, our results show that the hub proteins can provide useful guide for selecting candidate proteins for validation.

Example outputs

Each of the steps described above will output some visual images or tables. Here we use our main dataset results as an example to illustrate them.

Selection of soft threshold power

Scale-free topology fitting (R square) shows this dataset has a good scale-free topology fit (R square > 0.85). Figure S2 shows the change of the soft threshold fit and the mean connectivity as the power changes. There is a trade-off between the scale-free topology fit and the mean connectivity of the network. The soft-threshold power selected was 12.

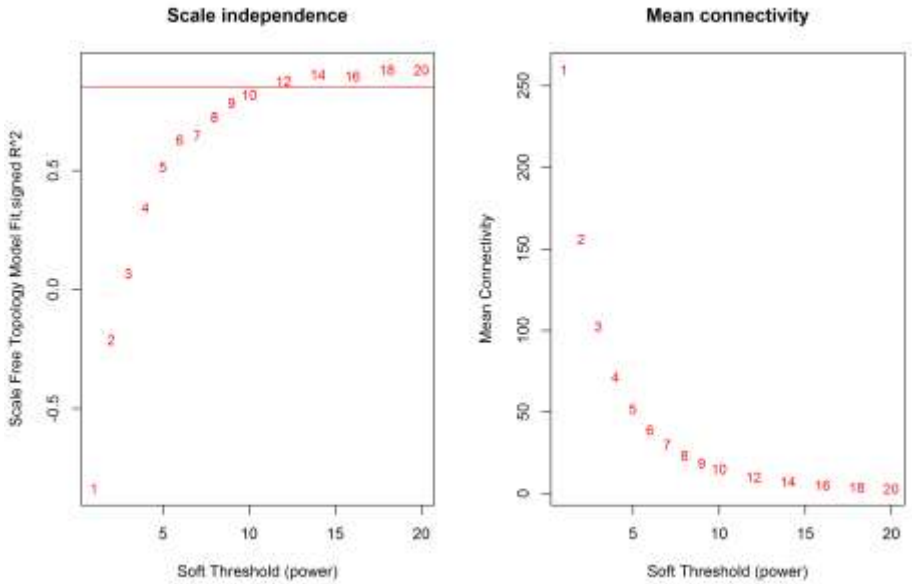

Figure S2. Soft threshold and scale-free topology fit. Left, scale-free topology fitting (R square) vs soft threshold. The red horizontal line represents the 0.85 default cut-off value. Right, the mean connectivity (number of connected nodes) vs soft threshold.

Cluster generation

Seven clusters (or modules) were produced after the dynamic tree cutting and cluster merging. Figure S3 shows the cluster dendrogram and network heatmap for those clusters.

