Introduction to Systems Biology
Class 02

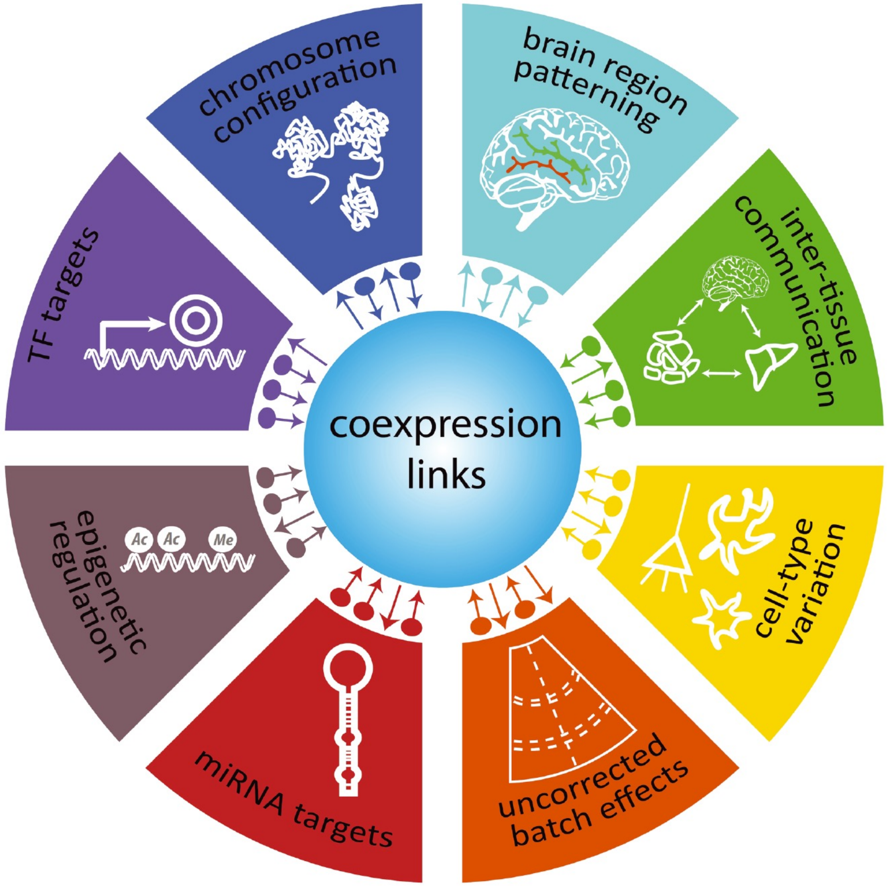
Katia de Paiva Lopes, PhD
Rush Alzheimer's Disease Center (RADC)
Instituto de Assistência Médica ao Servidor Público Estadual de São Paulo (IAMSPE)
Universidade Federal do Paraná (UFPR)

Agenda

- Network as a tool of System’s Biology
- Network basics and concepts
- WGCNA method for coexpression networks
- Noise in gene expression
- Example of a coexpression network
- Tutorials to follow

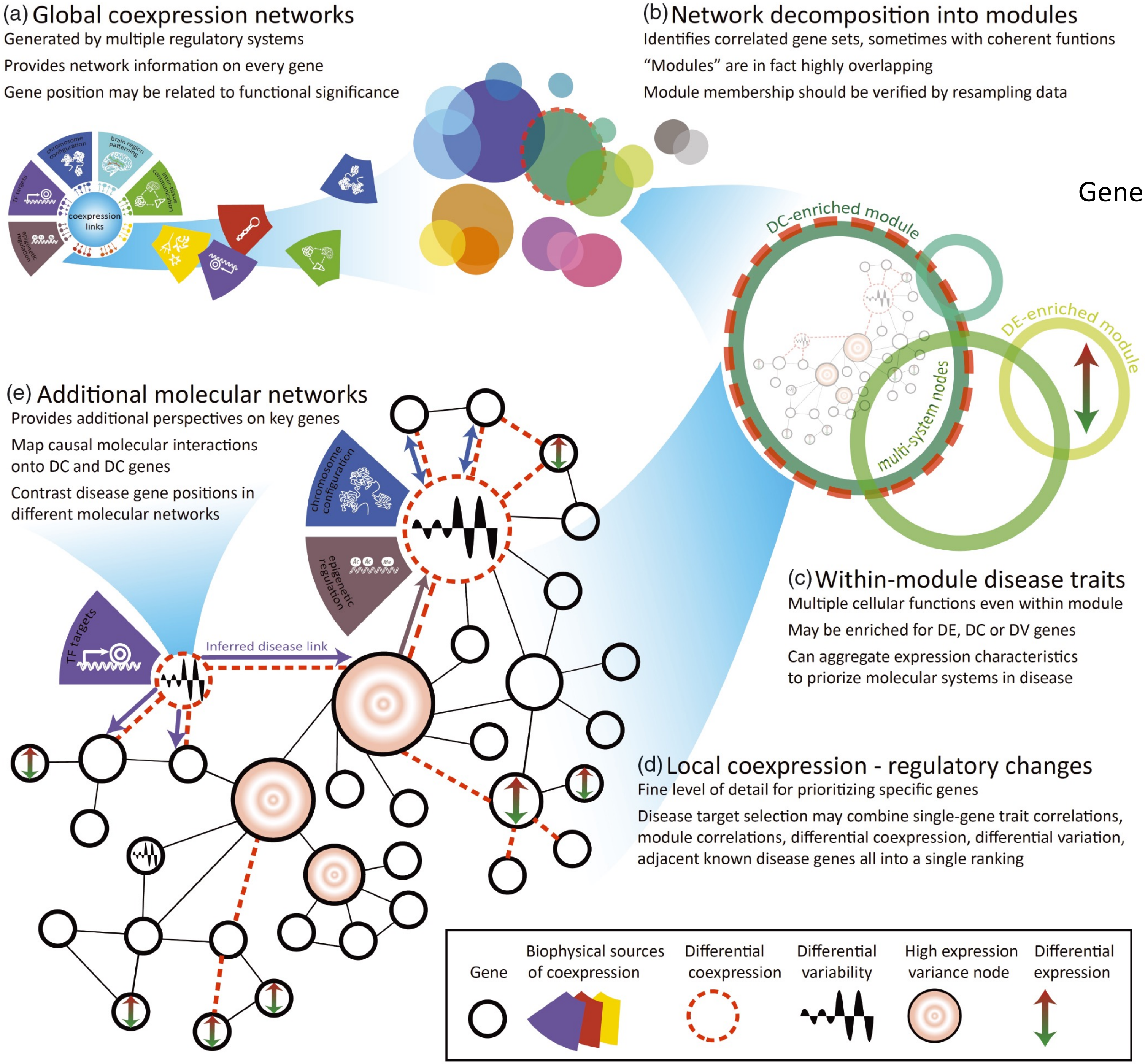
Hands on!

Networks as a tool in Systems Biology



Summary of molecular, cellular, tissue and technical regulatory sources of observed gene–gene correlations/ coexpression links.

Gaiteri et al. Genes, Brain and Behavior, 2014.



(a) Global coexpression networks

Generated by multiple regulatory systems

Provides network information on every gene

Gene position may be related to functional significance

(b) Network decomposition into modules

Identifies correlated gene sets, sometimes with coherent funtions

“Modules” are in fact highly overlapping

Module membership should be verified by resampling data

(c) Within-module disease traits

Multiple cellular functions even within module

May be enriched for DE, DC or DV genes

Can aggregate expression characteristics to priorize molecular systems in disease

(d) Local coexpression - regulatory changes

Fine level of detail for prioritizing specific genes

Disease target selection may combine single-gene trait correlations, module correlations, differential coexpression, differential variation, adjacent known disease genes all into a single ranking

(e) Additional molecular networks

Provides additional perspectives on key genes

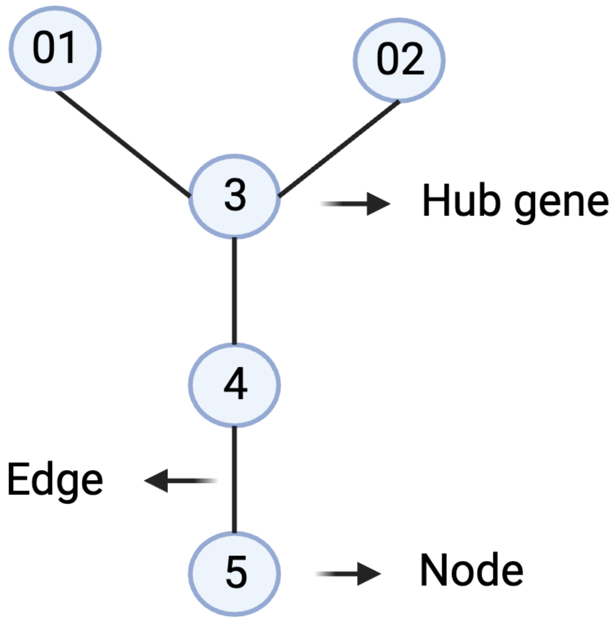
Map causal molecular interactions onto DC and DC genes

Contrast disease gene positions in different molecular networks

Gaiteri et al.
Genes, Brain and Behavior, 2014.

Key terminology

Networks are composed of nodes that are connected by edges (links).



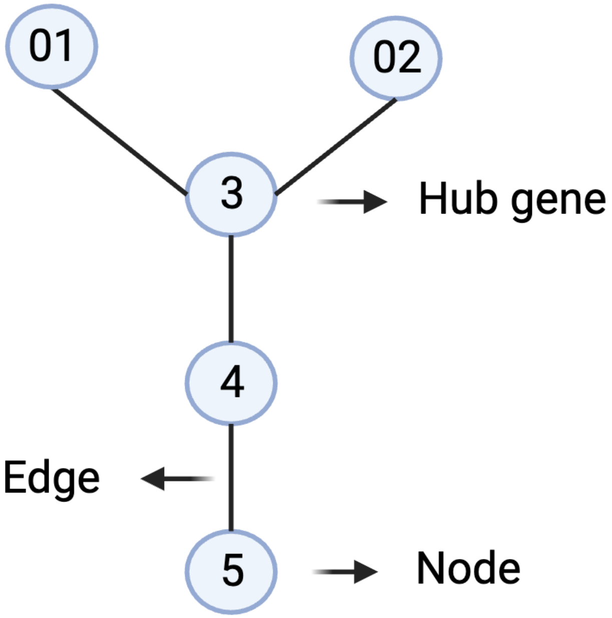
Adjacency matrix

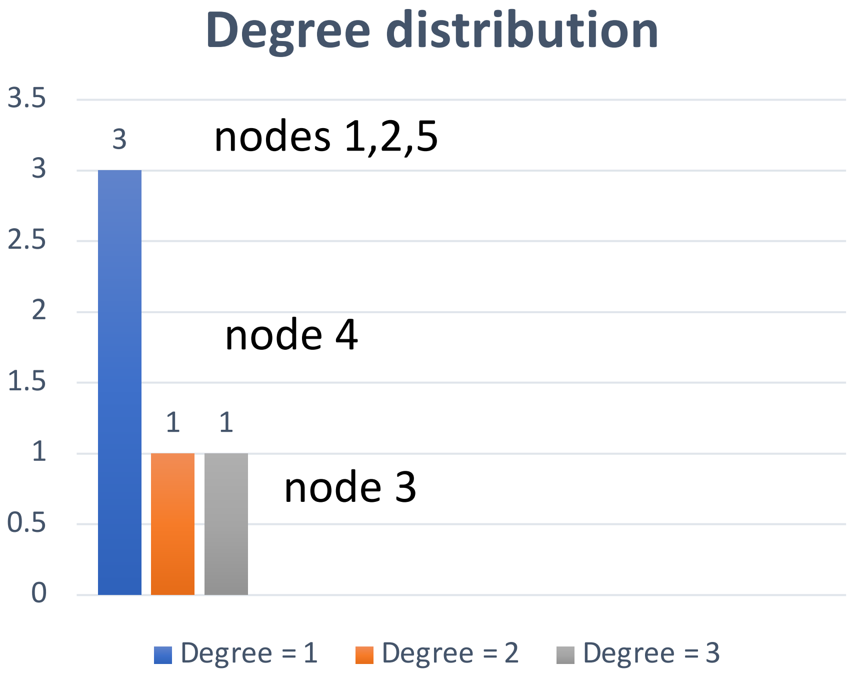
0	0	1	0	0
0	0	1	0	0
1	1	0	1	0
0	0	1	0	1
0	0	0	1	0

Loscalzo, Barabási, Silverman. Network Medicine Book, 2017.
Figure created with Biorender.

Key terminology

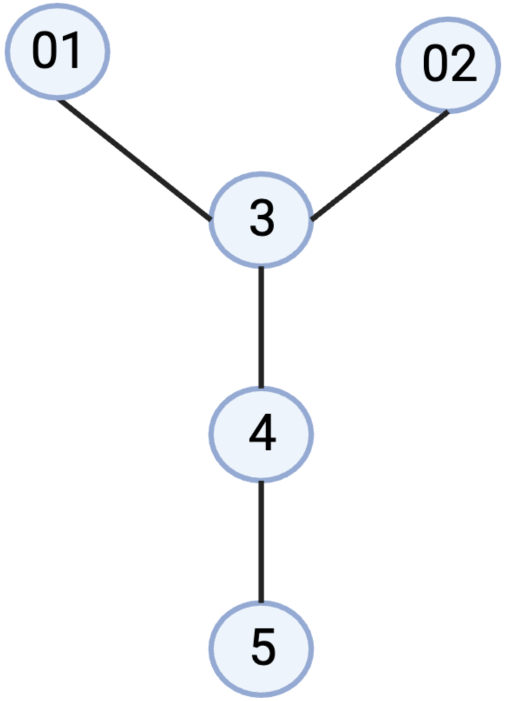
For a particular node, the number of edges directly linked to that node is the **degree**. The **degree distribution** is defined by the frequencies of edges in the network.





Loscalzo, Barabási, Silverman. Network Medicine Book, 2017.
Figure created with Biorender.

Key terminology



A path within a network is a connection between two nodes that follows the edges. The length of the path is quantified by the number of edges included in the path.

