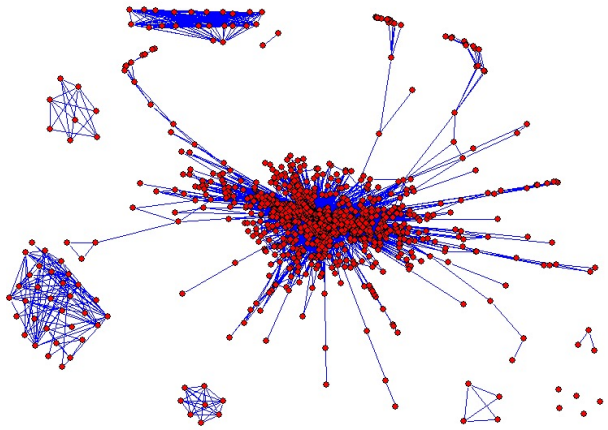


An Overview of Gene Co-Expression Network Analysis

Cheng He

04/27/2021

Bioinformatics Applications (PLPTH813)

Contents

- Introduction of gene co-expression network (GCN)
- Rationales of gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

Contents

- Introduction of gene co-expression network (GCN)
- Rationales of a gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

Does this map tell you which cities are important?

This one does!

The nodes with the largest number of links (connections) are most important!

**Slide courtesy of Paul Mischel and AL Barabasi

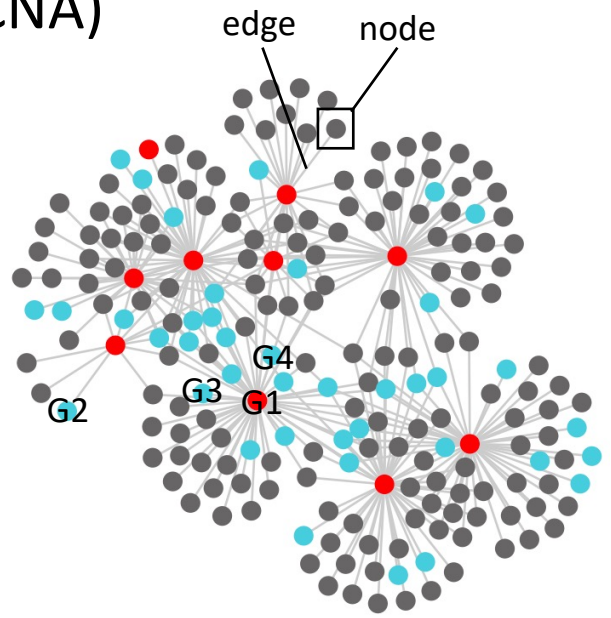
Gene co-expression network

- In gene co-expression networks, each gene corresponds to a node
- Two genes are connected by an edge if their expression values are highly correlated
- Definition of ”high” correlation can be decided by different methods:
  (1) Statistical significance (Pearson correlation)
  (2) Scale free topology criterion (WGCNA)

Gene expression matrix

Gene_ID	sample_1	sample_2	.....
G1	4	2	xx
G2	0	3	xx
G3	6	3	xx
G4	10	5	xx
⋮	xx	xx	xx
	xx	xx	xx

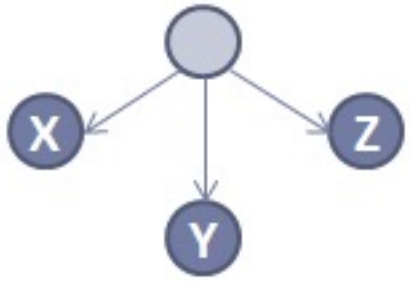
Gene co-expression network analysis →



A **gene co-expression network (GCN)** is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them.

(Stuart, Joshua M; Segal, Eran; Koller, Daphne; Kim, Stuart K (2003). "A gene co-expression network for global discovery of conserved genetic modules". *Science*. **302** (5643): 249–55.)

Undirected graph



Directed graph

Applications of Gene co-expression network

- Finding clusters or modules of highly relevant genes.
- Summarizing such modules using an intramodular hub gene or the module eigengene.
- Relating modules to one another and to external sample traits with eigengene network methodology.
- Calculating the measures of module membership.
- Gene Network Reverse Engineering.
- Plant Biology - Co-expression analyses have been extensively used to search for novel genes involved in specific plant pathways.

Contents

- Introduction of gene co-expression network (GCN)
- Rationales of gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

Gene co-expression similarity measure

- Pearson correlation was used to measure the similarity of 2 gene expressions (s_{ij})

Unsigned network: No gene repression and activation

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

Signed network: Consider gene repression and activation

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

G1 activate G2: $s_{12}^{signed} = 0.5 + 0.5*1 = 1$

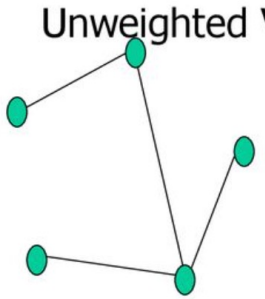
G1 repress G2: $s_{12}^{signed} = 0.5 + 0.5*(-1) = 0$

Network=Adjacency Matrix

- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.

Unweighted network:

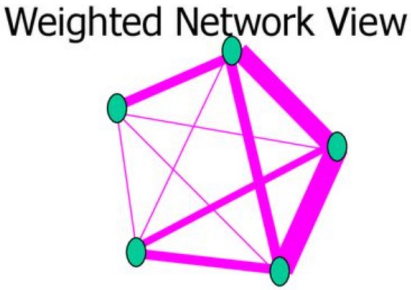
- Entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected)



Some genes are connected
All connections are equal

★ Weighted network:

- The adjacency matrix reports the connection strength between gene pairs



•All genes are connected
•Connection Widths=Connection strenghts

Generalized Connectivity

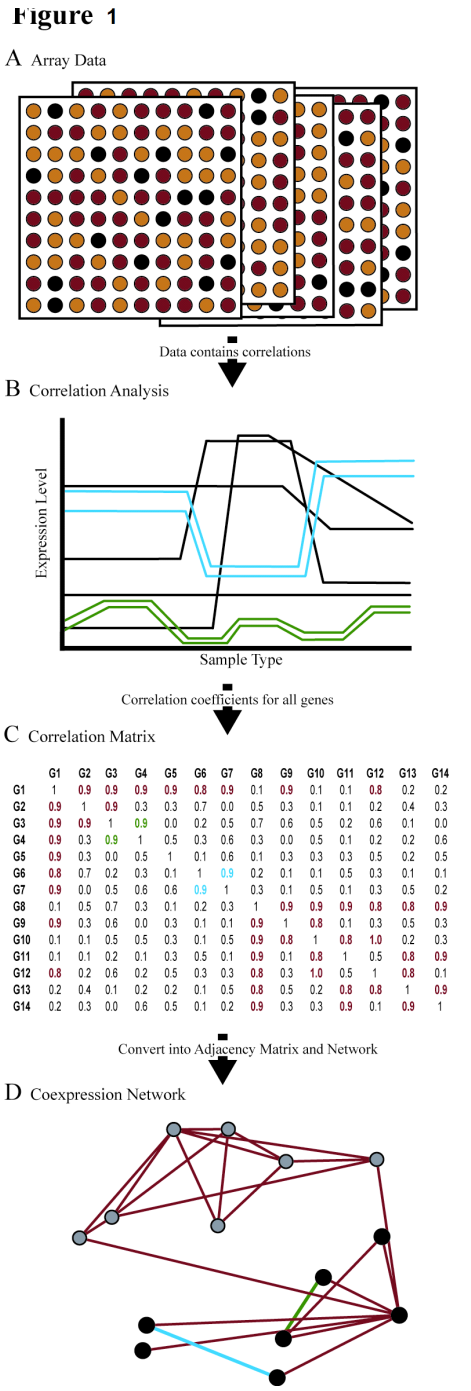
- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks = number of direct neighbors
  - For weighted networks = sum of connection strengths to other nodes