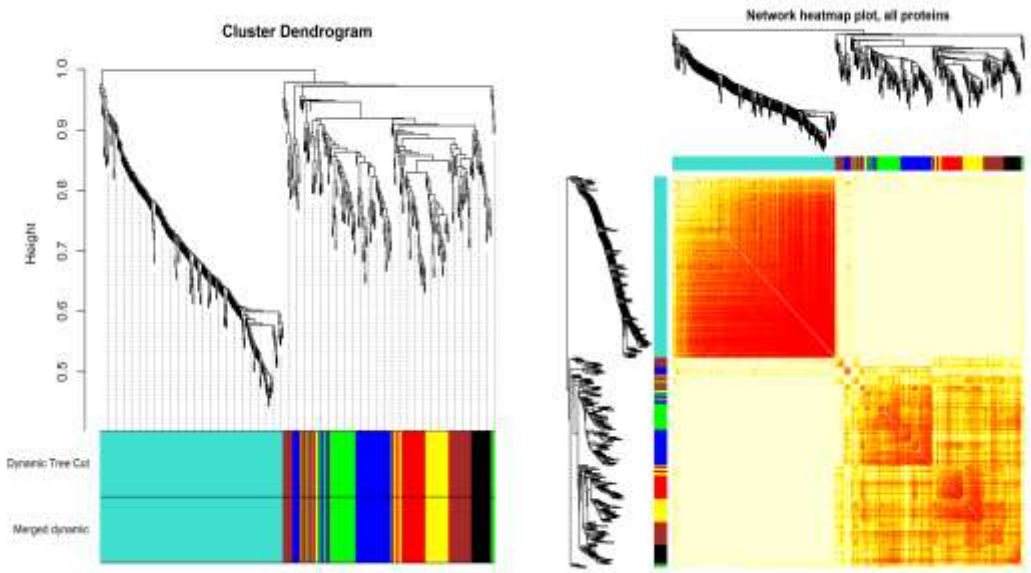

Figure S3. Cluster dendrogram and network heatmap.

The top colour row in the left panel of Figure S3 represents the clusters (modules) generated by using dynamic tree cut, and the bottom colour row represents the clusters generated by merging closely correlated modules. In the network heatmap (right panel), each cell represents the TOM similarity (topology overlap) of two proteins. The more red the cell the higher topology overlap they share in the network. The blue and green module are shown to have high TOM similarity.

Besides the default WGCNA plots, we added a set of additional cluster expression profile plots, each of which shows the abundance changing patterns with the experimental conditions/groups. Figure S4 shows the cluster expression profile for the seven clusters. Each grey line represents a protein abundance and the bold color line represent the average abundance of all proteins in that cluster; some proteins with high fold changes are visible.

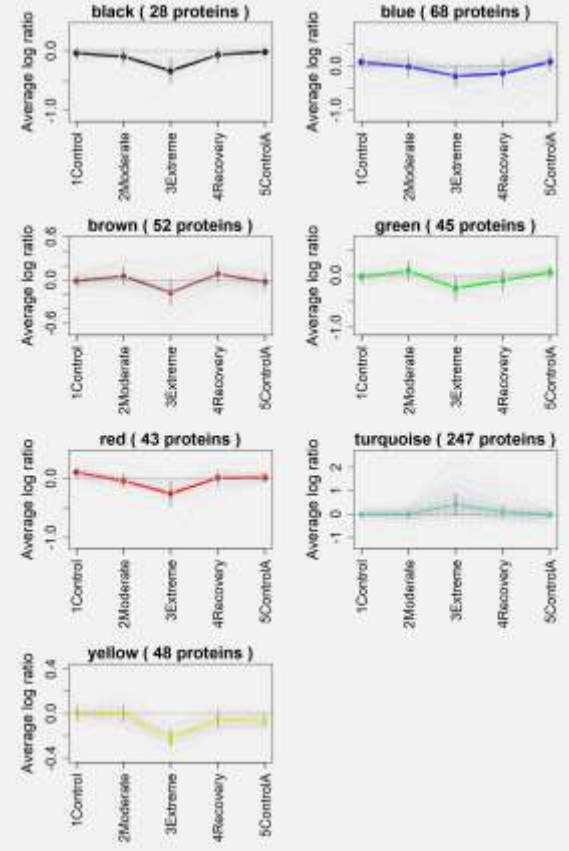

Figure S4. Cluster expression profile. each grey line represents one protein and the thick coloured line represents the average for all.

Eigenproteins

Figure S5 shows the eigenprotein dendrogram, heatmap and boxplots. The more red the colour, the more closely correlated they are. The plots in the left panel are from the WGCNA package, and the additional figure in the right panel capture via boxplots the expression pattern of each individual eigenprotein across experimental groups.