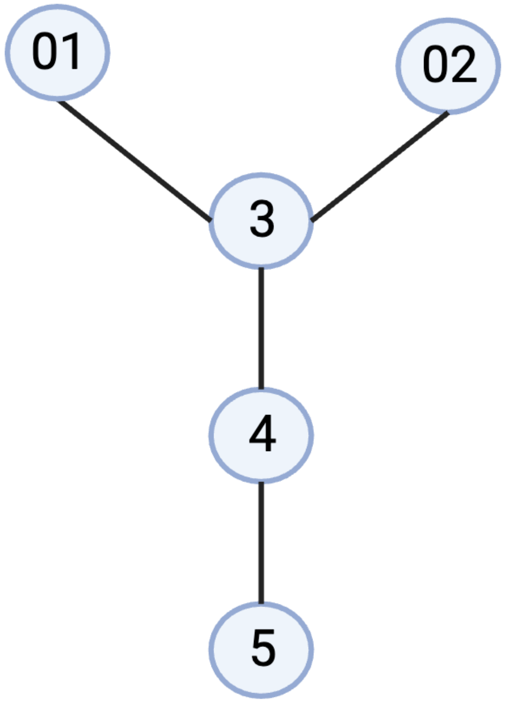
Shortest path lengths

Nodes	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**
Shortest path	1-3-2	1-3	1-3-4	1-3-4-5	2-3	2-3-4	2-3-4-5	3-4	3-4-5	4-5
Path length	2	1	2	3	1	2	3	1	2	1

Mean shortest length = 18/10 = 1.8

Loscalzo, Barabási, Silverman. Network Medicine Book, 2017.
Figure created with Biorender.

Key terminology



Small word effect = the path lengths between nodes are surprisingly small (Watts and Strogatz, 1998).

The betweenness of a node or edge assesses how often that network is present **within the group** of shortest paths in the network.

Betweenness centralities

Nodes	1	2	3	4	5
Shortest paths including node	4	4	9	7	4
Betweenness	0.4	0.4	0.9	0.7	0.4

$$Betweenness = \frac{(N\ of\ shortest\ paths\ including\ node)}{N\ of\ shortest\ path}$$

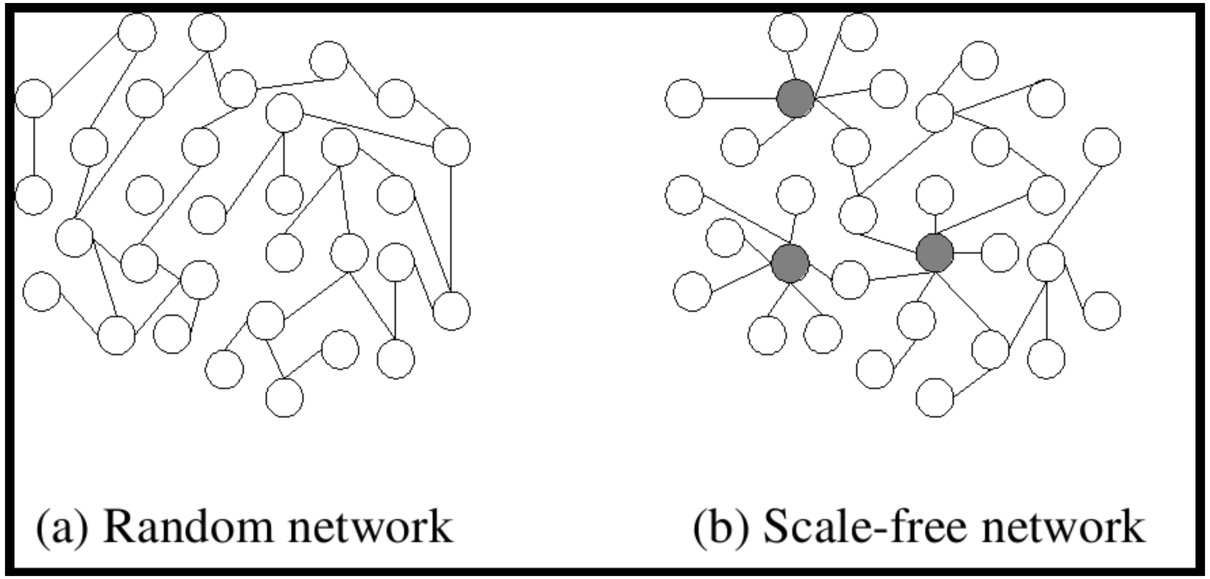
Loscalzo, Barabási, Silverman. Network Medicine Book, 2017.
Figure created with Biorender.

Graph properties of transcription networks

- Transcription networks are sparse!
- What is the maximal number of edges in a network with N nodes? Each node can have an outgoing edge to each of the N-1 other nodes for a total of **E_{max} = N(N-1) edges.**
- The number of edges found in transcription networks, **E**, is much smaller. Being sparse, in the sense that **$E/E_{max} << 1$**.
- Transcription networks are the product of evolutionary selection. It’s easy to lose and edge in a network.

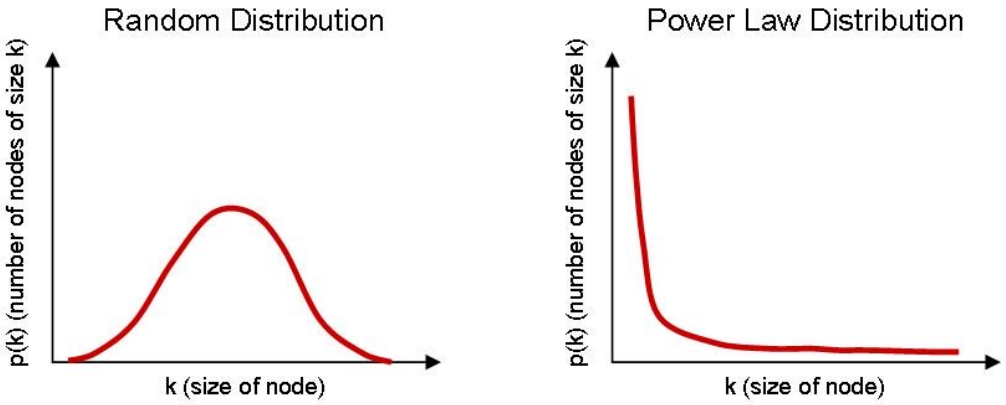
Uri Alon. An Introduction to Systems Biology. Book, 2015.

Network topology



Source: https://en.wikipedia.org/wiki/Hub_(network_science)

A scale-free network is a network whose distribution follows a power law. Barábasi et al. found many types of network in many domains to be approximately scale-free, including **metabolic and protein interaction.**



http://jitha.me/power-law-working-hard-enough/

Tasks

- Q1: Write some examples of what can alter links in a co-expression network.

- Q2: What is a hub gene?

The beauty of applying computational methods to biological data

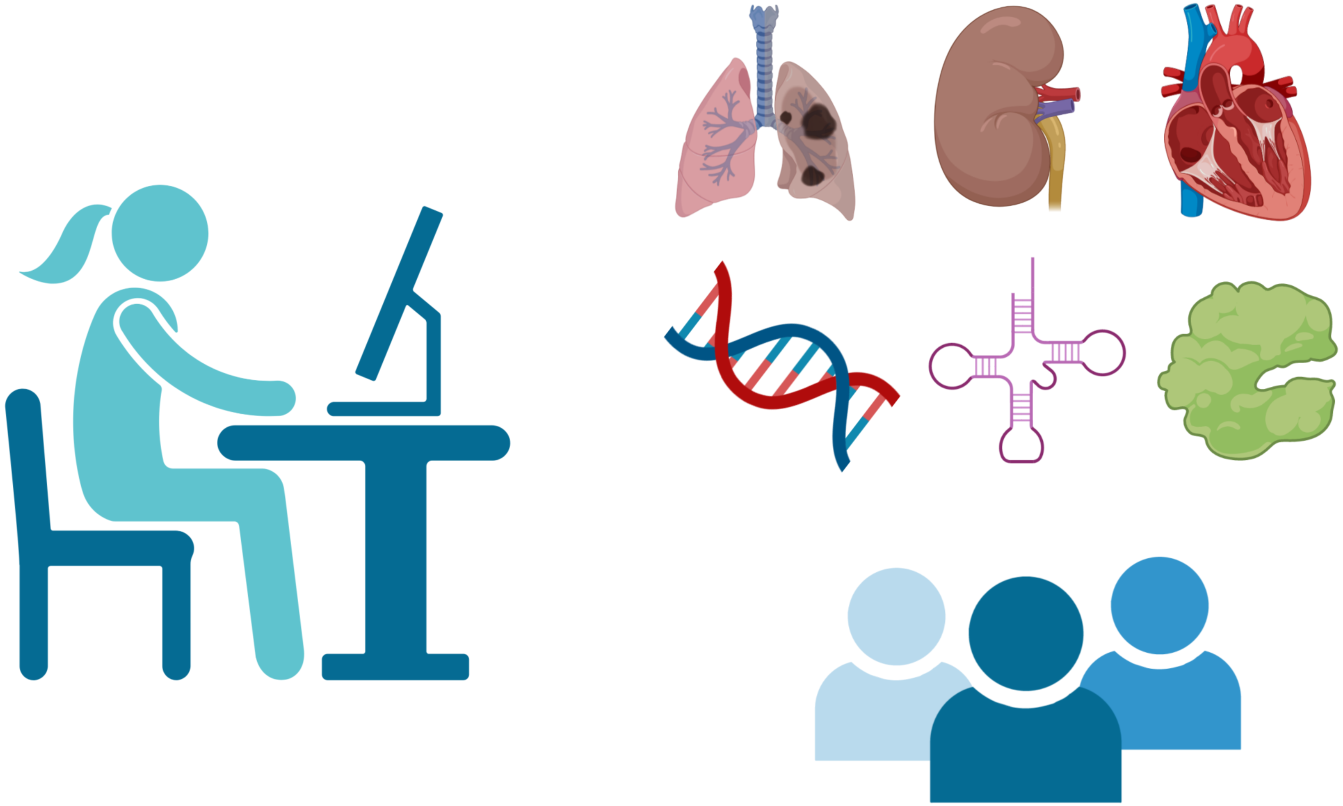
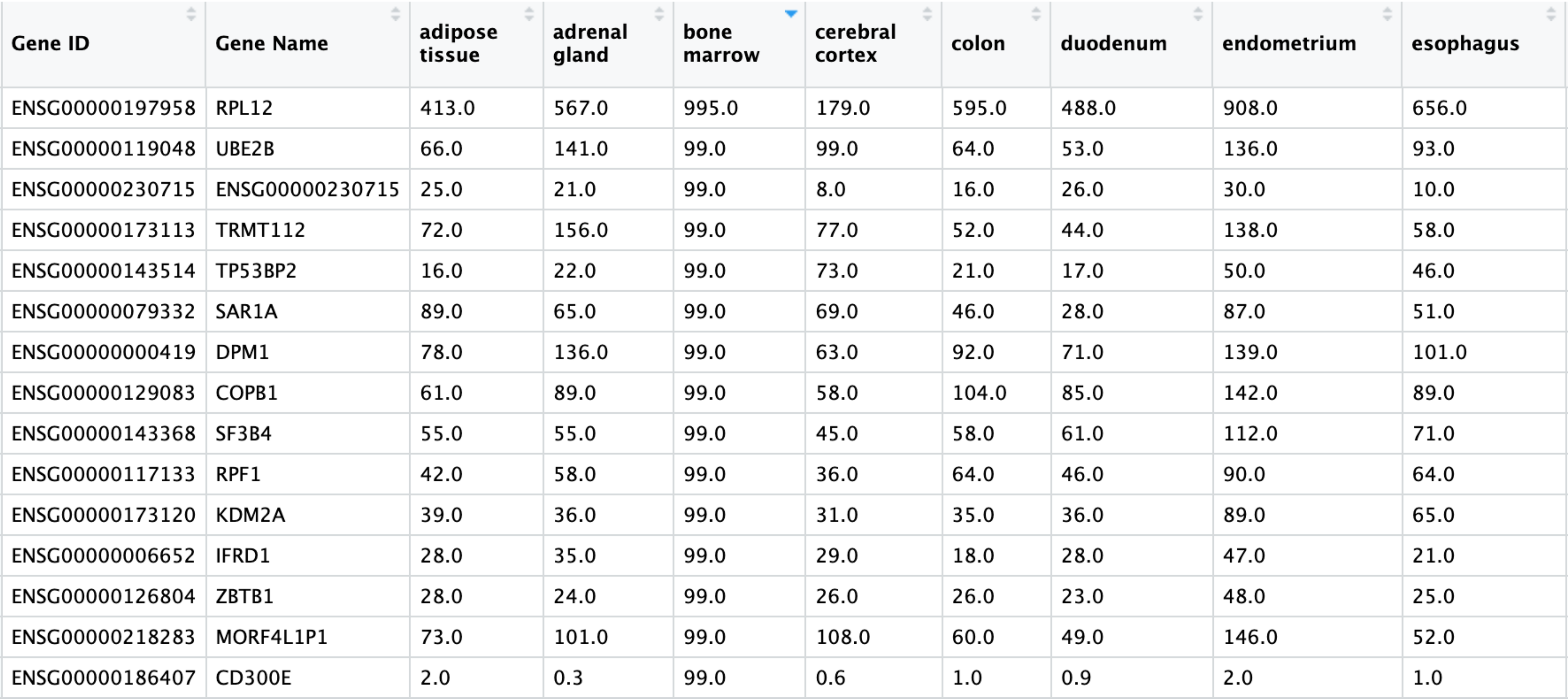


Figure generated with Biorender.