$$k_i = \sum_j a_{ij}$$

k_i : connectivity of gene i

a_{ij} : adjacency matrix between gene i and gene j

High gene connectivity usually indicates a hub gene



Steps for constructing a co-expression network

A) Microarray/RNA-Seq gene expression data

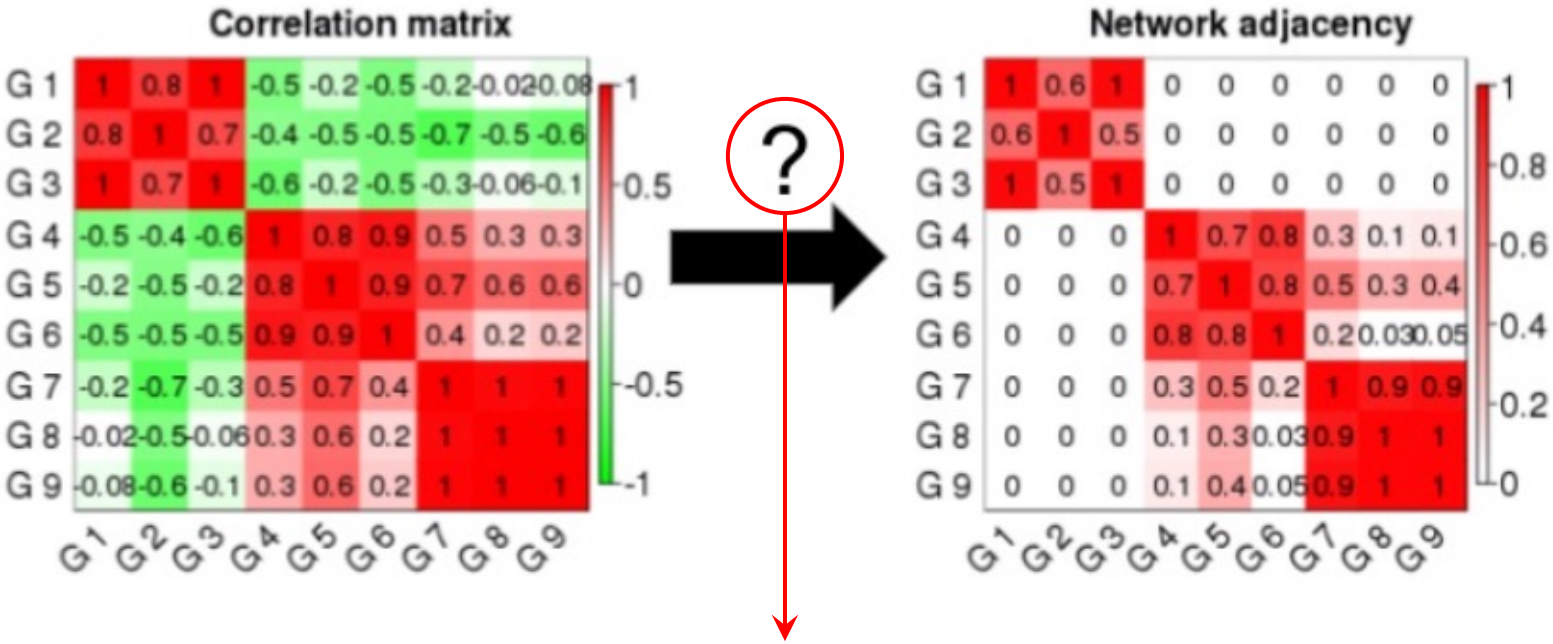
B) Measure similarity of gene expression with a Pearson correlation

C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network

Or transformed continuously with the power adjacency function → weighted network

D) The gene connectivity can be calculated from the adjacency matrix

How to turn correlation matrix to network adjacency is the key to construct a good gene co-expression network



Gene co-expression network analysis methods

Contents

- Introduction of gene co-expression network (GCN)
- Rationales of gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

Current Methods

WGCNA based methods

- Weighted Gene Co-expression Network Analysis (Horvath 2008)
- DiffCoEx (Bruno 2010)

Other methods

- THD-Module Extractor (Kakati 2016)

Weighted Gene Co-expression Network Analysis (WGCNA)

Bin Zhang and Steve Horvath

A General Framework for Weighted Gene Co-Expression Network Analysis

De Gruyter | Published online: August 12, 2005
DOI: https://doi.org/10.2202/1544-6115.1128

Software | Open Access | Published: 29 December 2008

WGCNA: an R package for weighted correlation network analysis

Peter Langfelder & Steve Horvath

BMC Bioinformatics **9**, Article number: 559 (2008) | Cite this article

Now WGCNA has been the most popular and reliable gene co-expression network analysis method

Procedures:

Step 1: Measure network adjacency by the power function

$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

a_{ij}: the adjacency matrix between gene i and j

β: soft thresholding parameter

Step 2: Cluster genes into network modules using TOM method

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

k_i: the connectivity of gene i

$$k_i = \sum_{u \neq i} a_{iu}$$

Notes:

1. Often β=6 or β=12 works well but in general WGCNA uses the “scale free topology criterion”

2. The topological overlap measure (TOM) combines the adjacency of two genes and the connection strengths these two genes share with other "third party" genes.

Advantages of WGCNA:

- It retains the connectivity of network nodes
- It has powerful analytical efficiency
- Standard data mining methods such as clustering analysis results can be transformed into weight networks (RNA-Seq)

Disadvantages of WGCNA:

- WGCNA doesn’t support dataset < 15 samples
- WGCNA takes many parameters with default values to control network construction and distinct module extraction
- WGCNA doesn’t assume scale-free topology for differentially expressed genes

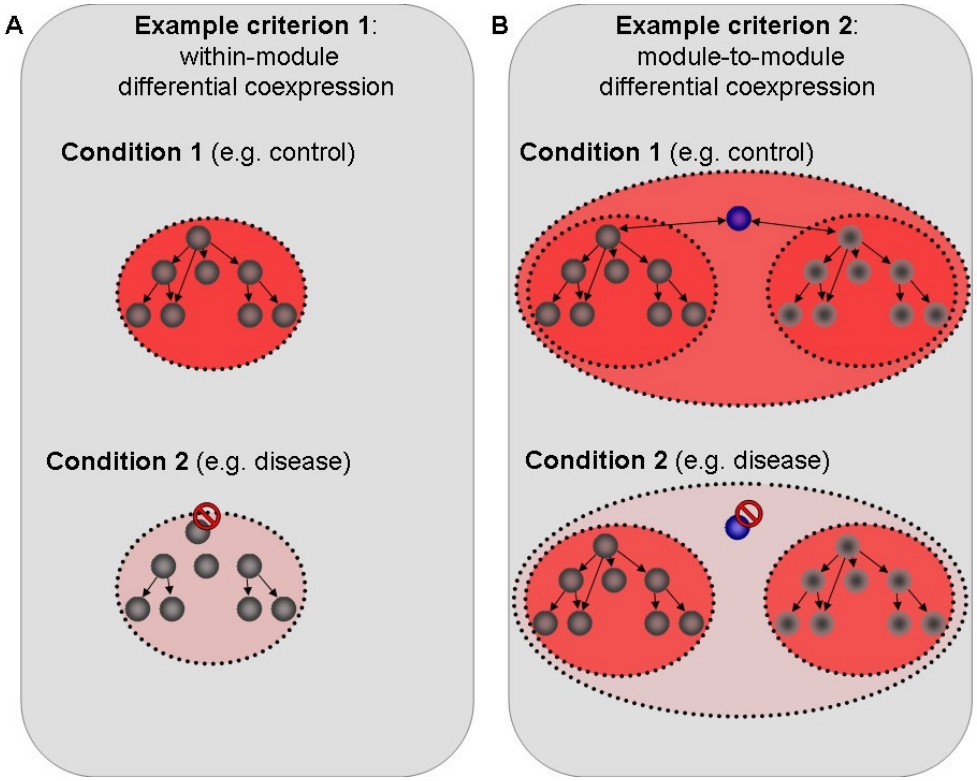
DiffCoEx

Methodology article | Open Access | Published: 06 October 2010

DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules

Bruno M Tesson, Rainer Breitling & Ritsert C Jansen

BMC Bioinformatics **11**, Article number: 497 (2010) | Cite this article



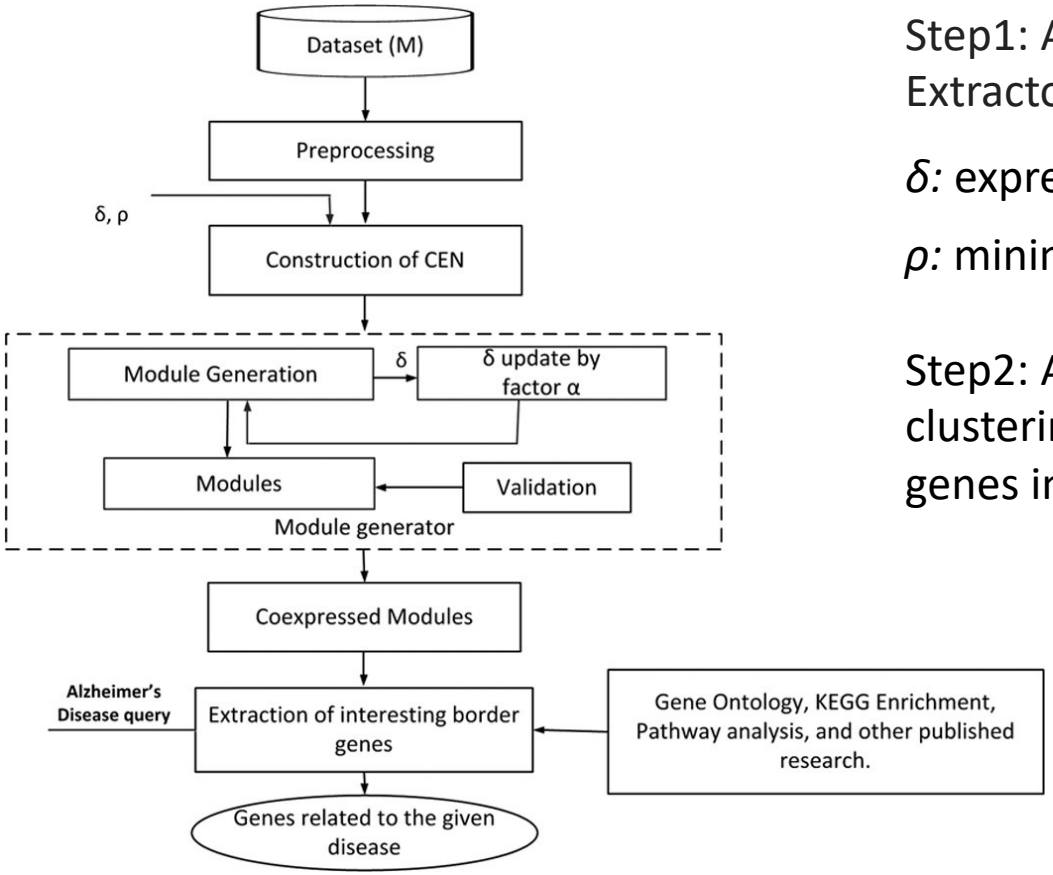
DiffCoEx can identify gene co-expression differences between multiple conditions.