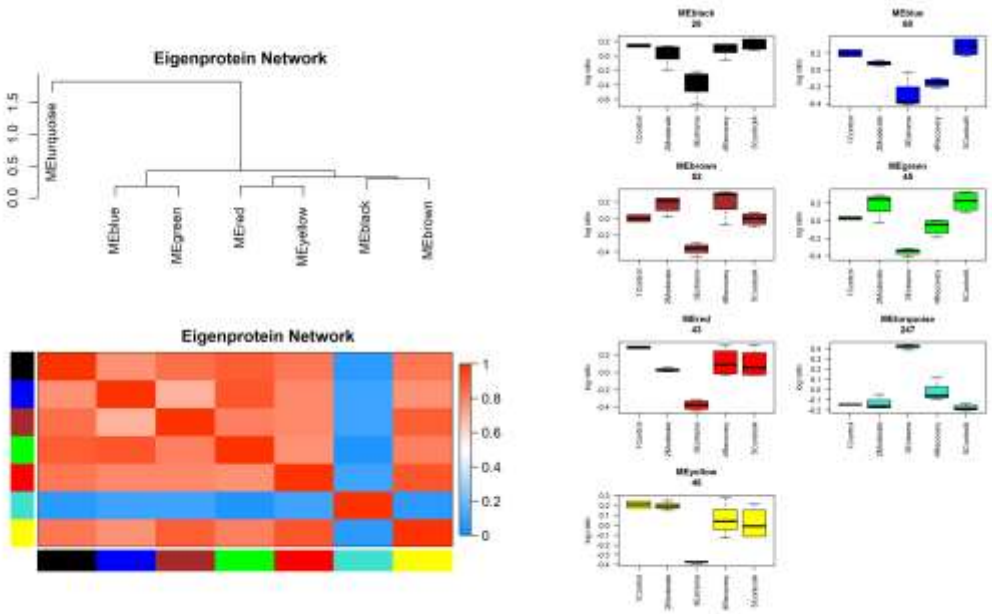

Figure S5. Eigenprotein dendrogram, heatmap and boxplots.

Hub proteins

The proteins in each cluster are ordered by their kME values decreasingly, and the proteins at the top of the list can be viewed as the hub proteins. Figure S6 shows an example of the boxplots the top 6 hub proteins in the red cluster. The top hub protein (Q53RM0 – a magnesium chelatase subunit) was identified in the original publication (3) as one of the relevant enzymes in the chlorophyll biosynthesis pathway.

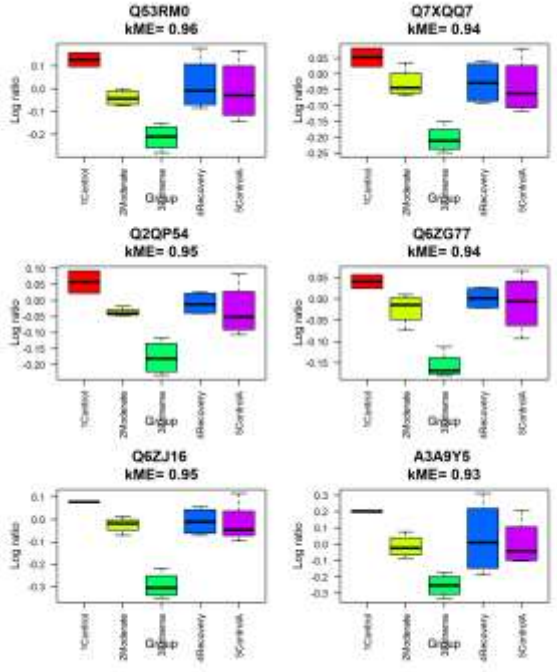

Figure S6. Top 6 hub proteins for the red cluster.

Discussion

The WGCNA workflow can be applied to either all proteins quantitative data or a subset of the data such as the differentially expressed proteins. In proteomics studies, it is common that statistical analysis is performed first to identify differentially expressed proteins (via ANOVA or t-tests or other methodology) which is then followed by unsupervised clustering and functional analysis. In our experience, performing WGCNA analysis on all proteins (usually a few thousands) often gives a large number of clusters among which the patterns are less informative or harder to interpret than on differentially expressed proteins. However, we have performed the WGCNA analysis on both all proteins and differentially expressed proteins of our main dataset, the TMT rice leaves, and provided both results on our Github repository (https://github.com/APAFbioinformatics/PloGO2_R_Package). The clusters obtained with differentially expressed proteins have (not surprisingly) clearer cluster profile patterns. However, the

results obtained using the full dataset include a grey module (proteins unassigned to any module, which may be interesting in their own right) while the results from the DEP dataset does not.