Example of gene expression data (RNASeq)



Gene ID	Gene Name	adipose tissue	adrenal gland	bone marrow	cerebral cortex	colon	duodenum	endometrium	esophagus
ENSG00000197958	RPL12	413.0	567.0	995.0	179.0	595.0	488.0	908.0	656.0
ENSG00000119048	UBE2B	66.0	141.0	99.0	99.0	64.0	53.0	136.0	93.0
ENSG00000230715	ENSG00000230715	25.0	21.0	99.0	8.0	16.0	26.0	30.0	10.0
ENSG00000173113	TRMT112	72.0	156.0	99.0	77.0	52.0	44.0	138.0	58.0
ENSG00000143514	TP53BP2	16.0	22.0	99.0	73.0	21.0	17.0	50.0	46.0
ENSG00000079332	SAR1A	89.0	65.0	99.0	69.0	46.0	28.0	87.0	51.0
ENSG00000000419	DPM1	78.0	136.0	99.0	63.0	92.0	71.0	139.0	101.0
ENSG00000129083	COPB1	61.0	89.0	99.0	58.0	104.0	85.0	142.0	89.0
ENSG00000143368	SF3B4	55.0	55.0	99.0	45.0	58.0	61.0	112.0	71.0
ENSG00000117133	RPF1	42.0	58.0	99.0	36.0	64.0	46.0	90.0	64.0
ENSG00000173120	KDM2A	39.0	36.0	99.0	31.0	35.0	36.0	89.0	65.0
ENSG00000006652	IFRD1	28.0	35.0	99.0	29.0	18.0	28.0	47.0	21.0
ENSG00000126804	ZBTB1	28.0	24.0	99.0	26.0	26.0	23.0	48.0	25.0
ENSG00000218283	MORF4L1P1	73.0	101.0	99.0	108.0	60.0	49.0	146.0	52.0
ENSG00000186407	CD300E	2.0	0.3	99.0	0.6	1.0	0.9	2.0	1.0

https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2836/Results

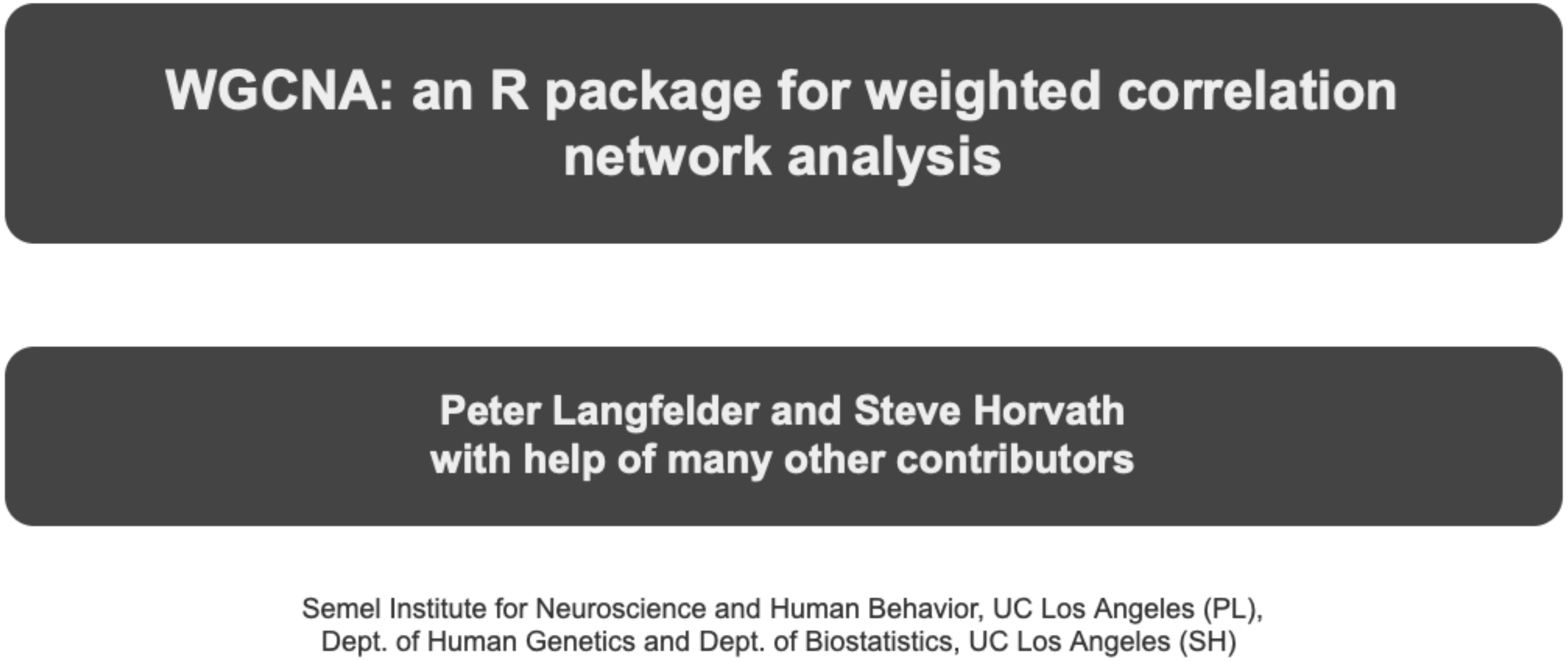
Types of RNASeq downstream analysis

- Differentially Expressed Genes (DEG)
- Age-related analysis (continuous data)
- Sample clusterization
- Functional Enrichment Analysis (FEA)
- Networks

Weighted Gene Co-expression Network Analysis

(WGCNA)

Background



WGCNA: an R package for weighted correlation network analysis

Peter Langfelder and Steve Horvath
with help of many other contributors

Semel Institute for Neuroscience and Human Behavior, UC Los Angeles (PL),
Dept. of Human Genetics and Dept. of Biostatistics, UC Los Angeles (SH)

https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/

Background

WGCNA analysis is a systems biology method for describing the correlation patterns among genes across samples.

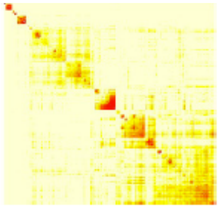
It can be used for:

- Finding **modules** of highly correlated genes
- For summarizing clusters using the module **eigengene** or an intramodular hub gene
- For relating modules to one another and to external **sample traits**
- For calculating **module membership** measures

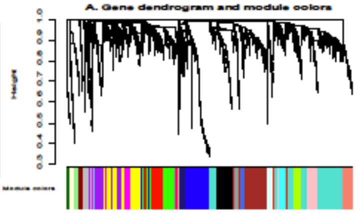
https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/

Overview

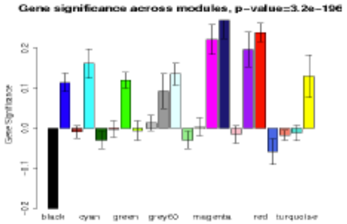
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



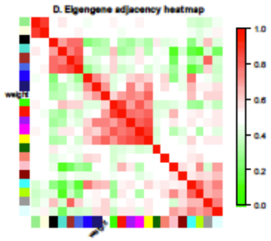
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



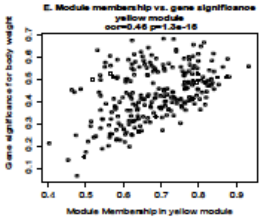
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



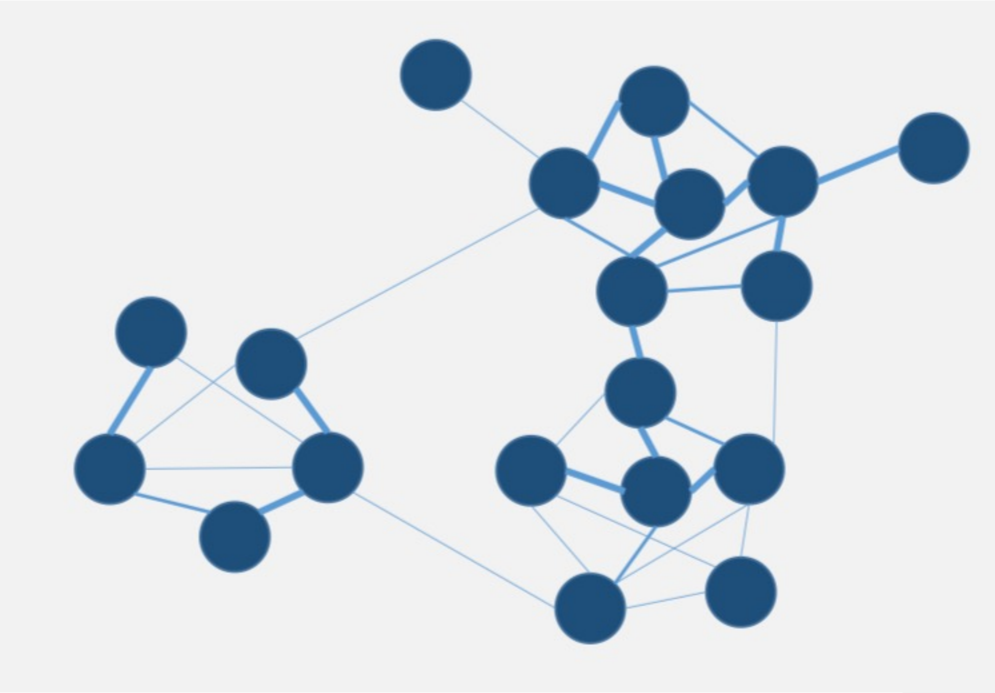
Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



https://pdfs.semanticscholar.org/dc7e/b33db056e083c04d90ee7c5d8e7650889643.pdf

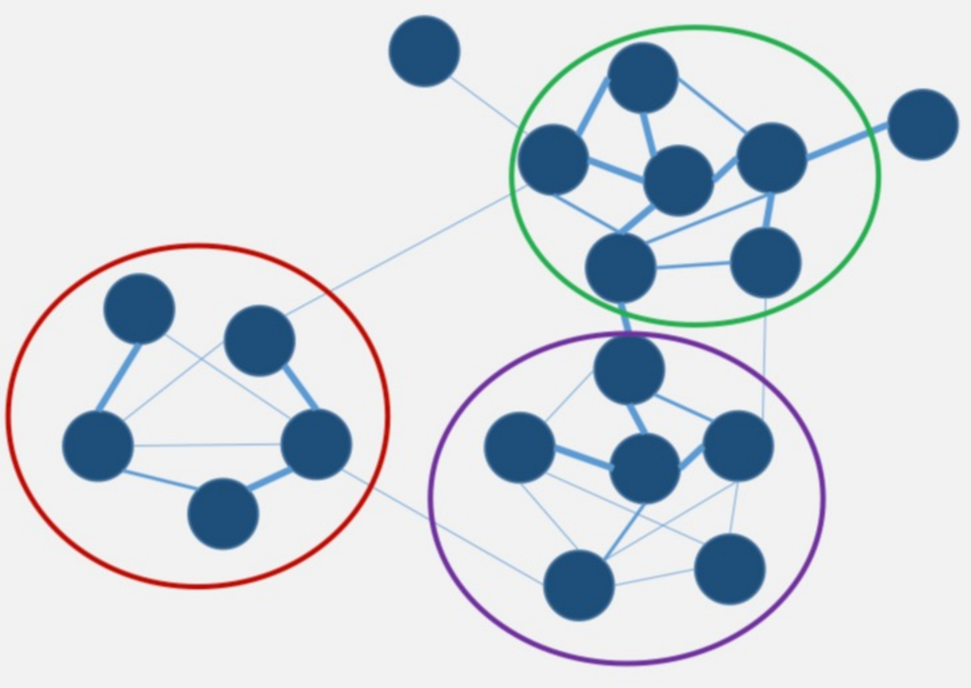
Background

Construct a gene-gene similarity network



Divide network into modules

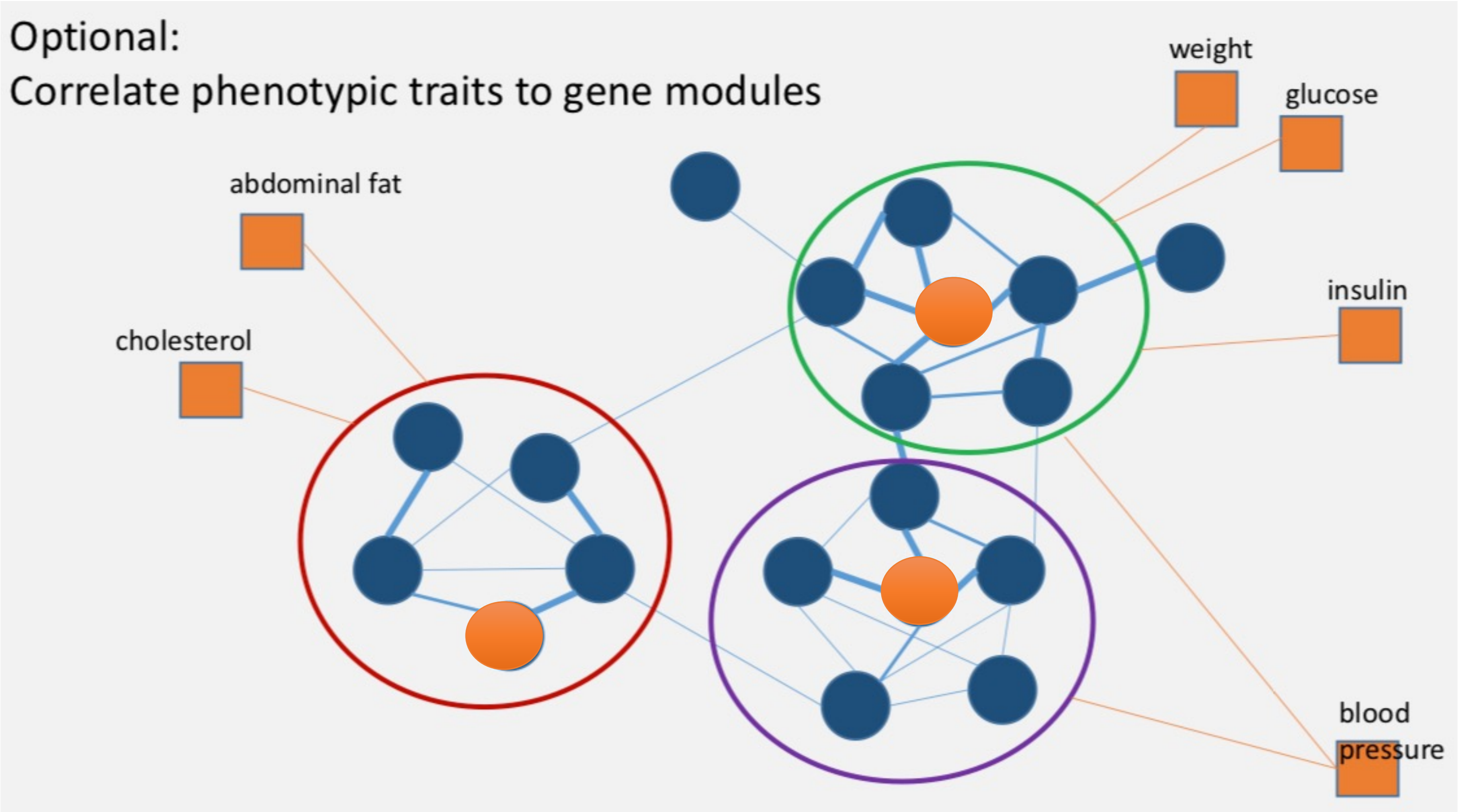
Group genes with similar expression



Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Background

Identify "hub" genes in modules



Optional:
Correlate phenotypic traits to gene modules

Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Background

Hypothesis

Genes with similar expression patterns are interesting because they may be:

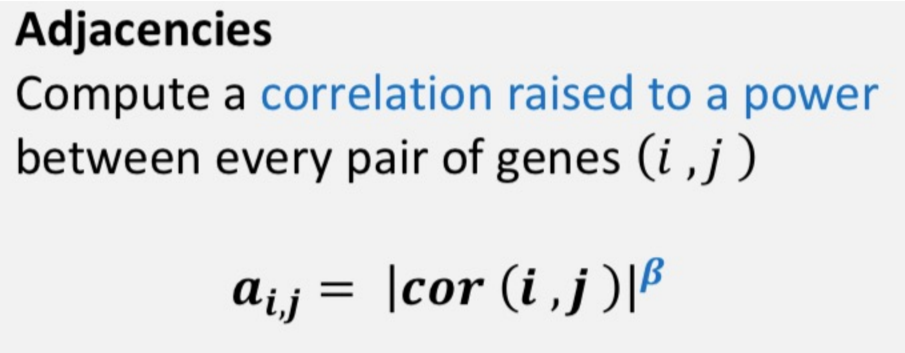
- Tightly **co-regulated**
- **Functionally** related
- Members of the same **pathway**

WGCNA encourages hypotheses about genes based on their close network neighbors.

Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Glossary – Co-expression network

The Basis of WGCNA: Weighted Correlation Network of Genes

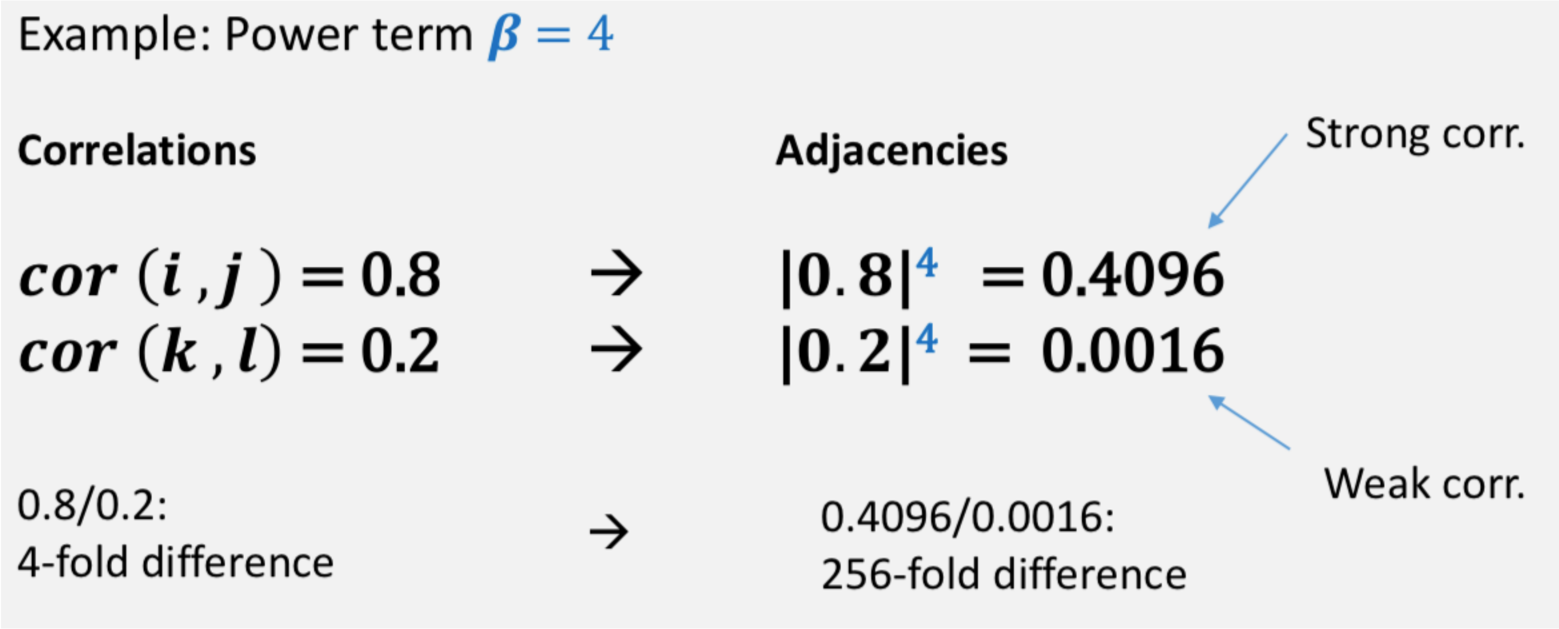


Adjacencies

Compute a correlation raised to a power between every pair of genes $(i\ ,j\ )$

$$a_{i,j} = |cor\ (i\ ,j\ )|^{\beta}$$

The method amplifies **disparity** between strong and weak correlations



Example: Power term $\beta = 4$

Correlations		**Adjacencies**
$cor\ (i\ ,j\ ) = 0.8$	→	$\lvert 0.8\rvert^{4} = 0.4096$
$cor\ (k\ ,l) = 0.2$	→	$\lvert 0.2\rvert^{4} = 0.0016$
0.8/0.2: 4-fold difference	→	0.4096/0.0016: 256-fold difference

Strong corr.

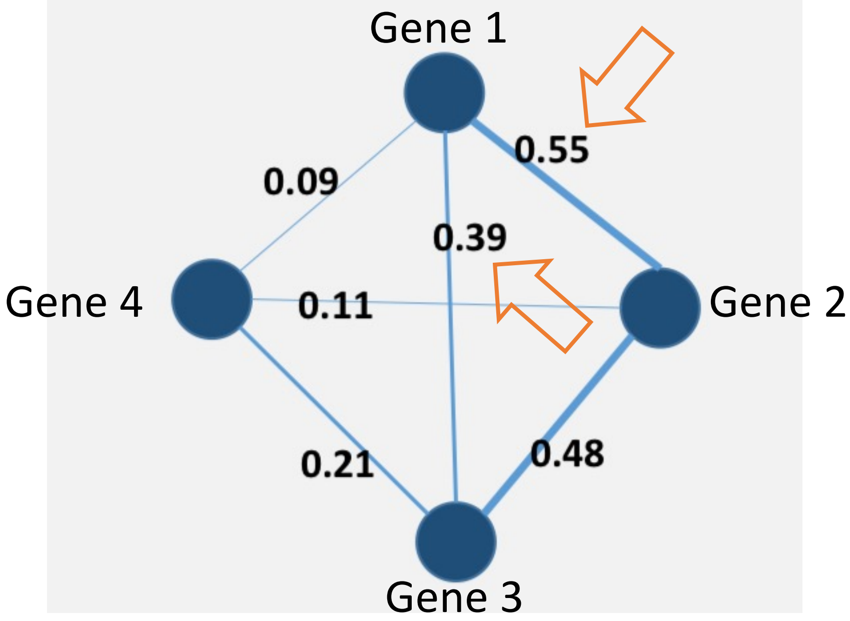
Weak corr.

Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

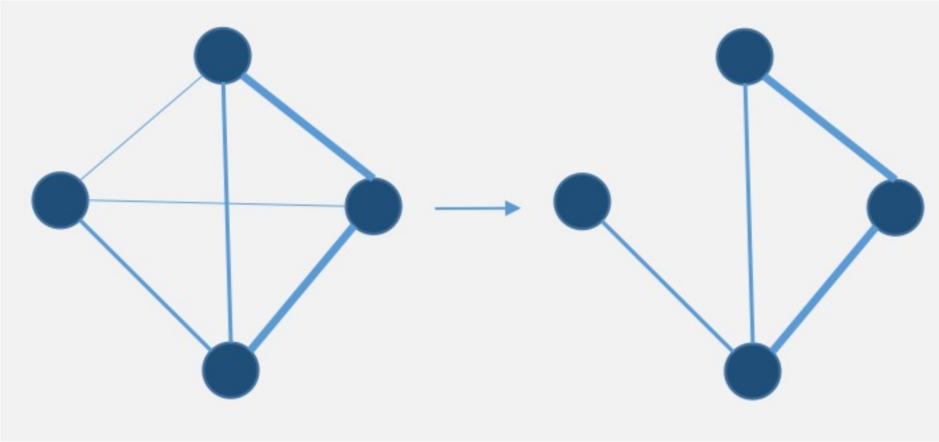
Background

Adjacency matrix of 4 genes

$a_{i,j}$	gene1	gene2	gene3	gene4
gene1	1	0.55	0.39	0.09
gene2	0.55	1	0.48	0.11
gene3	0.39	0.48	1	0.21
gene4	0.09	0.11	0.21	1

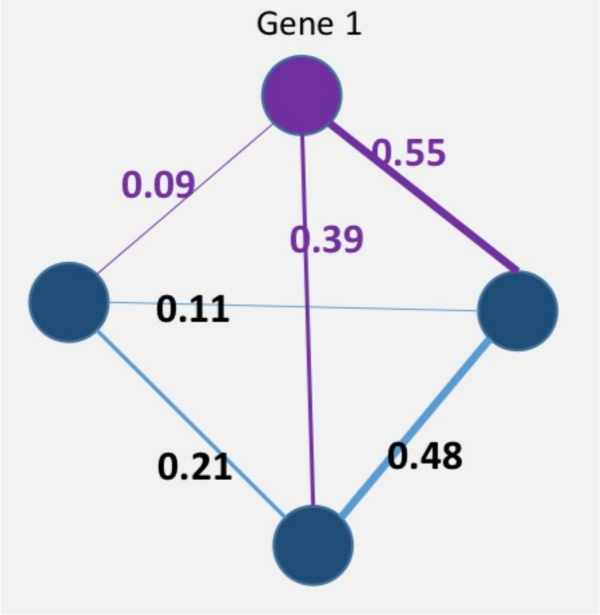


Remove the **weakest** links.



Work with all edges of the fully connected network.

Connectivity (degree) in a weighted network.



Example for **connectivity (k)** of gene 1:

0.55 + **0.39** + **0.09** = 1.03

Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Background

According to WGCNA the co-expression matrix is not enough! The similarity between genes should be reflected at the expression and the network topology level.

Compute similarity/dissimilarity between genes

Topological Overlap Measure (**TOM**):

- Is a pairwise similarity measure between network nodes (genes)
- **TOM**(i,j) is high if genes i,j have many shared neighbors
- A high **TOM**(i,j) implies that genes have similar expression patterns

Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Background

Signed TOM needs as input not only the connection strengths (aij – adjacency matrix), but also the sign of the correlations. The modified adjacency matrix:

$$\tilde{a}_{ij} = a_{ij} \times \operatorname{sign}\left(\operatorname{cor}(x_i, x_j)\right) . \tag{1}$$

The signed TOM is then defined as

Connectivities of nodes

Adjacency matrix

$$TOM_{ij}^{signed} = \frac{\left|\boxed{a_{ij}} + \boxed{\sum_{u \neq i,j} \tilde{a}_{iu} \tilde{a}_{uj}}\right|}{\min\boxed{(k_i, k_j)} + 1 - \left|\boxed{a_{ij}}\right|} , \tag{2}$$

Where k_i and k_j denote the connectivities of nodes i and j:

$$k_i = \sum_{u \neq i} |\tilde{a}_{ui}| . \tag{3}$$

K = connectivity degree based on neighbors.

In contrast, unsigned TOM uses absolute values in the numerator:

$$TOM_{ij} = \frac{|a_{ij}| + \sum_{u \neq i,j} |\tilde{a}_{iu} \tilde{a}_{uj}|}{\min(k_i, k_j) + 1 - |a_{ij}|} . \tag{4}$$

Source: Peter Langfelder, 2013. Signed vs Unsigned. Technical report.

Glossary – TOM

$$TOM_{ij}^{signed} = \frac{\left|a_{ij} + \sum_{u \neq i,j} \tilde{a}_{iu}\tilde{a}_{uj}\right|}{\min(k_i, k_j) + 1 - |a_{ij}|},$$

1 – Count numbers of shared neighbors:

Using the connectivity degree **(k)**

2 – Normalize values between 0 and 1:

TOM(i,j) = 0: no overlap of network neighbors

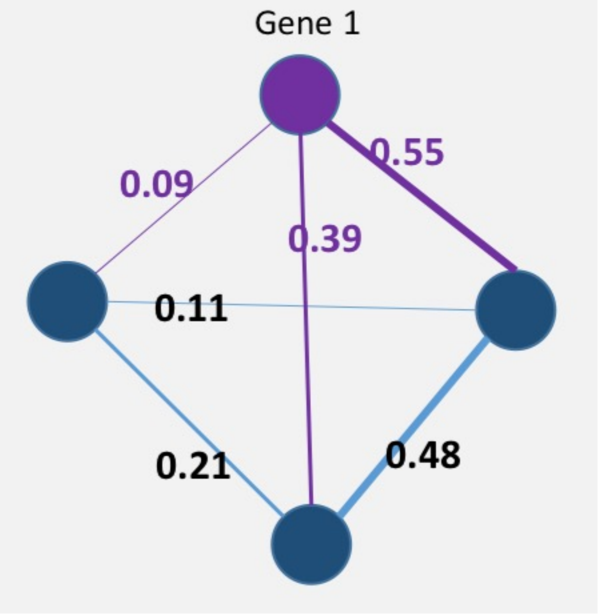
TOM(i,j) = 1: identical set of network neighbors

3 – Then, we can calculate the (dis)similarity measure **distTOM = 1-TOM**.

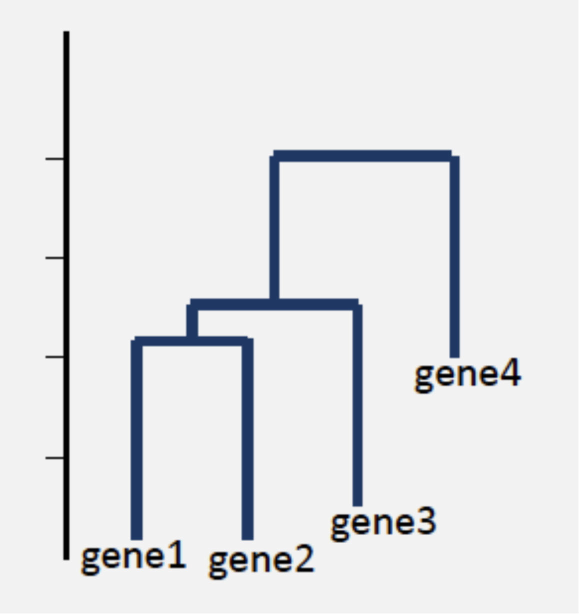
Source: Peter Langfelder, 2013. Signed vs Unsigned. Technical report.
Leonore Wigger with Frédéric Burdet and Mark Ibberson