THD-Module Extractor

THD-Module Extractor: An Application for CEN Module Extraction and Interesting Gene Identification for Alzheimer's Disease

Tulika Kakati, Hirak Kashyap & Dhruba K. Bhattacharyya

Scientific Reports **6**, Article number: 38046 (2016) | Cite this article



Step1: A schematic diagram of THD-Module Extractor framework.

δ: expression similarity threshold

ρ: minimum neighborhood threshold

Step2: A Shifting-and-scaling correlation based clustering algorithm (SSSim) is used to cluster genes into modules

Advantages of THD-Module Extractor:

- THD-Module Extractor considers the important issue of analysis of CEN using both gene expression similarity and semantic similarity

Semantic similarity between two genes can be measured on the basis of information content (*Ic*)

$$Ic = -ln(P(t))$$

$$P(t) = \frac{number\ of\ annotations\ involving\ a\ GO\ term\ 't'}{number\ of\ genes}$$

Comparison of four CEN module extraction techniques

Method	Working principle	Measure used in CEN construction	Parameters	Language	Datasets used
WGCNA	Hierarchical clustering	Dissimilarity measure based on TOM	Soft threshold parameter and tree cutting threshold	R	Microarray and RNA-Seq
DiffCoEx	Differential Co-expression analysis	Dissimilarity measure based on TOM	Soft threshold parameter and tree cutting threshold	R	Microarray
THD-Module Extractor	Density-based clustering	SSSim	Expression similarity threshold (δ), minimum neighborhood threshold (ρ), and δ updating factor α	R and MATLAB	Microarray

Kakati, T., Bhattacharyya, D.K., Barah, P. and Kalita, J.K., 2019. Comparison of methods for differential co-expression analysis for disease biomarker prediction. *Computers in biology and medicine*, *113*, p.103380.

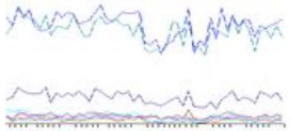
Contents

- Introduction of gene co-expression network (GCN)
- Rationales of gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

General Framework for WGCNA

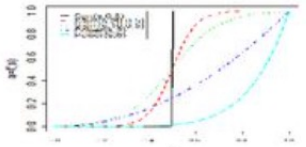
(Pearson correlation)

Define a Gene Co-expression Similarity



(scale-free topology criterion)

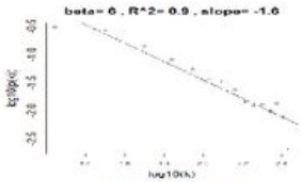
Define a soft thresholding power (β)



β

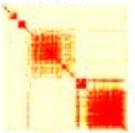
$a_{ij} = |cor(x_i, x_j)|^{\beta}$

Calculate Adjacency Matrix with β



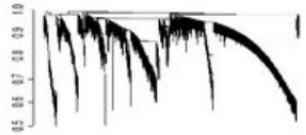
$DistTOM_{ij} = 1 - TOM_{ij}$

Define a Measure of Node Dissimilarity

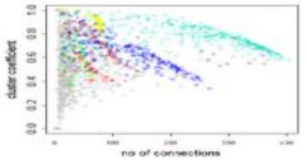


(hierarchical clustering)

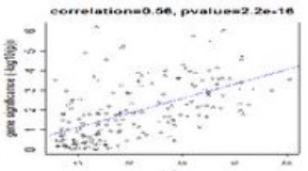
Identify Network Modules (Clustering)



Visualization of Co-expression network



Relate the Network Concepts to External Gene or Sample Information



Define a Gene Co-expression Similarity

	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

Gene expression values

$|r(G_i, G_j)|$

Pearson correlation

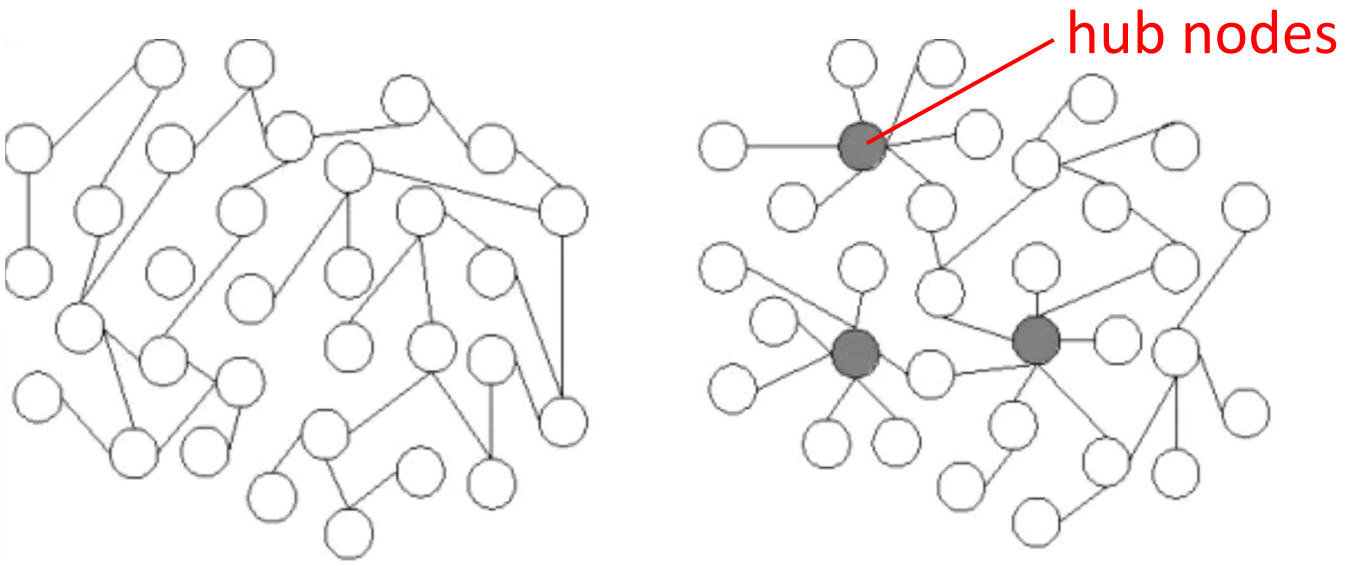
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score