Hub proteins are proteins that are highly connected within a module. Intromodular connectivity kME can be used as a measurement. However, due to its loose definition, it is hard to give a strict cut-off value for selecting hub proteins, and various papers use the top 10, top 6, or other kME cut-off thresholds. The ranking provided by kME can give useful guide for selecting candidate proteins for validation. In addition to kME measurement, correlation with biological traits, if exists, can be another indicator(13).

Reference

1. Langfelder, P.; Horvath, S., WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **2008,** 9, (1), 559.
2. Zhang, B.; Horvath, S., A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **2005,** 4, (1).
3. Wu, Y.; Mirzaei, M.; Pascovici, D.; Chick, J. M.; Atwell, B. J.; Haynes, P. A., Quantitative proteomic analysis of two different rice varieties reveals that drought tolerance is correlated with reduced abundance of photosynthetic machinery and increased abundance of ClpD1 protease. *Journal of Proteomics* **2016,** 143, 73-82.
4. Horvath, P. L. a. S. Tutorials for the WGCNA package. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/
5. Neilson, K. A.; Keighley, T.; Pascovici, D.; Cooke, B.; Haynes, P. A., Label-free quantitative shotgun proteomics using normalized spectral abundance factors. In *Proteomics for Biomarker Discovery*, Springer: 2013; pp 205-222.
6. Wu, J. X.; Song, X.; Pascovici, D.; Zaw, T.; Care, N.; Krisp, C.; Molloy, M. P., SWATH mass spectrometry performance using extended peptide MS/MS assay libraries. *Molecular & Cellular Proteomics* **2016,** 15, (7), 2501-2514.
7. Mirzaei, M.; Pascovici, D.; Wu, J. X.; Chick, J.; Wu, Y.; Cooke, B.; Haynes, P.; Molloy, M. P., TMT one-stop shop: from reliable sample preparation to computational analysis platform. In *Proteome Bioinformatics*, Springer: 2017; pp 45-66.
8. Pascovici, D.; Song, X.; Wu, J.; Zaw, T.; Molloy, M., Practical Integration of Multi-Run iTRAQ Data. In *Mass Spectrometry of Proteins*, Springer: 2019; pp 199-215.
9. Wu, J. X.; Pascovici, D.; Ignjatovic, V.; Song, X.; Krisp, C.; Molloy, M. P. J. P., Improving Protein Detection Confidence Using SWATH‐Mass Spectrometry with Large Peptide Reference Libraries. **2017,** 17, (19), 1700174.
10. Boja, E.; Težak, Ž.; Zhang, B.; Wang, P.; Johanson, E.; Hinton, D.; Rodriguez, H. J. N. m., Right data for right patient—a precisionFDA NCI–CPTAC Multi-omics Mislabeling Challenge. **2018,** 24, (9), 1301-1302.
11. Plubell, D. L.; Wilmarth, P. A.; Zhao, Y.; Fenton, A. M.; Minnier, J.; Reddy, A. P.; Klimek, J.; Yang, X.; David, L. L.; Pamir, N., Extended multiplexing of tandem mass tags (TMT) labeling reveals age and high fat diet specific proteome changes in mouse epididymal adipose tissue. *Molecular & Cellular Proteomics* **2017,** 16, (5), 873-890.

12. Umoh, M. E.; Dammer, E. B.; Dai, J.; Duong, D. M.; Lah, J. J.; Levey, A. I.; Gearing, M.; Glass, J. D.; Seyfried, N. T., A proteomic network approach across the ALS‐FTD disease spectrum resolves clinical phenotypes and genetic vulnerability in human brain. *EMBO molecular medicine* **2018,** 10, (1), 48-62.
13. Yuan, L.; Chen, L.; Qian, K.; Qian, G.; Wu, C.-L.; Wang, X.; Xiao, Y., Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). *Genomics data* **2017,** 14, 132-140.