Background

Weighted correlation network from gene expression data



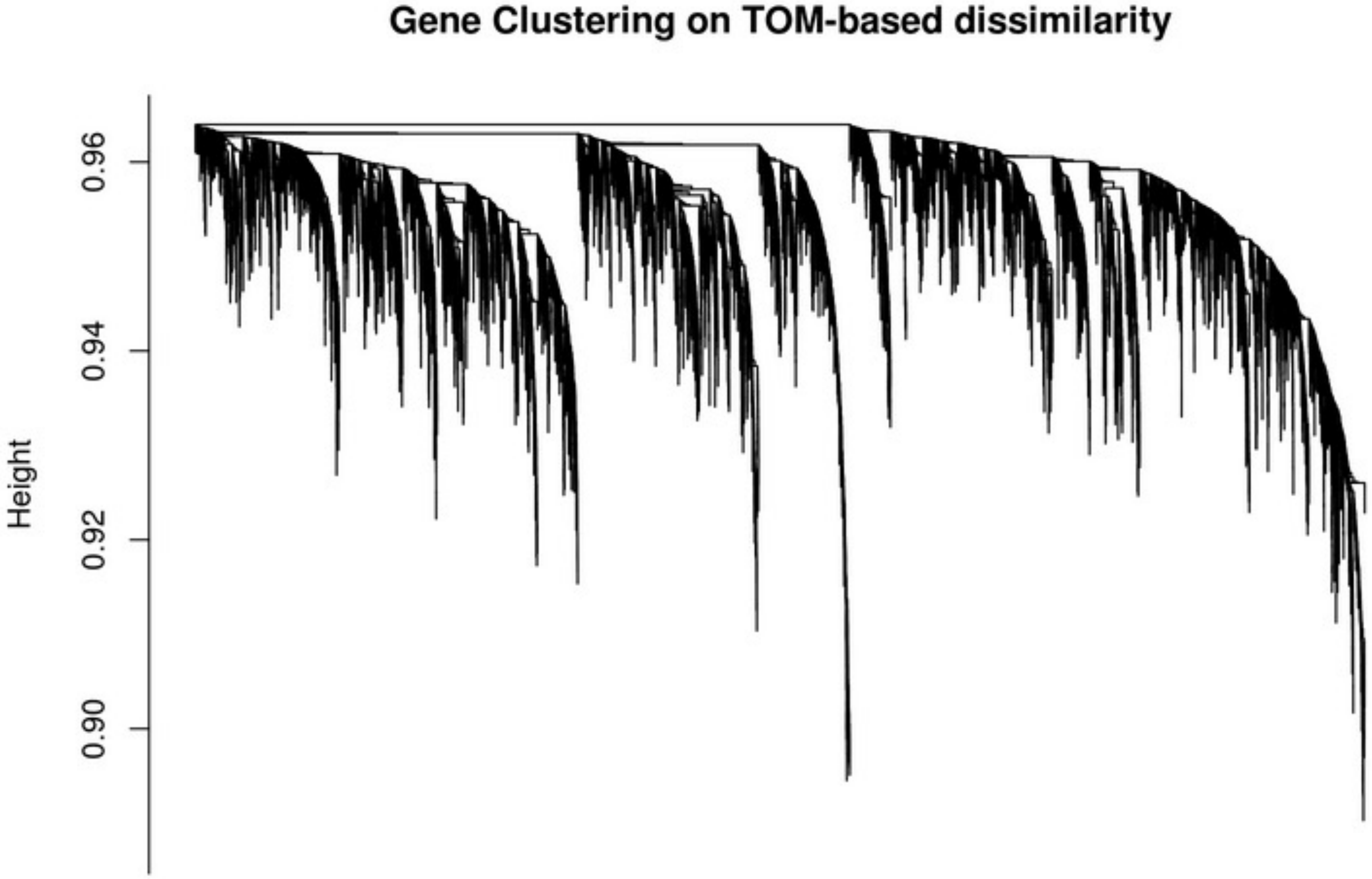
Gene clustering dendrogram



(dis)similarity between genes:
Topological Overlap Measure (TOM)

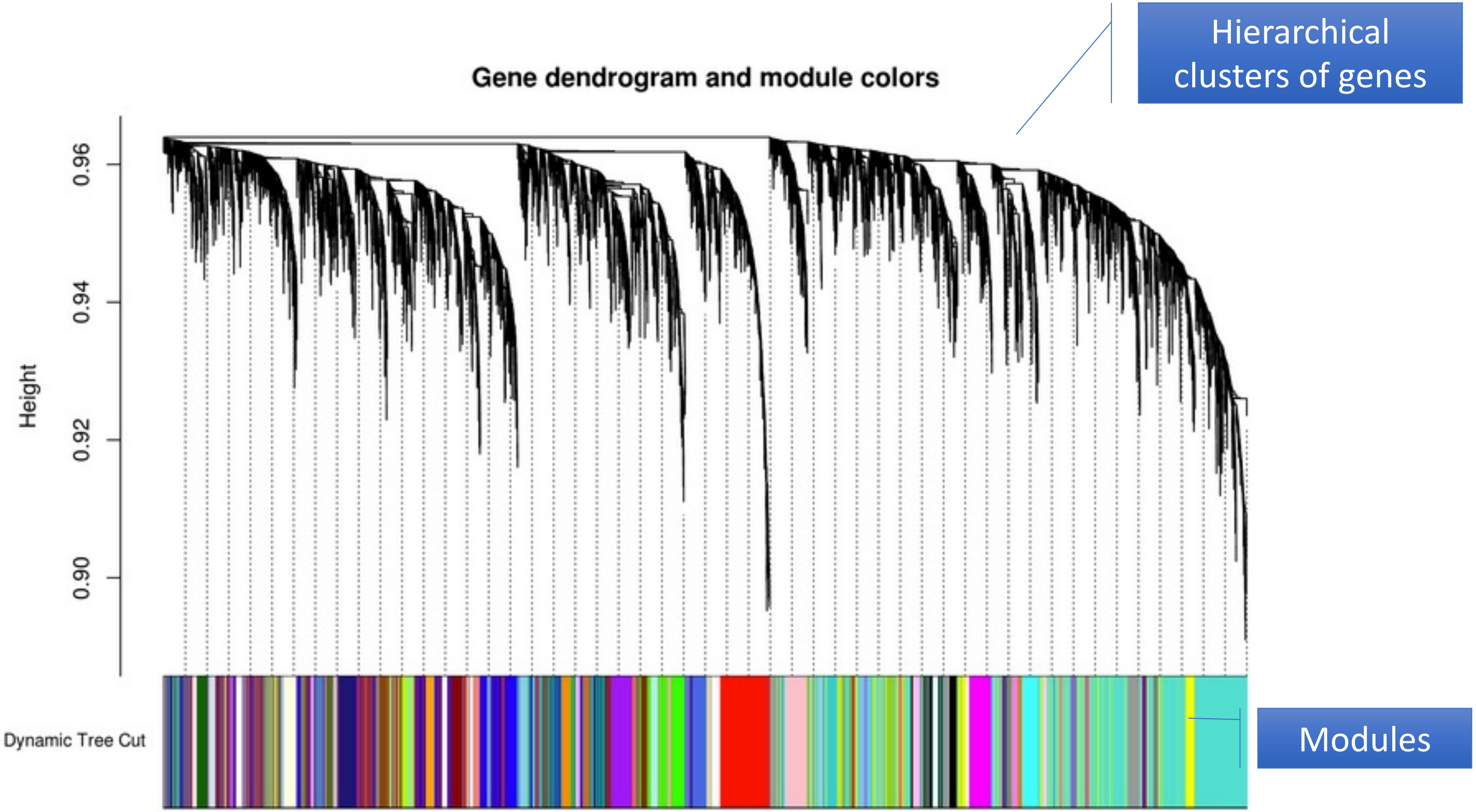
Source: Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Background – Signed network



Background – Signed network

Divide clustered genes into modules using the Dynamic tree cut algorithm.



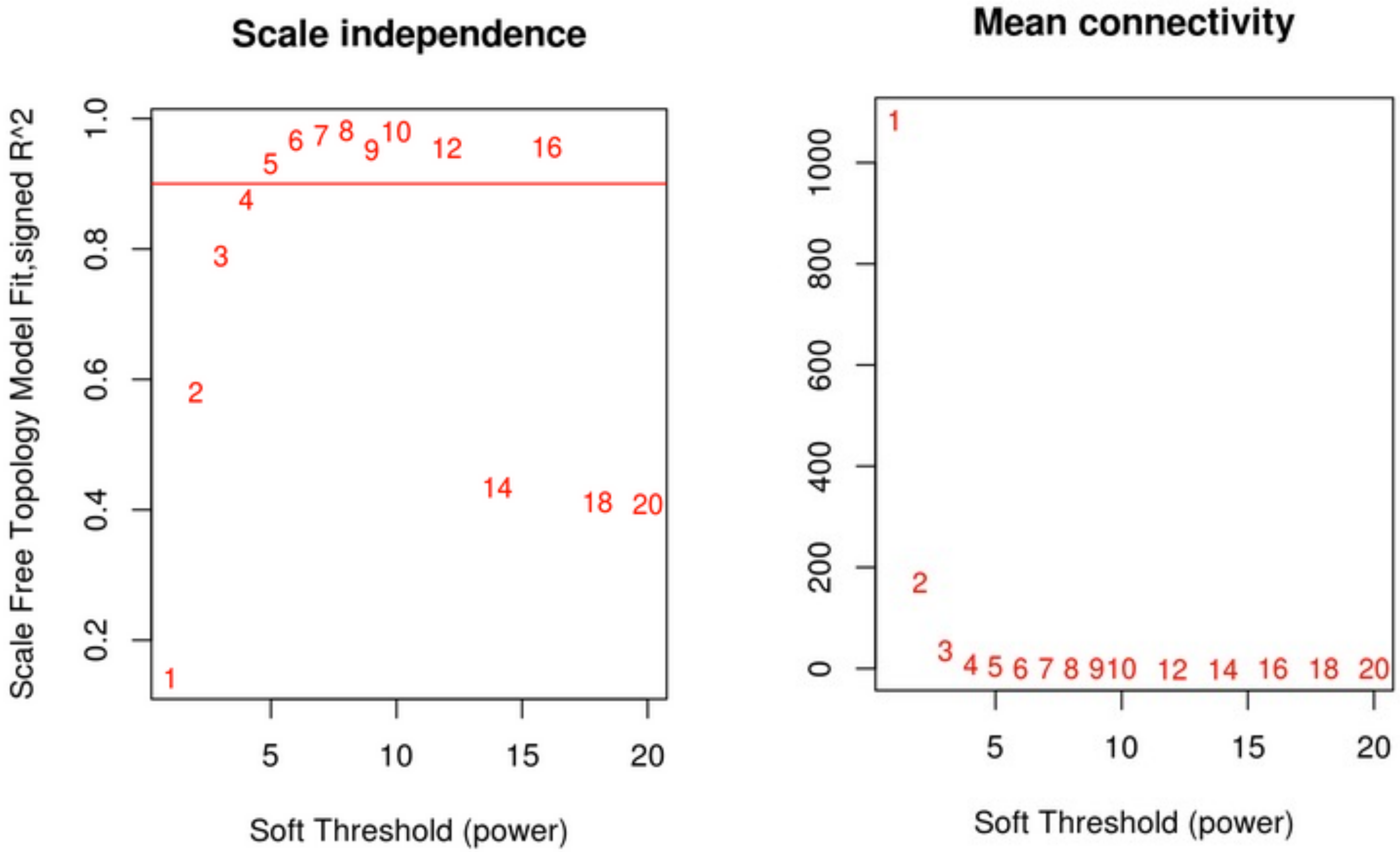
Hierarchical clusters of genes

Modules

WGCNA has a visual way to pick a power term:

We need to choose a soft thresholding power that approximately fits a scale free network. It means, the lowest power on or above the red horizontal line.

Mean connectivity plot: mean connectivity drops as power goes up.



Glossary – Module Eigengene

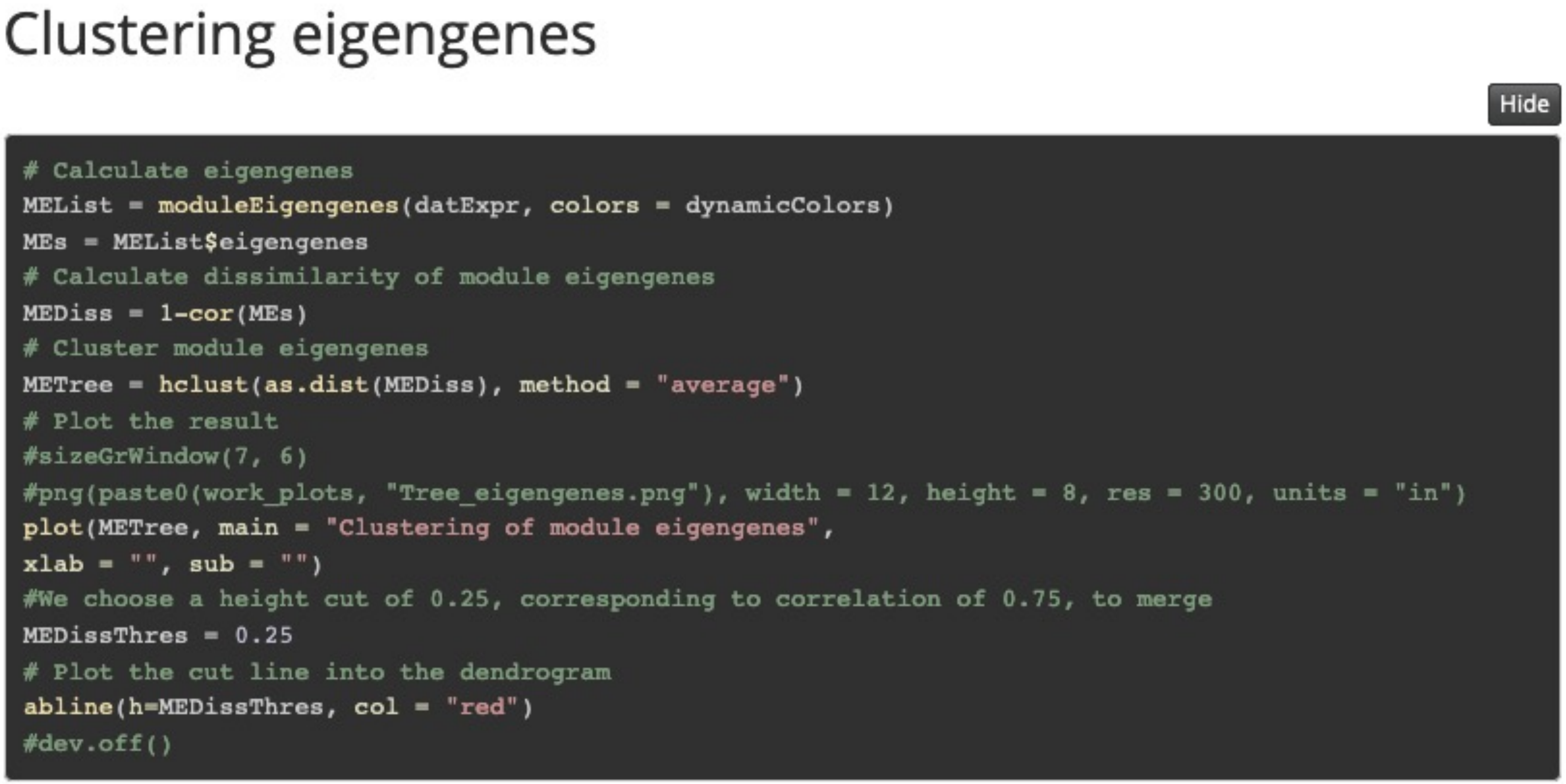
Next step: **merge** very similar modules using the eigengenes.

Eigengene 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. It's a way to summarize the expression data from a module.