Microarray or RNA-Seq

$$a_{ij} = |cor(x_i, x_j)|^{\beta}$$

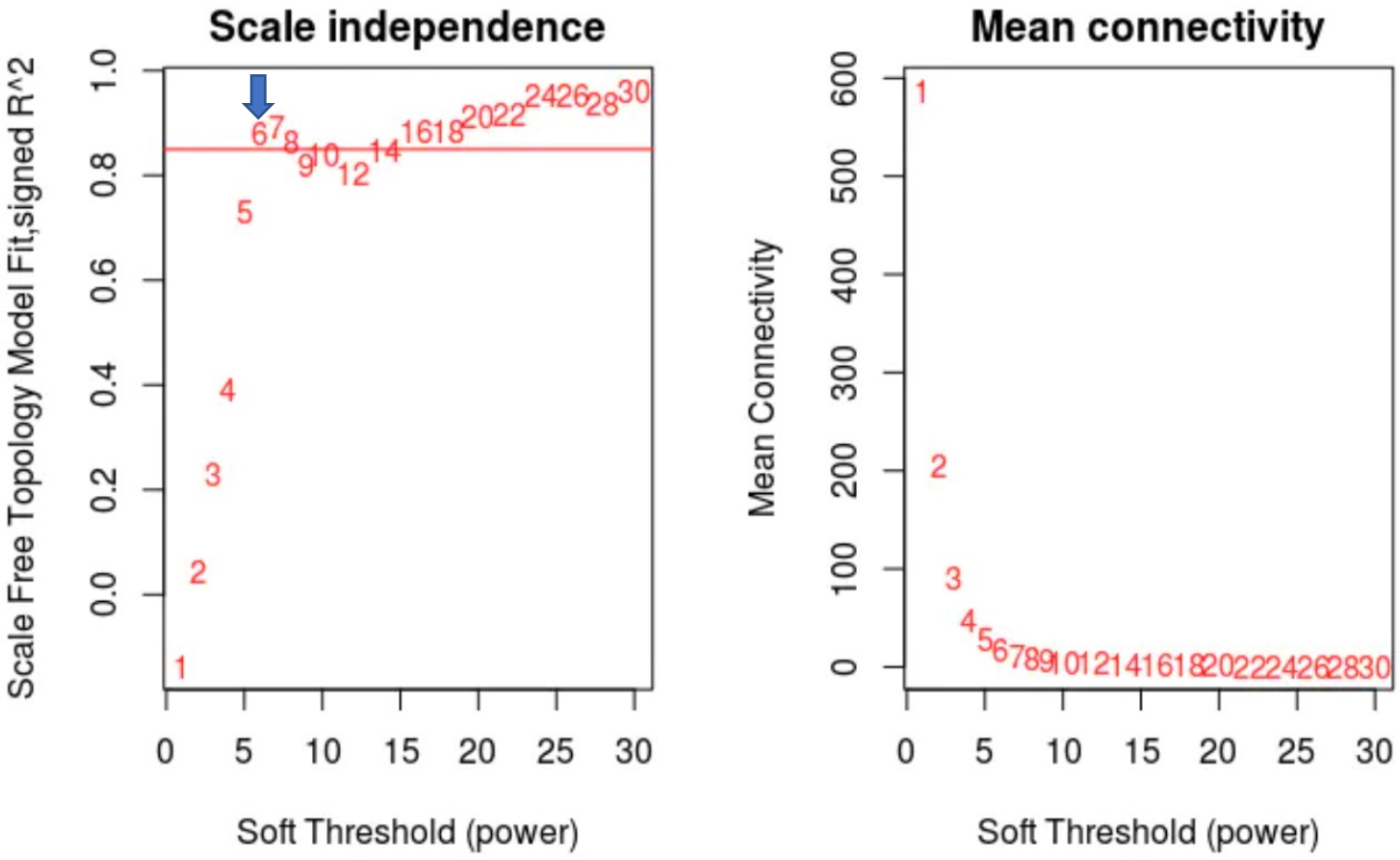
Define a soft thresholding power (β)



(a) Random network (b) Scale-free network

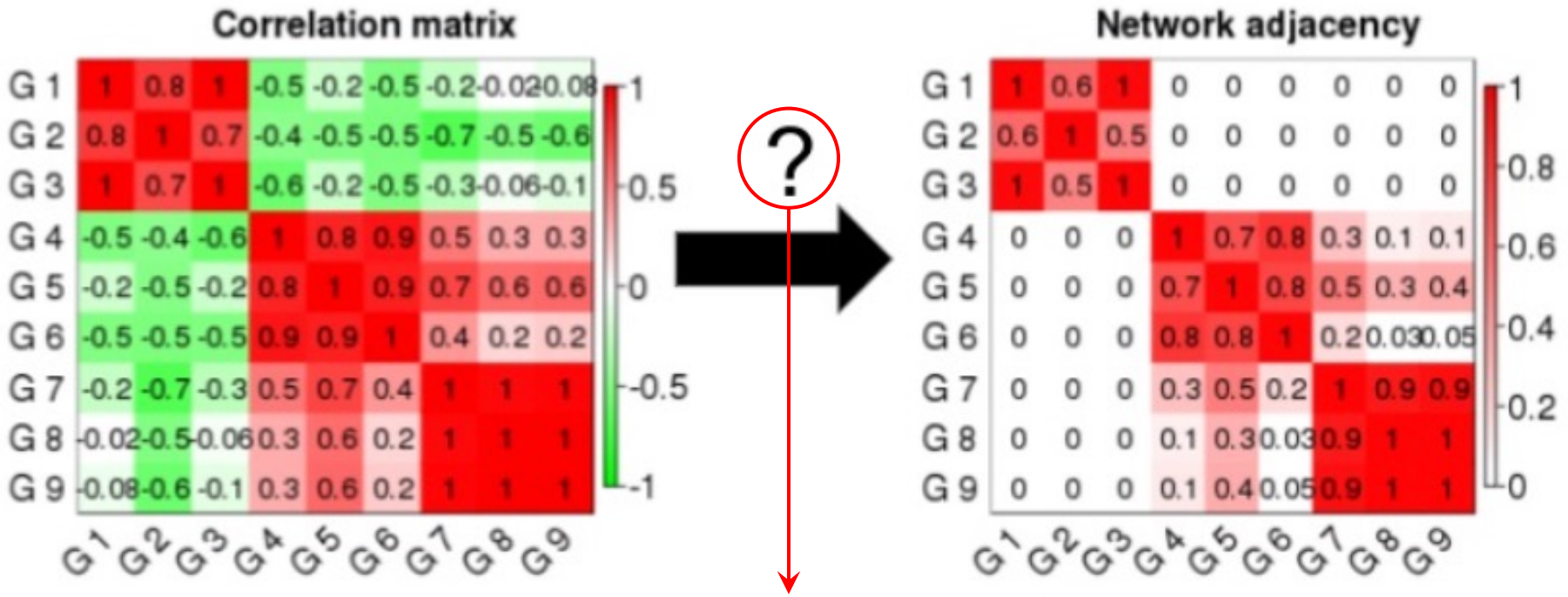
- A scale-free network is a network whose degree distribution follows a power law.
- Unlike random network, some nodes in scale-free network have many more connections than others, which is called “hubs”
- Real-world networks are often claimed to be scale free

scale-free topology criterion



We need to find a best β that leads to a network that satisfies scale free topology most (R^2). Consequently, the lowest β value that leads to a R^2 of 0.8 (or the highest possible model fit) was used.

Calculate Adjacency Matrix with β



$$a_{ij} = | cor(x_i, x_j) |^{\beta}$$

a_{ij} : adjacency matrix between gene i and gene j

k_i : connectivity of gene i

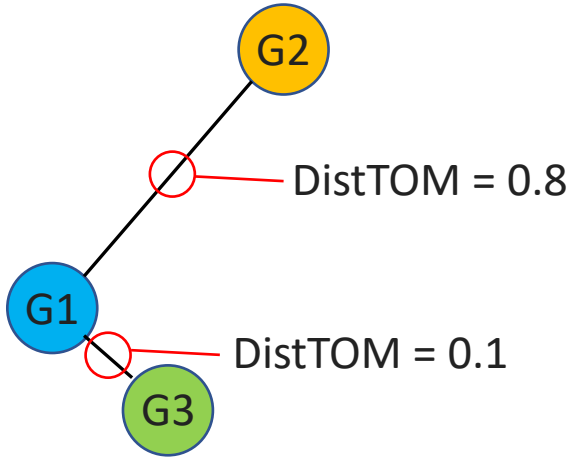
$$k_i = \sum_j a_{ij}$$

Define a Measure of Mode Dissimilarity

Topological Overlap leads to a network distance measure (Ravasz et al., 2002)

$$TOM_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

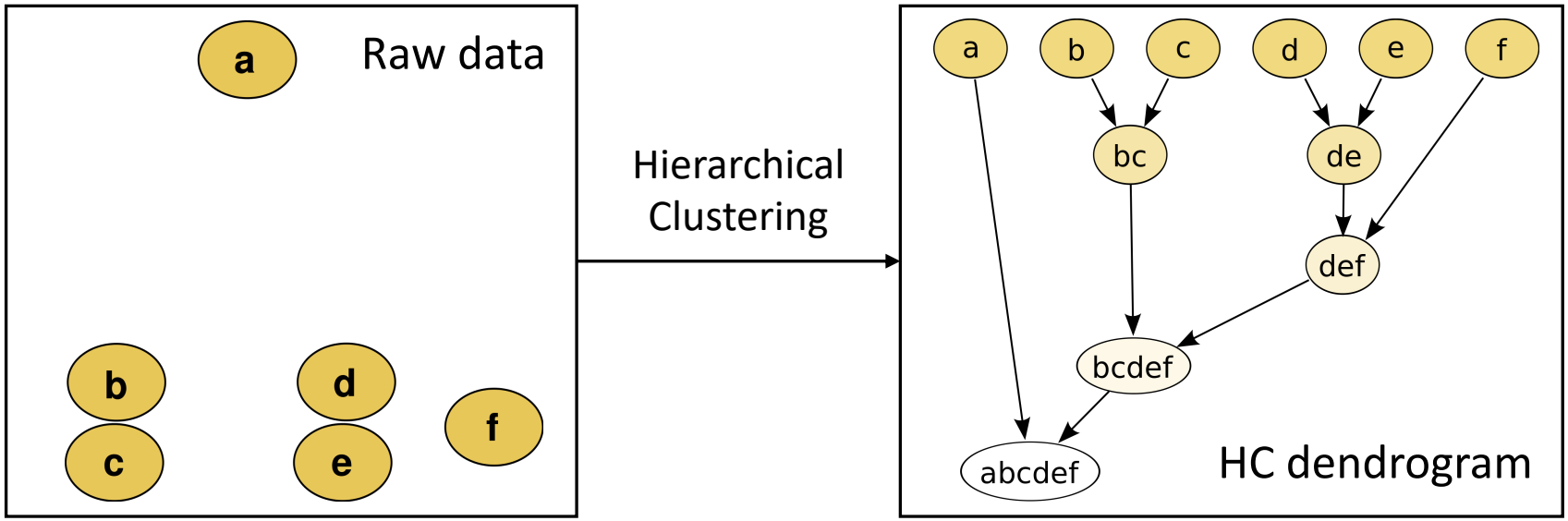
$$DistTOM_{ij} = 1 - TOM_{ij}$$



- The TOM describes how well connected the genes are in respect of how many neighbors they share
- The TOM matrix was transformed to a dissimilarity TOM (distTOM, 1-TOM), where a high connectivity produces a low number and no connectivity gives us a value of 1, or close.

Identify Network Modules

Hierarchical Clustering (HC) is a method of cluster analysis which seeks to build a hierarchy of clusters (Rokach 2005. 321-352.)



(https://en.wikipedia.org/wiki/Hierarchical_clustering)

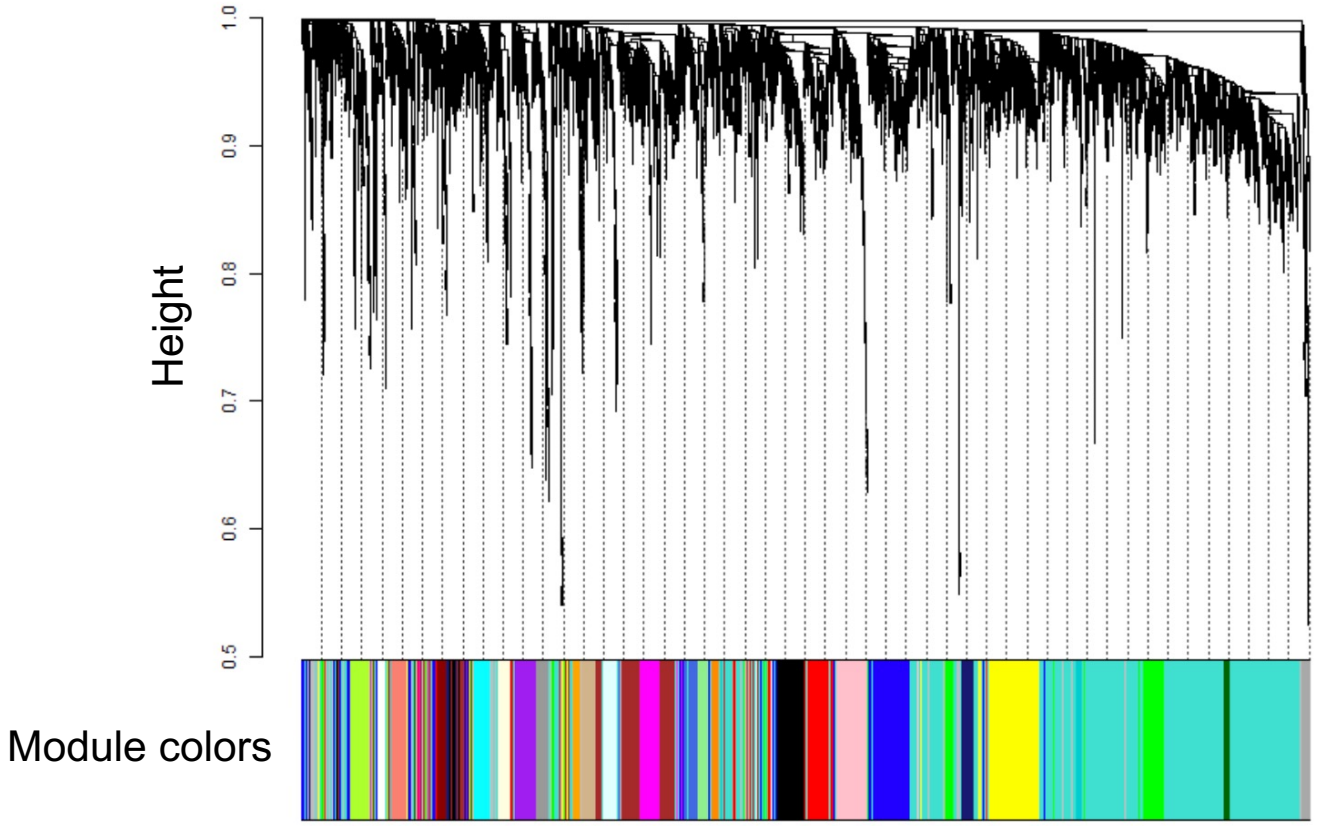
Euclidean distance: $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$

a_i : *DistTOM* of gene a

b_i : *DistTOM* of gene b

Clustering of the genes in the distTOM matrix produces clusters of genes sharing many common neighbors into the same module.

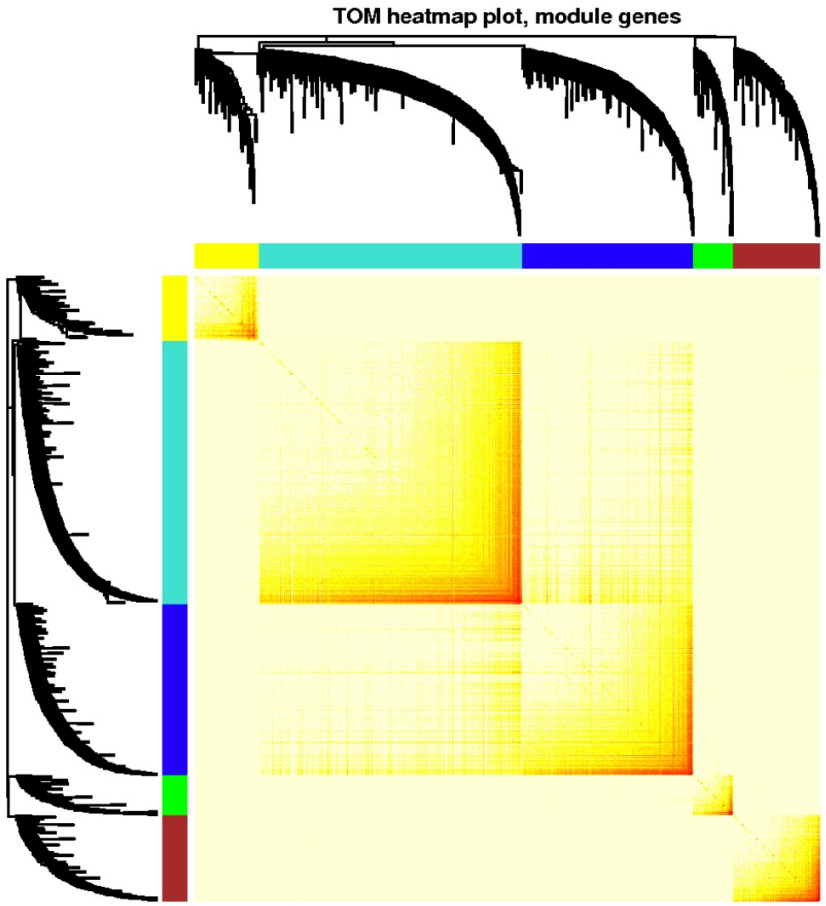
Hierarchical Cluster Dendrogram



The 1st principal component for each co-expression module was defined as module **eigengene**, which represents the whole gene expression patterns for each module

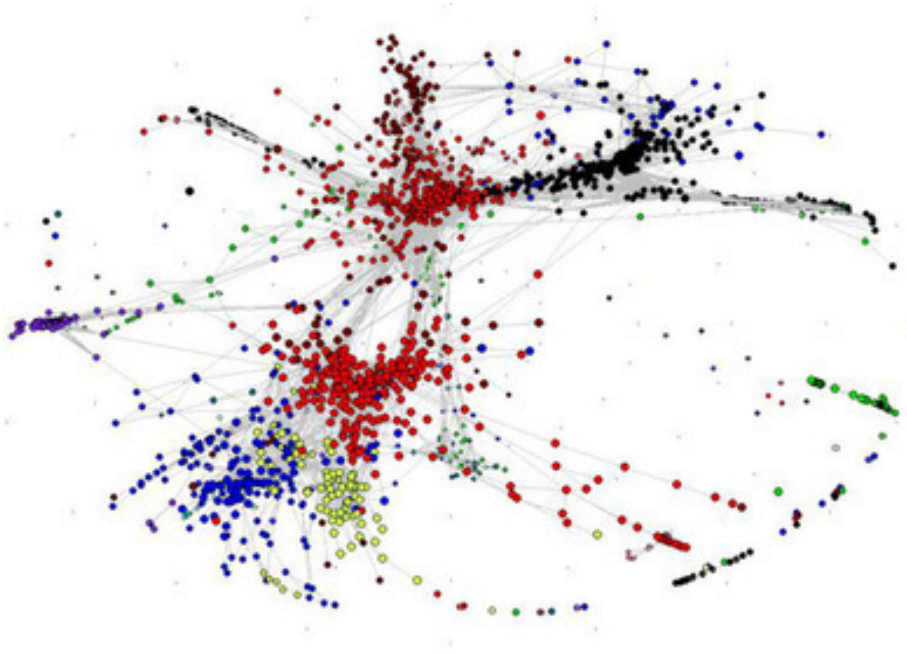
Visualization of Co-expression network

Heatmap plot of the TOM



(https://Horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/)