Eigengenes are used for:

- Modules can be correlated with one another
- Modules can be correlated with external traits

Steps to calculate the eigengenes

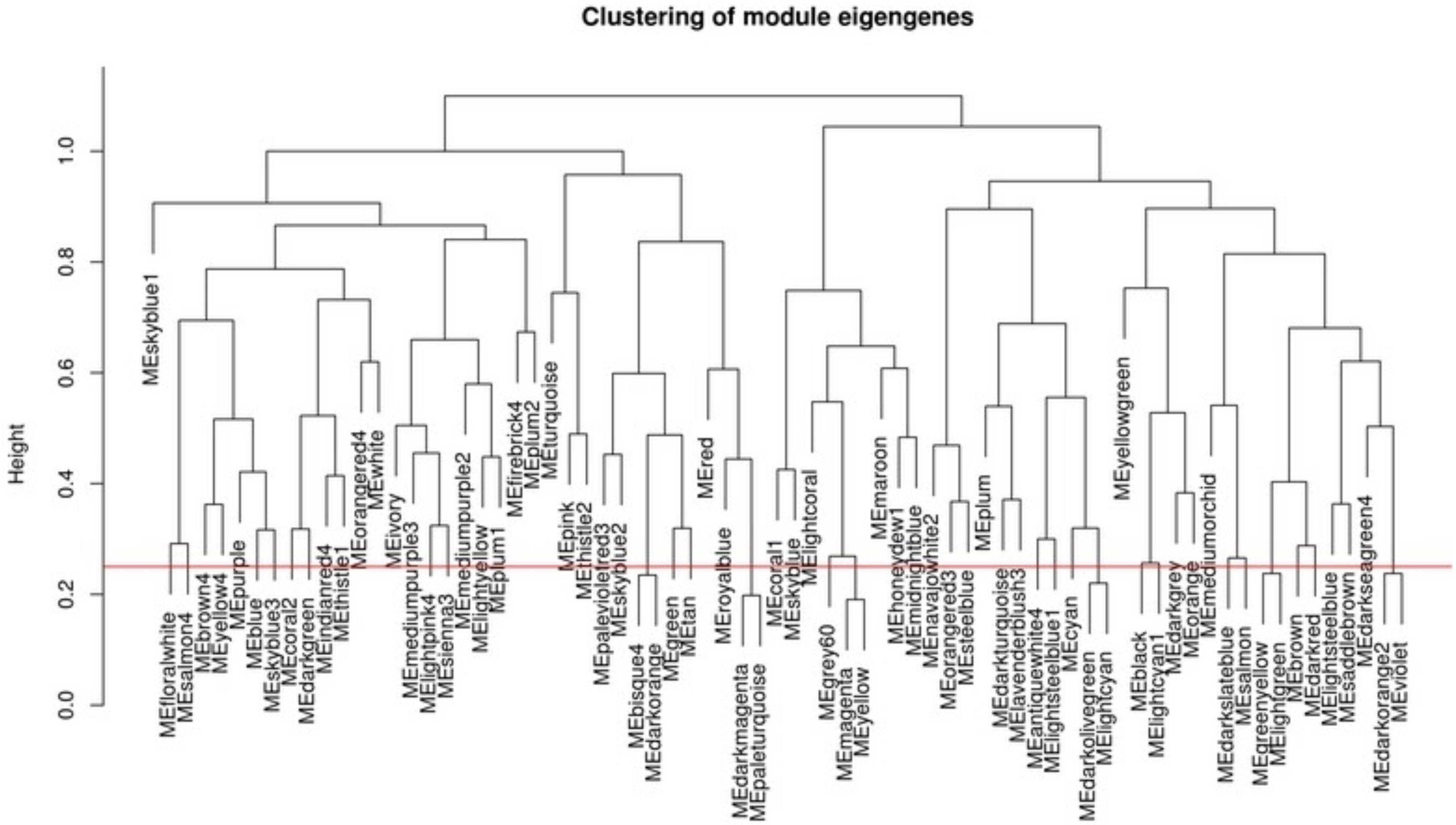


Clustering eigengenes

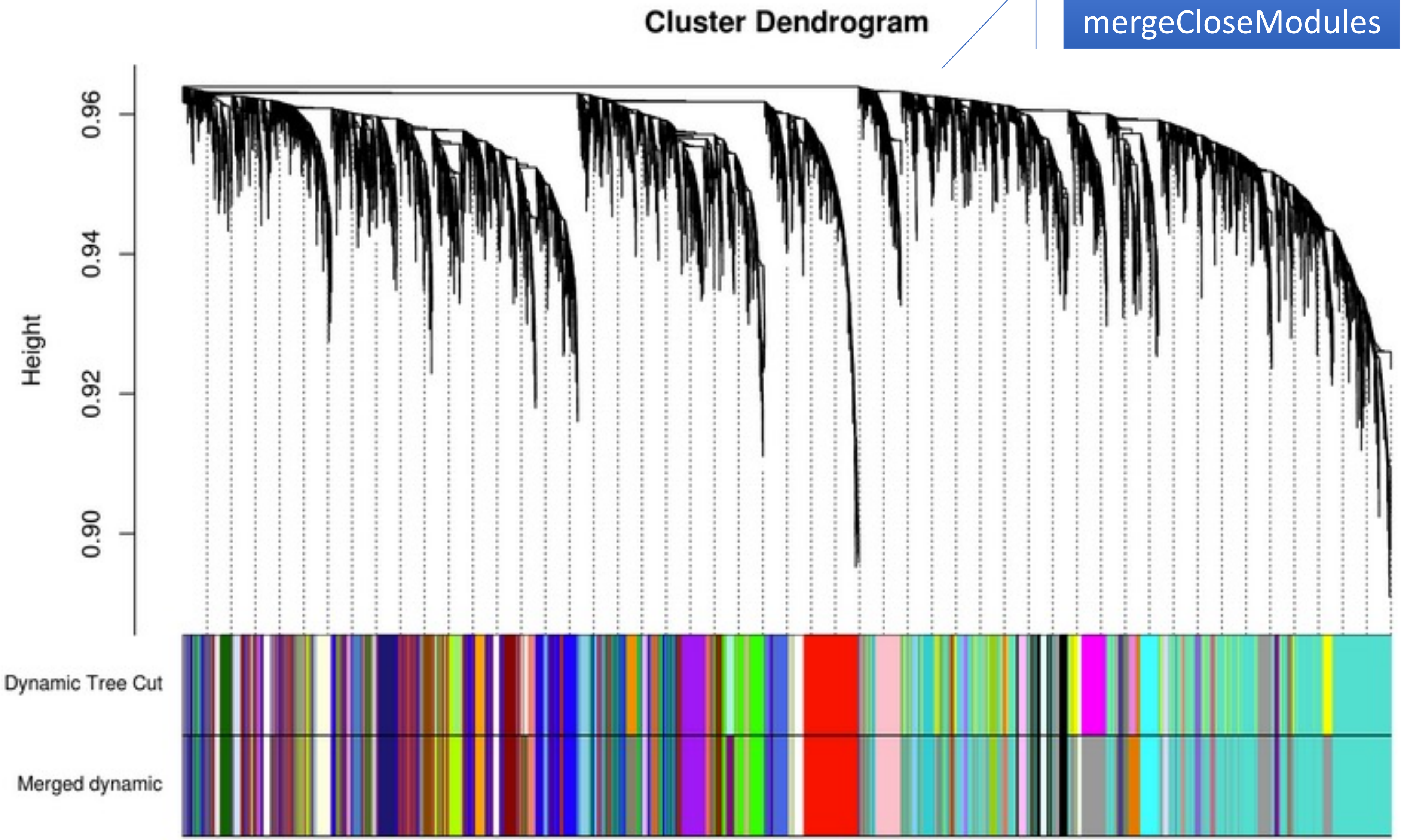
```
# Calculate eigengenes
MEList = moduleEigengenes(datExpr, colors = dynamicColors)
MEs = MEList$eigengenes
# Calculate dissimilarity of module eigengenes
MEDiss = 1-cor(MEs)
# Cluster module eigengenes
METree = hclust(as.dist(MEDiss), method = "average")
# Plot the result
#sizeGrWindow(7, 6)
#png(paste0(work_plots, "Tree_eigengenes.png"), width = 12, height = 8, res = 300, units = "in")
plot(METree, main = "Clustering of module eigengenes",
xlab = "", sub = "")
#We choose a height cut of 0.25, corresponding to correlation of 0.75, to merge
MEDissThres = 0.25
# Plot the cut line into the dendrogram
abline(h=MEDissThres, col = "red")
#dev.off()
```