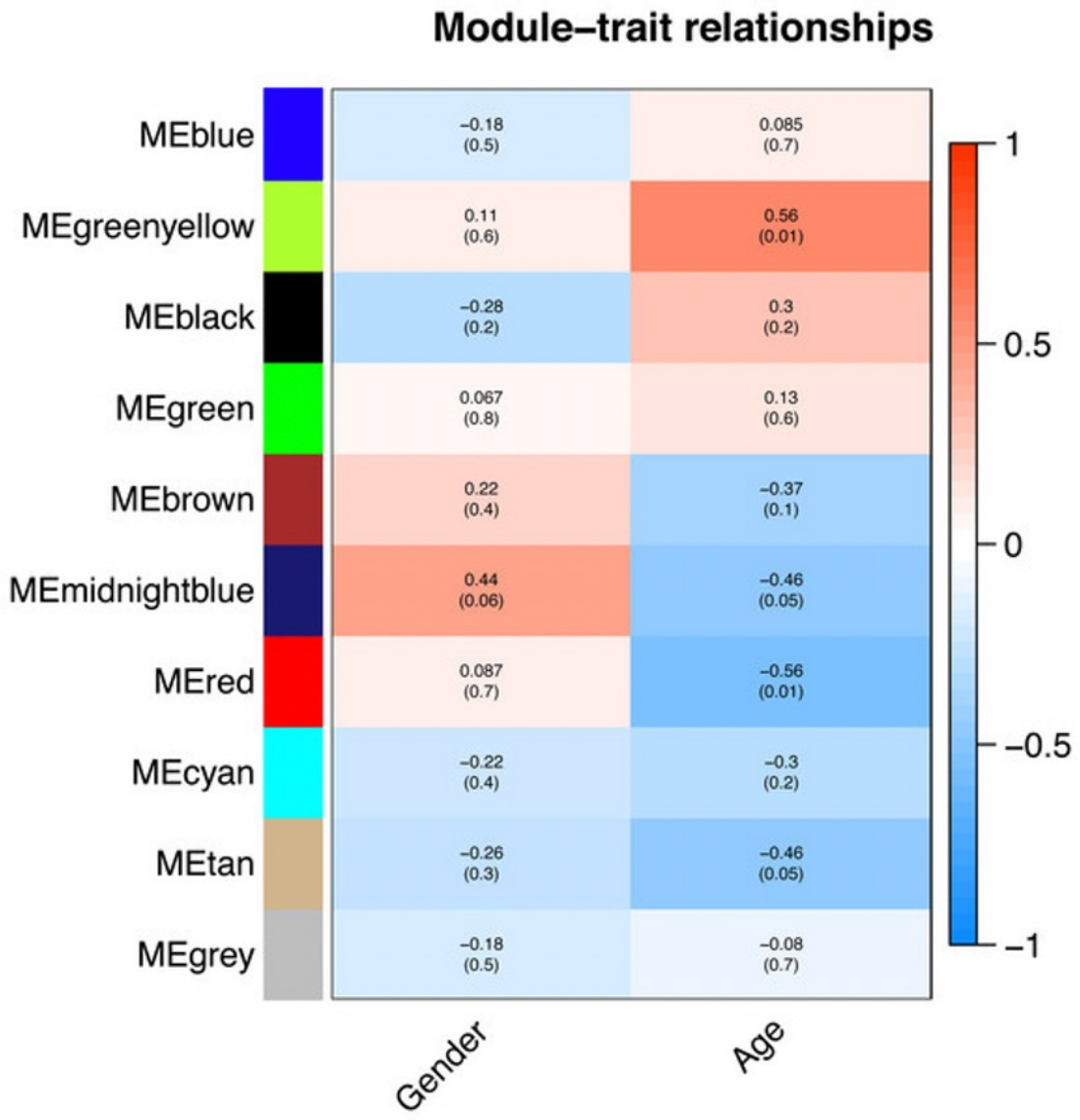
Network Visualization



Cytoscape (https://cytoscape.org)

Gephi (https://gephi.org/)

Relate the Network Concepts to External Information



- Module-trait relationships analysis
- GO terms analysis of genes in specific modules
- Functional analysis of hub genes
- …….

Contents

- Introduction of gene co-expression network (GCN)
- Steps to construct a gene co-expression network
- Gene co-expression network analysis methods
- Weighted Gene Co-Expression Network Analysis (WGCNA)
- Tutorial: Run WGCNA on your own laptop

Software requirement

R project or R Studio (https://www.rstudio.com/)

Gephi (https://gephi.org/)

Tutorial input data

Gene expressions of 56 breast cancer cell lines (GSE48213_gene_expression_fpkm.txt)

Traits of 56 breast cancer cell lines (GSE48213_trait.txt)

Tutorial codes

WGCNA_GSE48213_coexp.R

(github)

1. Install & load WGCNA package

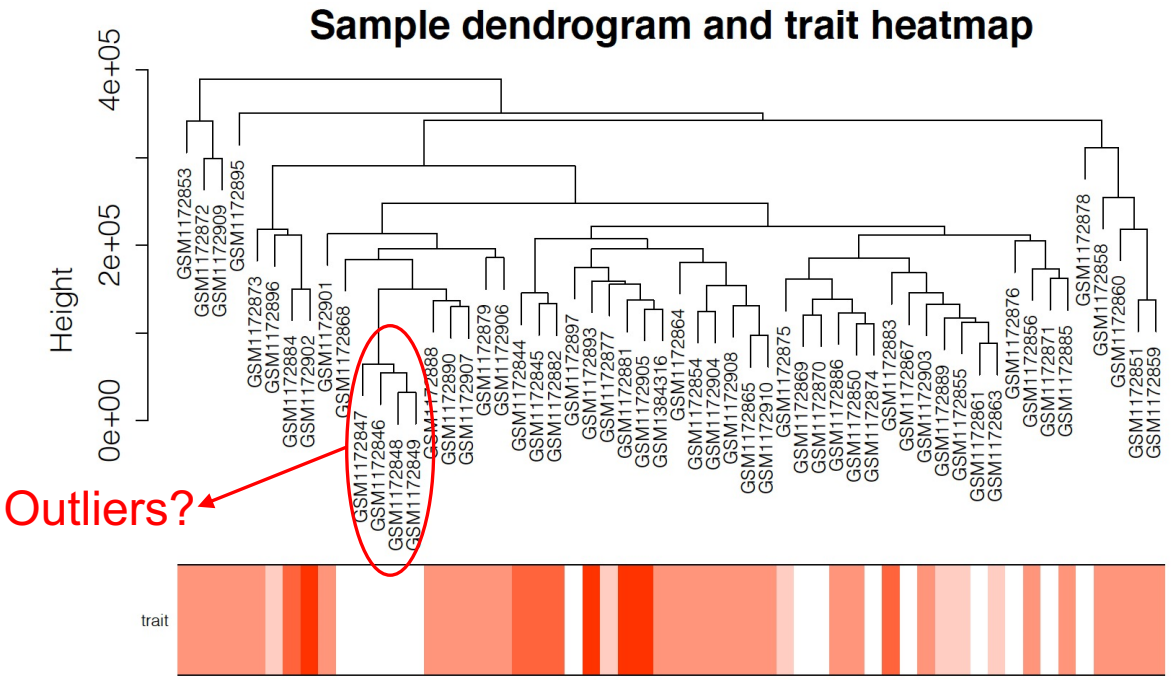
Code Line 1 - Line 27

2. Load tutorial input data

Code Line 29 – Line 44

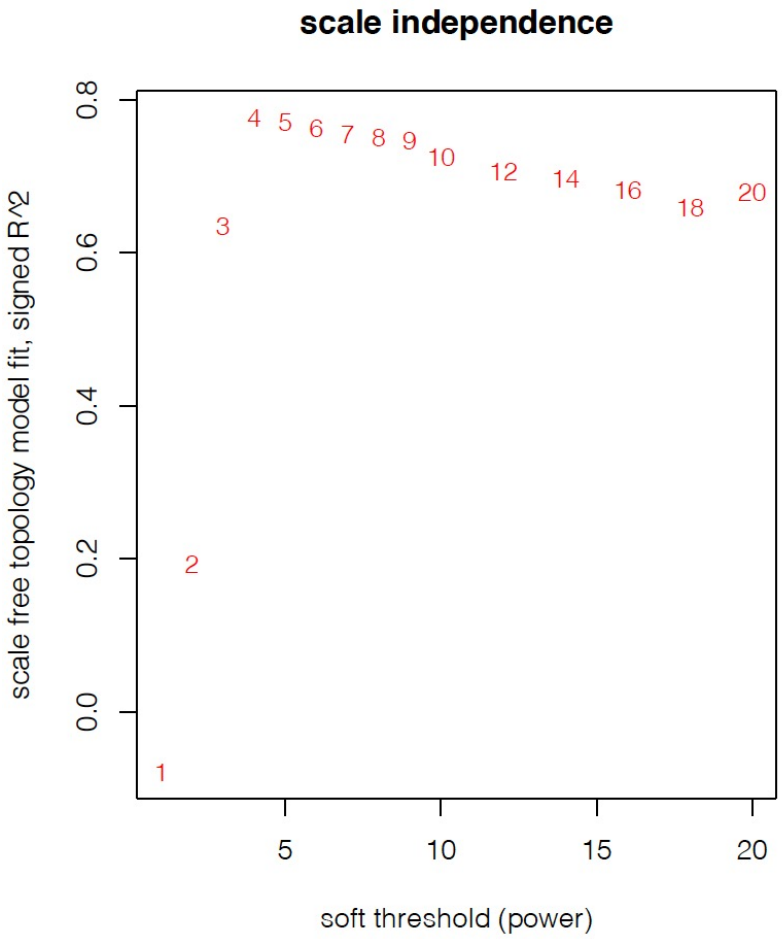
3. Sample clustering to identify outliers

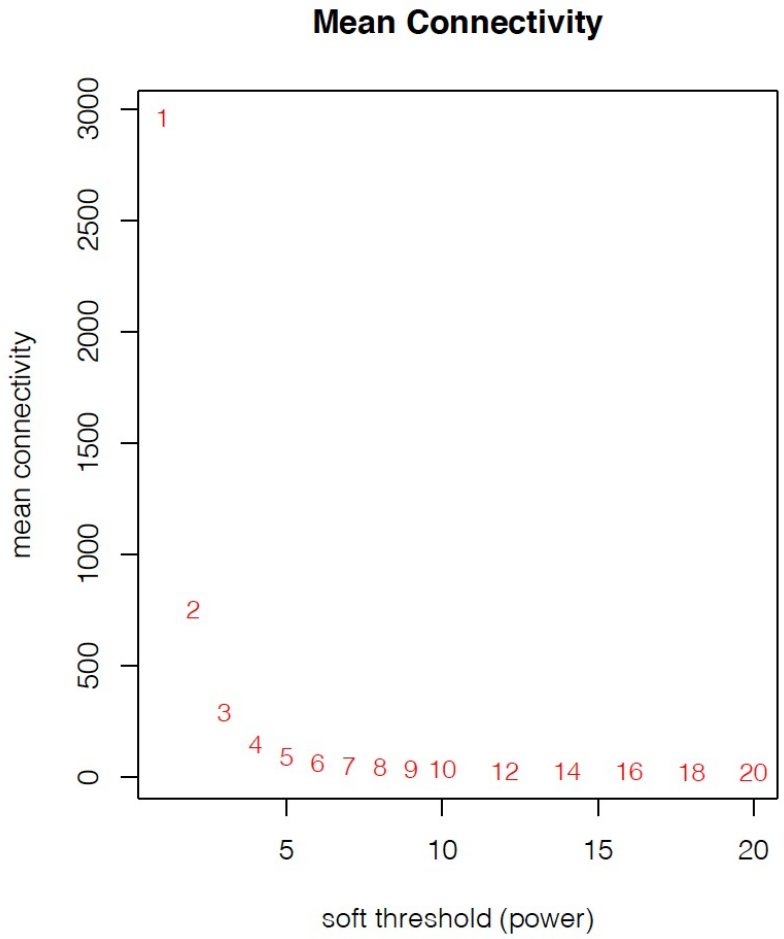
Code Line 44 – Line 54



4. Define the soft thresholding power (β)

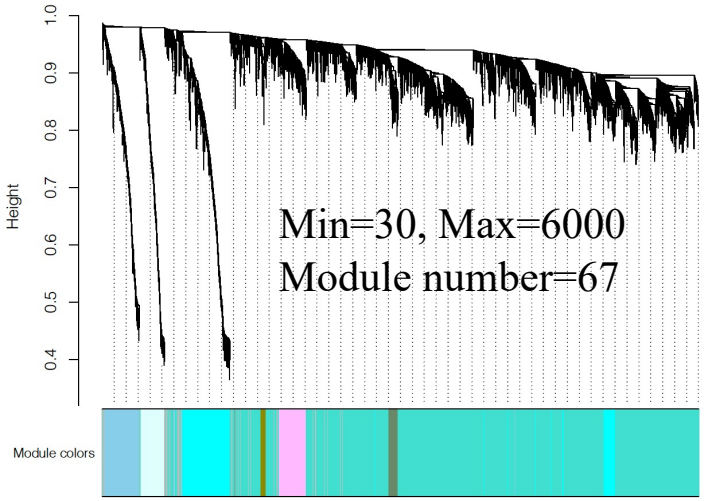
Code Line 56 – Line 76

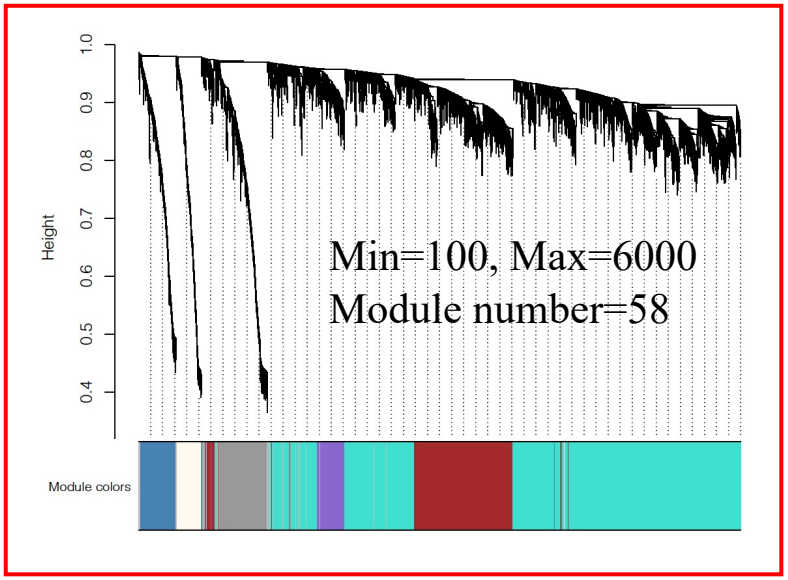


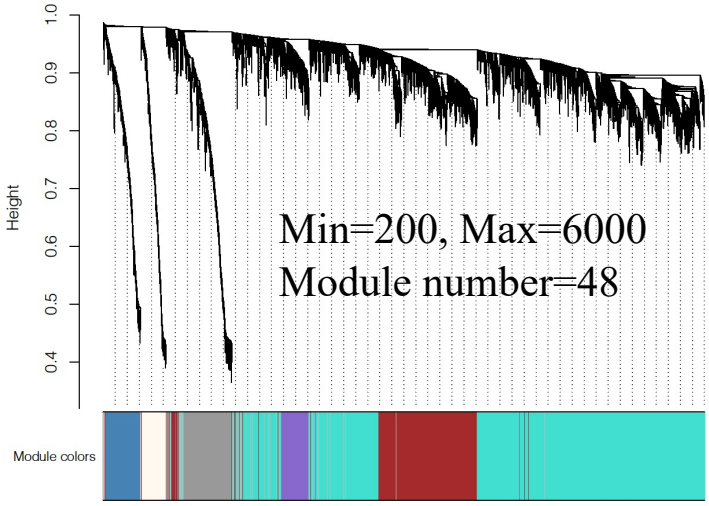


4. Construction gene co-expression modules

Code Line 78 – Line 89







The parameter minModuleSize and maxBlockSize can be used to limit the size of each module, which will result in different module numbers finally