Clustering eigengenes

Height cut of **0.25**, corresponding to correlation of **0.75** to merge



Merge modules

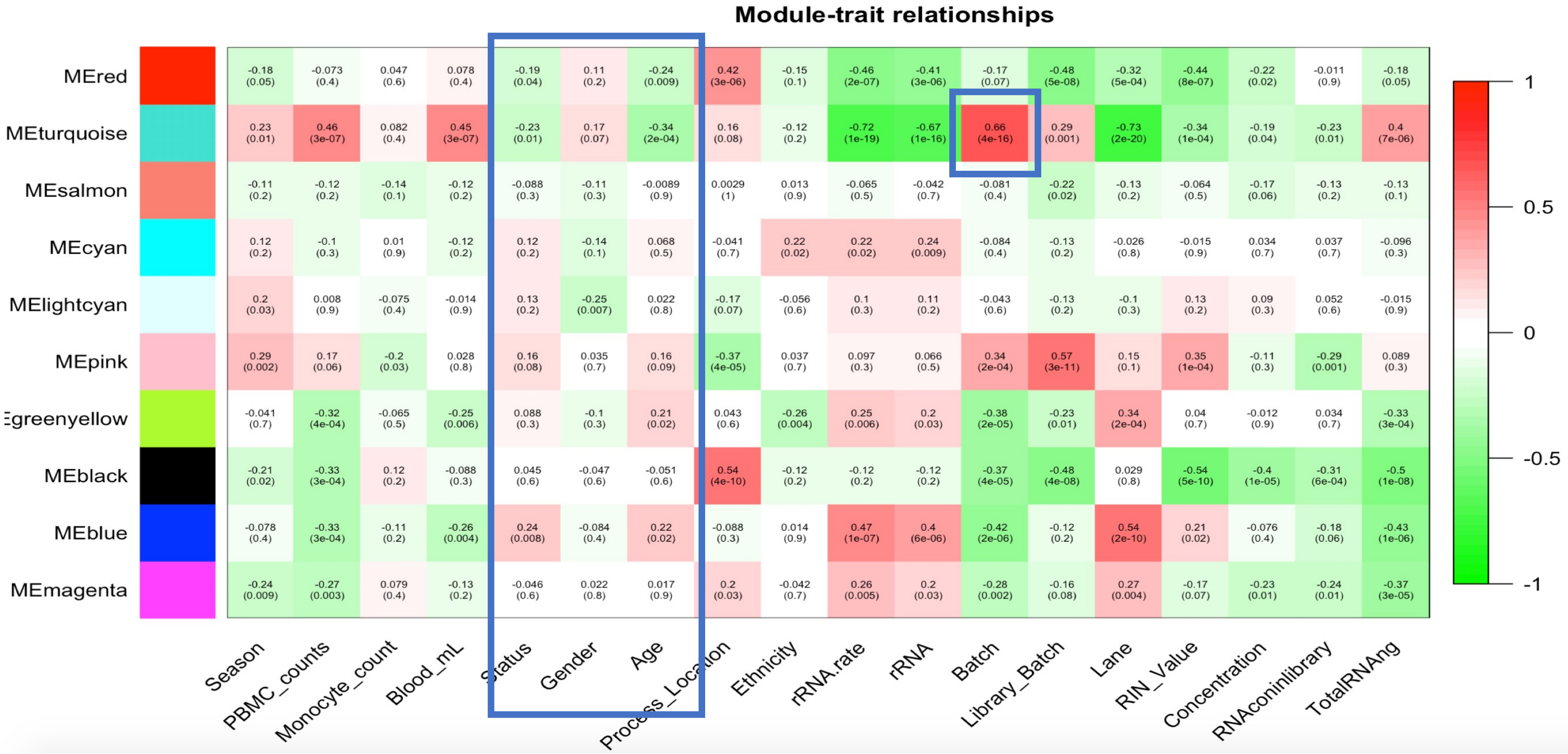


Using the mergeCloseModules

Module-trait relationships
TPM data without correction

(p values)
Pearson correlation

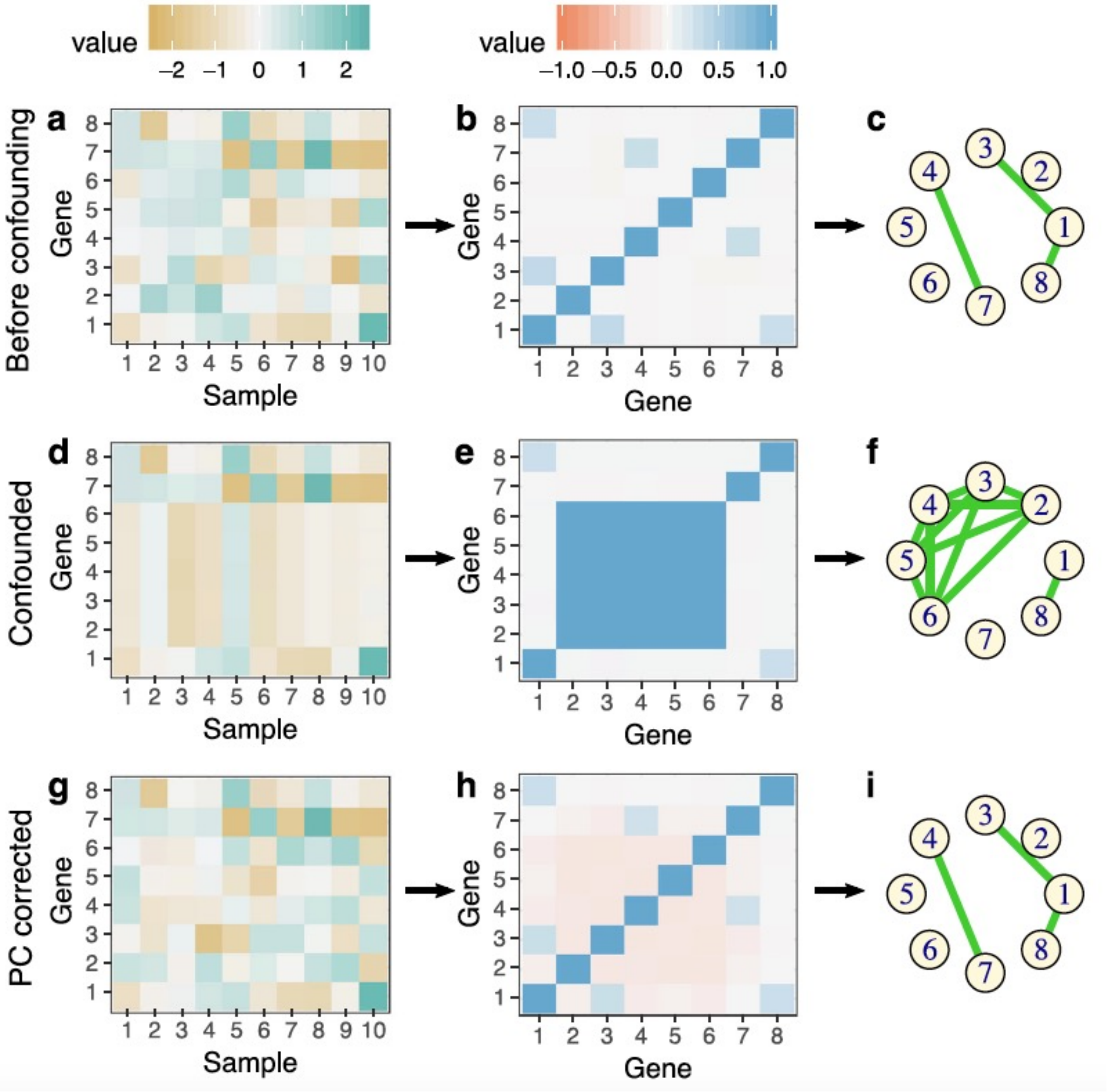
Data not adjusted



Module-trait relationships

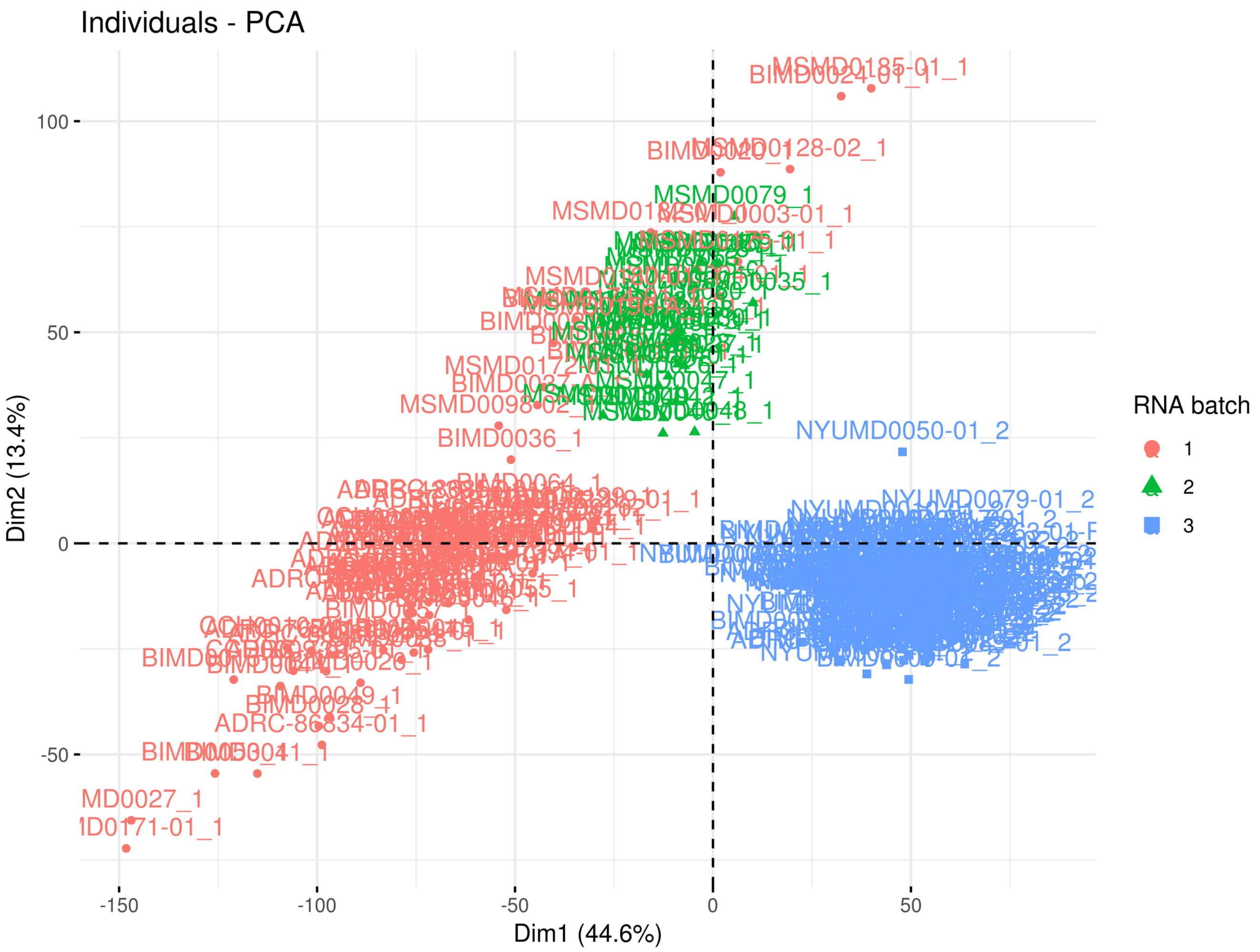
	Season	PBMC_counts	Monocyte_count	Blood_mL	Status	Gender	Age	Process_Location	Ethnicity	rRNA.rate	rRNA	Batch	Library_Batch	Lane	RIN_Value	Concentration	RNAconinlibrary	TotalRNAng
MEred	-0.18 (0.05)	-0.073 (0.4)	0.047 (0.6)	0.078 (0.4)	-0.19 (0.04)	0.11 (0.2)	-0.24 (0.009)	0.42 (3e-06)	-0.15 (0.1)	-0.46 (2e-07)	-0.41 (3e-06)	-0.17 (0.07)	-0.48 (5e-08)	-0.32 (5e-04)	-0.44 (8e-07)	-0.22 (0.02)	-0.011 (0.9)	-0.18 (0.05)
MEturquoise	0.23 (0.01)	0.46 (3e-07)	0.082 (0.4)	0.45 (3e-07)	-0.23 (0.01)	0.17 (0.07)	-0.34 (2e-04)	0.16 (0.08)	-0.12 (0.2)	-0.72 (1e-19)	-0.67 (1e-16)	0.66 (4e-16)	0.29 (0.001)	-0.73 (2e-20)	-0.34 (1e-04)	-0.19 (0.04)	-0.23 (0.01)	0.4 (7e-06)
MEsalmon	-0.11 (0.2)	-0.12 (0.2)	-0.14 (0.1)	-0.12 (0.2)	-0.088 (0.3)	-0.11 (0.3)	-0.0089 (0.9)	0.0029 (1)	0.013 (0.9)	-0.065 (0.5)	-0.042 (0.7)	-0.081 (0.4)	-0.22 (0.02)	-0.13 (0.2)	-0.064 (0.5)	-0.17 (0.06)	-0.13 (0.2)	-0.13 (0.1)
MEcyan	0.12 (0.2)	-0.1 (0.3)	0.01 (0.9)	-0.12 (0.2)	0.12 (0.2)	-0.14 (0.1)	0.068 (0.5)	-0.041 (0.7)	0.22 (0.02)	0.22 (0.02)	0.24 (0.009)	-0.084 (0.4)	-0.13 (0.2)	-0.026 (0.8)	-0.015 (0.9)	0.034 (0.7)	0.037 (0.7)	-0.096 (0.3)
MElightcyan	0.2 (0.03)	0.008 (0.9)	-0.075 (0.4)	-0.014 (0.9)	0.13 (0.2)	-0.25 (0.007)	0.022 (0.8)	-0.17 (0.07)	-0.056 (0.6)	0.1 (0.3)	0.11 (0.2)	-0.043 (0.6)	-0.13 (0.2)	-0.1 (0.3)	0.13 (0.2)	0.09 (0.3)	0.052 (0.6)	-0.015 (0.9)
MEpink	0.29 (0.002)	0.17 (0.06)	-0.2 (0.03)	0.028 (0.8)	0.16 (0.08)	0.035 (0.7)	0.16 (0.09)	-0.37 (4e-05)	0.037 (0.7)	0.097 (0.3)	0.066 (0.5)	0.34 (2e-04)	0.57 (3e-11)	0.15 (0.1)	0.35 (1e-04)	-0.11 (0.3)	-0.29 (0.001)	0.089 (0.3)
Egreenyellow	-0.041 (0.7)	-0.32 (4e-04)	-0.065 (0.5)	-0.25 (0.006)	0.088 (0.3)	-0.1 (0.3)	0.21 (0.02)	0.043 (0.6)	-0.26 (0.004)	0.25 (0.006)	0.2 (0.03)	-0.38 (2e-05)	-0.23 (0.01)	0.34 (2e-04)	0.04 (0.7)	-0.012 (0.9)	0.034 (0.7)	-0.33 (3e-04)
MEblack	-0.21 (0.02)	-0.33 (3e-04)	0.12 (0.2)	-0.088 (0.3)	0.045 (0.6)	-0.047 (0.6)	-0.051 (0.6)	0.54 (4e-10)	-0.12 (0.2)	-0.12 (0.2)	-0.12 (0.2)	-0.37 (4e-05)	-0.48 (4e-08)	0.029 (0.8)	-0.54 (5e-10)	-0.4 (1e-05)	-0.31 (6e-04)	-0.5 (1e-08)
MEblue	-0.078 (0.4)	-0.33 (3e-04)	-0.11 (0.2)	-0.26 (0.004)	0.24 (0.008)	-0.084 (0.4)	0.22 (0.02)	-0.088 (0.3)	0.014 (0.9)	0.47 (1e-07)	0.4 (6e-06)	-0.42 (2e-06)	-0.12 (0.2)	0.54 (2e-10)	0.21 (0.02)	-0.076 (0.4)	-0.18 (0.06)	-0.43 (1e-06)
MEmagenta	-0.24 (0.009)	-0.27 (0.003)	0.079 (0.4)	-0.13 (0.2)	-0.046 (0.6)	0.022 (0.8)	0.017 (0.9)	0.2 (0.03)	-0.042 (0.7)	0.26 (0.005)	0.2 (0.03)	-0.28 (0.002)	-0.16 (0.08)	0.27 (0.004)	-0.17 (0.07)	-0.23 (0.01)	-0.24 (0.01)	-0.37 (3e-05)

Artifacts in reconstruction of gene co-expression networks

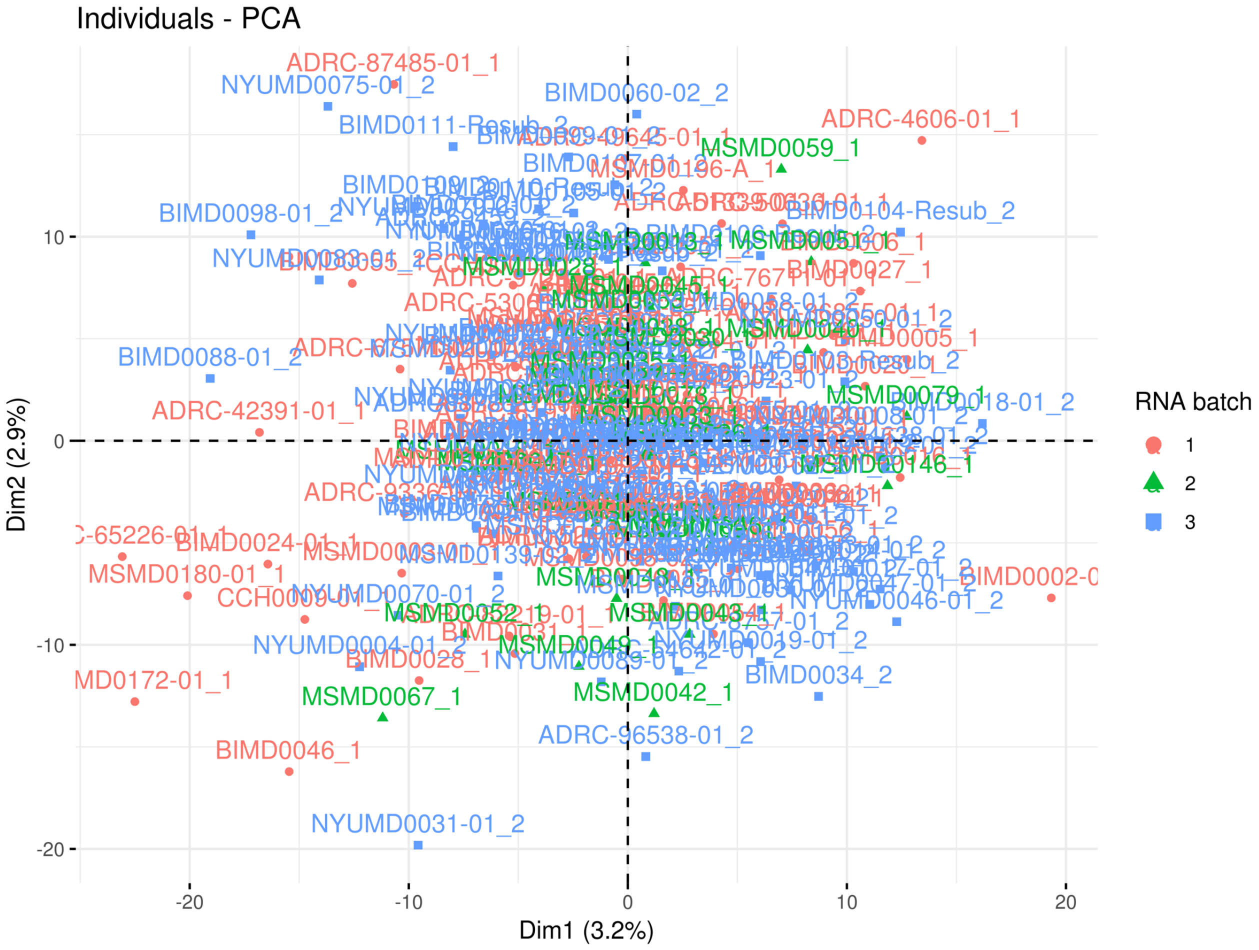


Parsana et al. Genome Biology, 2019.

Noise in gene expression



Noise in gene expression



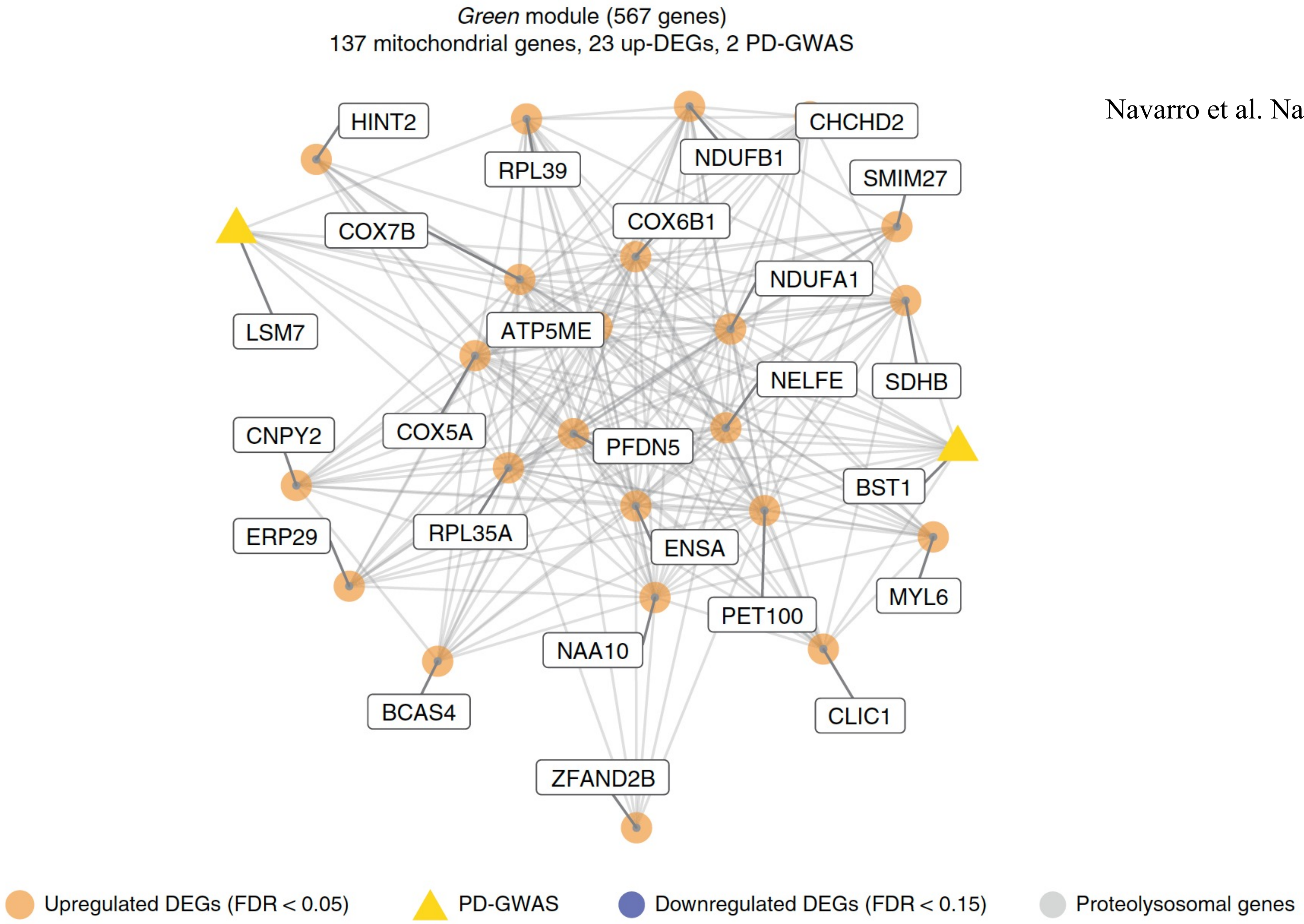
Tasks

- Q1: What are the input data for the WGCNA pipeline?

- Q2: Why is it so important to take care of noise in the data?

- Q3: What is the hypothesis behind a co-expression network?

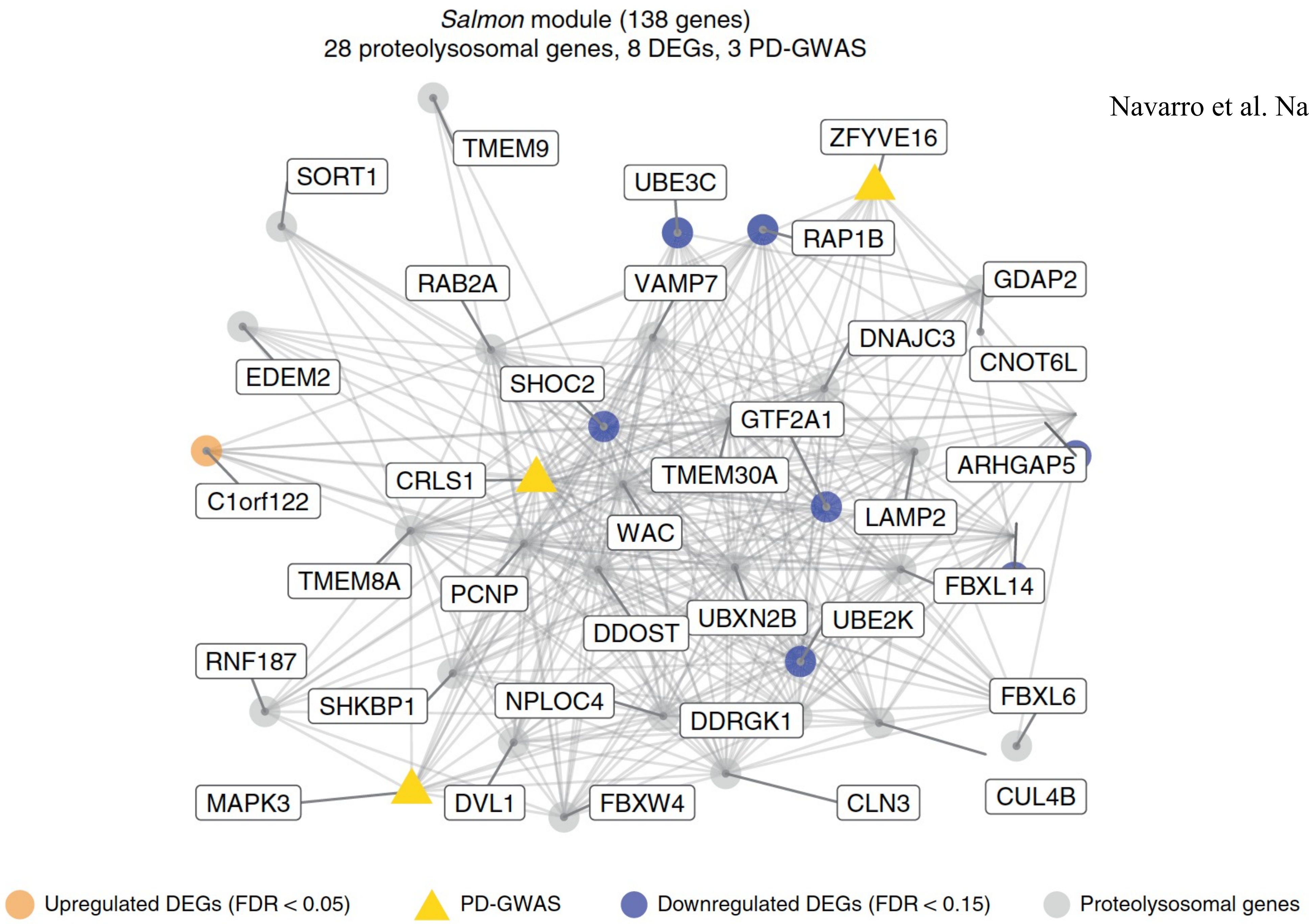
Networks for the *“Myeloid cells in Neurodegenerative Diseases”* (**MyND**) project



Green module (567 genes)
137 mitochondrial genes, 23 up-DEGs, 2 PD-GWAS

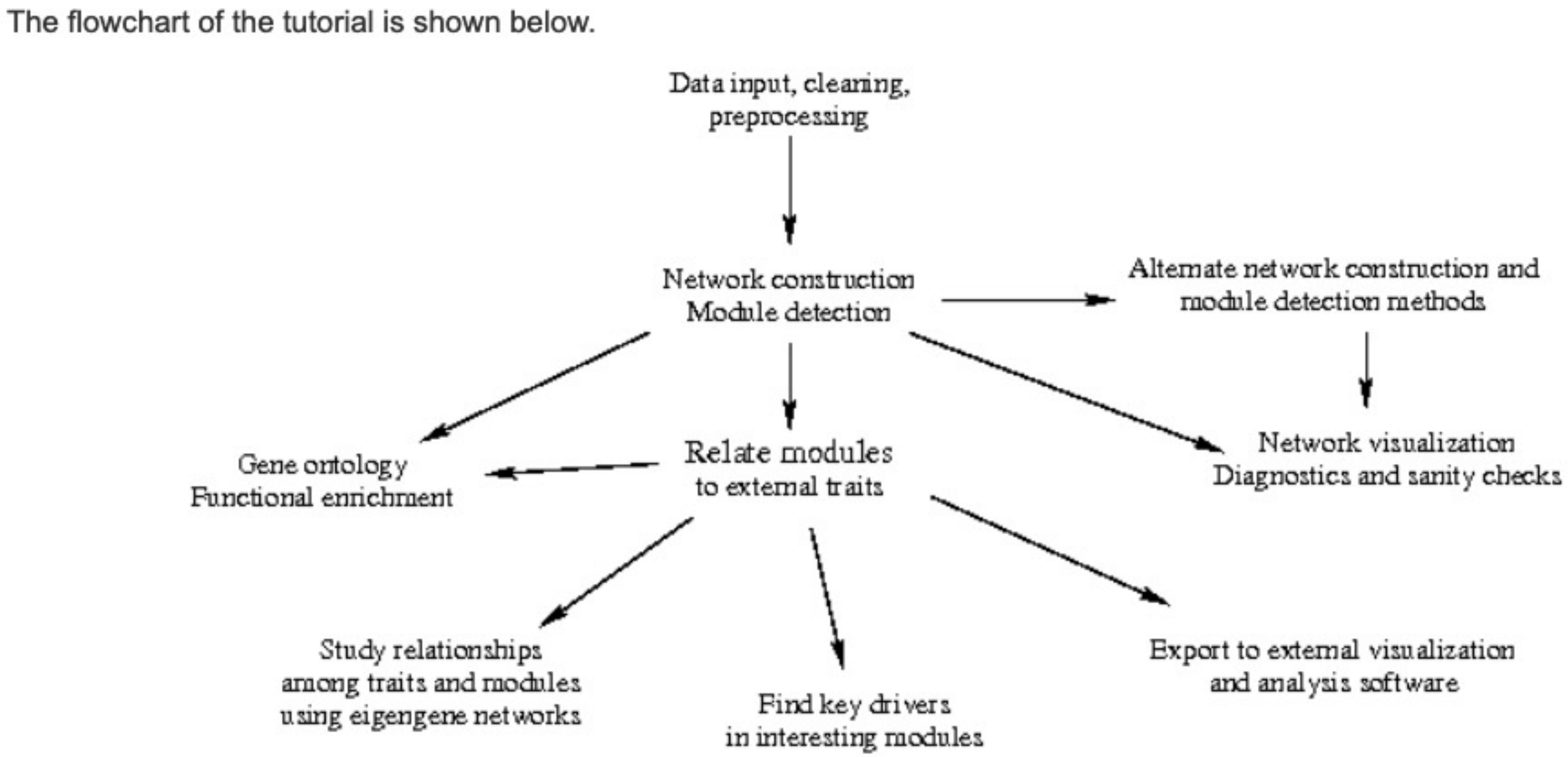
Navarro et al. Nat Aging, 2021.

Navarro et al. Nat Aging, 2021.



Tutorials to follow

WGCNA tutorials

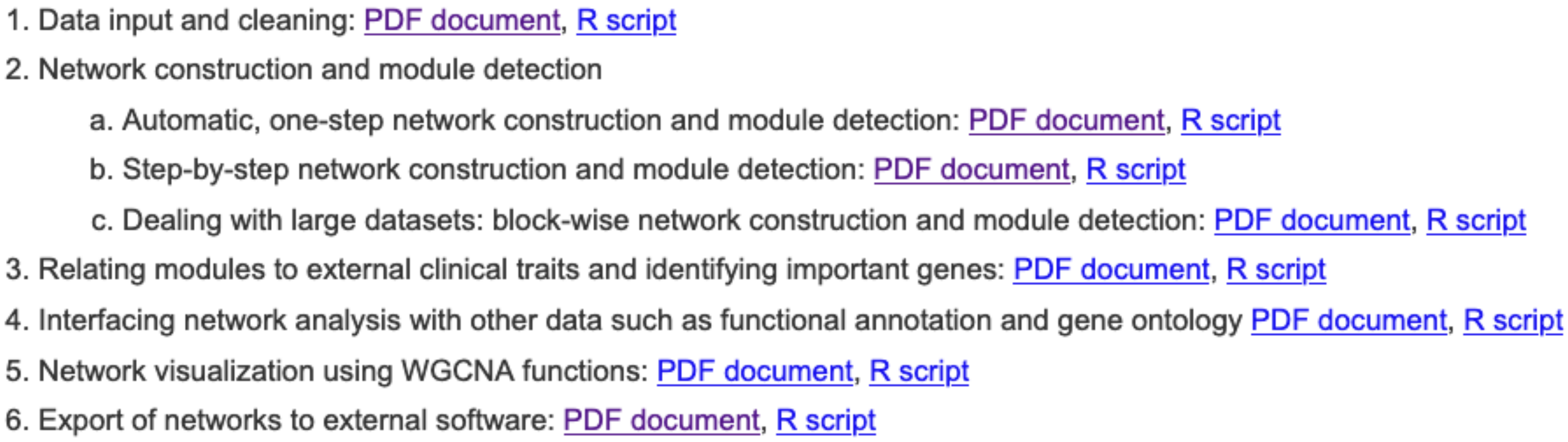


The flowchart of the tutorial is shown below.

Data input, cleaning, preprocessing
Network construction Module detection
Alternate network construction and module detection methods
Gene ontology Functional enrichment
Relate modules to external traits
Network visualization Diagnostics and sanity checks
Study relationships among traits and modules using eigengene networks
Find key drivers in interesting modules
Export to external visualization and analysis software

https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

WGCNA tutorials



1. Data input and cleaning: PDF document, R script
2. Network construction and module detection
    a. Automatic, one-step network construction and module detection: PDF document, R script
    b. Step-by-step network construction and module detection: PDF document, R script
    c. Dealing with large datasets: block-wise network construction and module detection: PDF document, R script
3. Relating modules to external clinical traits and identifying important genes: PDF document, R script
4. Interfacing network analysis with other data such as functional annotation and gene ontology PDF document, R script
5. Network visualization using WGCNA functions: PDF document, R script
6. Export of networks to external software: PDF document, R script

https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/

To do – Networks final project

R code:

WGCNA: monocytes dataset of individuals diagnosed with Parkinson's Disease. https://rushalz.github.io/Intro_Systems_Biology/WGCNA_rnaseq_monocytes.html

WGCNA: Thoracic spinal cord RNASeq of individuals diagnosed with Amyotrophic Lateral Sclerosis. https://rushalz.github.io/Intro_Systems_Biology/WGCNA_rnaseq.html

Thank you!

katiaplopes@gmail.com
@lopeskp