5. Save module results

Code Line 91 – Line 101

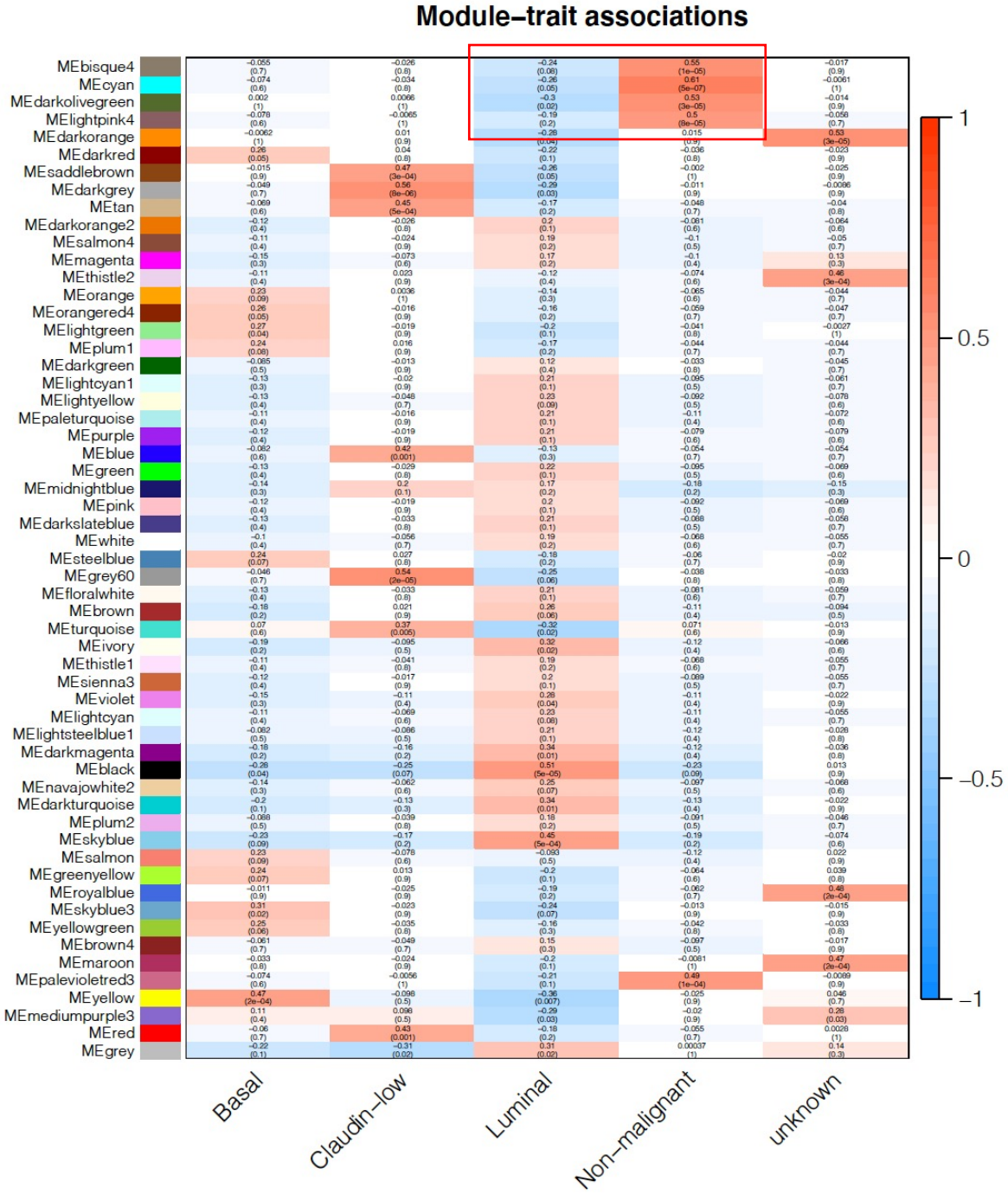
Coexp_modules_for_each_gene_P4M100.txt ——— Module information for each gene

Summary_of_coexp_modules_P4M100.txt ——— Summary of gene numbers in each module

Eigengene_for_each_module_P4M100.txt ——— Eigengene expressions for each module (ME)

6. Quantifying module–trait associations

Code Line 103 – Line 136

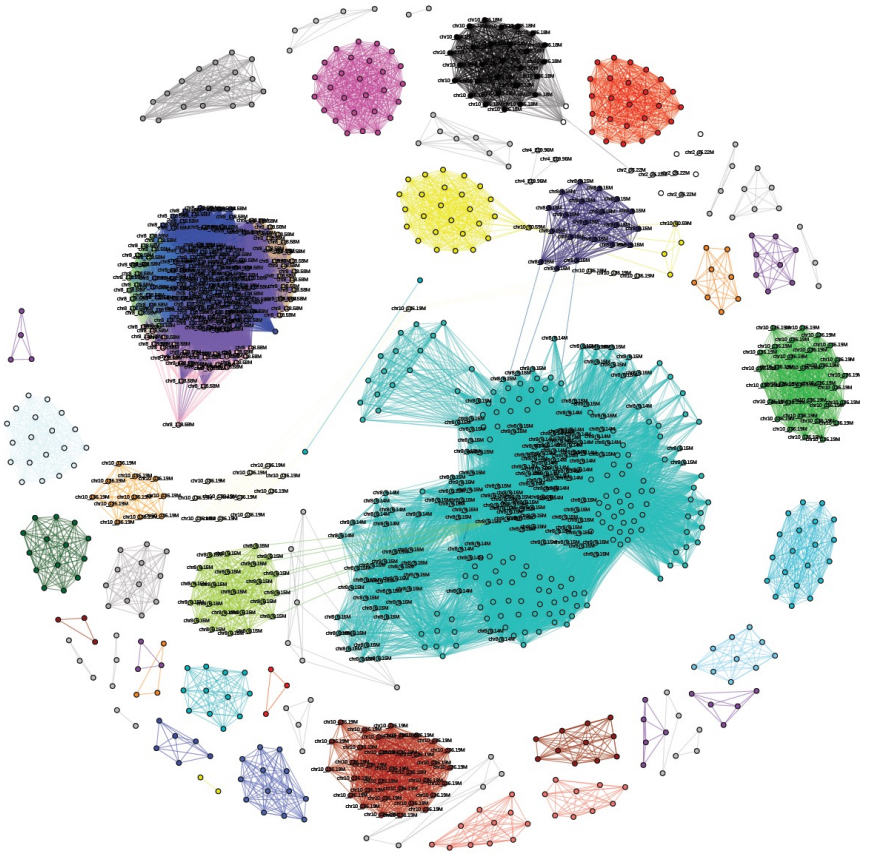

Module−trait associations
MEbisque4
MEcyan
MEdarkolivegreen
MElightpink4
MEdarkorange
MEdarkred
MEsaddlebrown
MEdarkgrey
MEtan
MEdarkorange2
MEsalmon4
MEmagenta
MEthistle2
MEorange
MEorangered4
MElightgreen
MEplum1
MEdarkgreen
MElightcyan1
MElightyellow
MEpaleturquoise
MEpurple
MEblue
MEgreen
MEmidnightblue
MEpink
MEdarkslateblue
MEwhite
MEsteelblue
MEgrey60
MEfloralwhite
MEbrown
MEturquoise
MEivory
MEthistle1
MEsienna3
MEviolet
MElightcyan
MElightsteelblue1
MEdarkmagenta
MEblack
MEnavajowhite2
MEdarkturquoise
MEplum2
MEskyblue
MEsalmon
MEgreenyellow
MEroyalblue
MEskyblue3
MEyellowgreen
MEbrown4
MEmaroon
MEpalevioletred3
MEyellow
MEmediumpurple3
MEred
MEgrey
Basal
Claudin−low
Luminal
Non−malignant
unknown
1
0.5
0
−0.5
−1

7. Exporting Files for Cytoscape Visualization

Code Line 138 – Line 148 Cytoscape (https://cytoscape.org)

8. Network visualization with Gephi

Folder: gephi_visualization



A short methodological summary of the publications.

- WGCNA methods
  - Horvath S (2011) Weighted Network Analysis. Applications in Genomics and Systems Biology. Springer Book. ISBN: 978-1-4419-8818-8
  - Zhang B, Horvath S (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17
  - Langfelder P, Horvath S (2008) WGCNA: an R package for Weighted Correlation Network Analysis. BMC Bioinformatics. 2008 Dec 29;9(1):559. PMID: 19114008 PMCID: PMC2631488
  - Langfelder P et al (2011) Is my network module preserved and reproducible? PloS Comp Biol. 7(1): e1001057. PMID: 21283776
- Math and WGCNA:
  - Horvath S, Dong J (2008) Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology. 4(8): e1000117. PMID: 18704157
- Empirical evaluation of WGCNA
  - Langfelder P, et al (2013) When Is Hub Gene Selection Better than Standard Meta-Analysis? PLoS ONE 8(4): e61505.
  - Song L, Langfelder P, Horvath S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices.BMC Bioinformatics;13(1):328. PMID: 23217028
- What is the topological overlap measure? Empirical studies of the robustness of the topological overlap measure:
  - Yip A, Horvath S (2007) Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics 8:22
- Dynamic branch cutting:
  - Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics.;24(5):719-20. PMID: 18024473
- Gene screening based on intramodular connectivity identifies brain cancer genes that validate.
  - Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Shu, Q, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target", PNAS | November 14, 2006 | vol. 103 | no. 46 | 17402-17407
- How to integrate SNP markers into weighted gene co-expression network analysis?
  - Plaisier CL et al Pajukanta P (2009) A systems genetics approach implicates USF1, FADS3 and other causal candidate genes for familial combined hyperlipidemia. PloS Genetics;5(9):e1000642 PMID: 19750004
- Differential network analysis:
  - Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S (2007) "Weighted Gene Co-expression Network Analysis Strategies Applied to Mouse Weight", Mammalian Genome.

More WGCNA R-package tutorials:
http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

My E-mail address:
ksuhecheng90@gmail.com

Thank you