Weighted Correlation Network Analysis (WGCNA)

SIB course, Nov. 15-17, 2016
Introduction to Biological Network Analysis

Leonore Wigger
with Frédéric Burdet and Mark Ibberson

Agenda

Overview of WGCNA

Theory 1: Weighted correlation network, split into modules

Practical

LUNCH BREAK

Theory 2: Identify modules and genes of interest

Practical

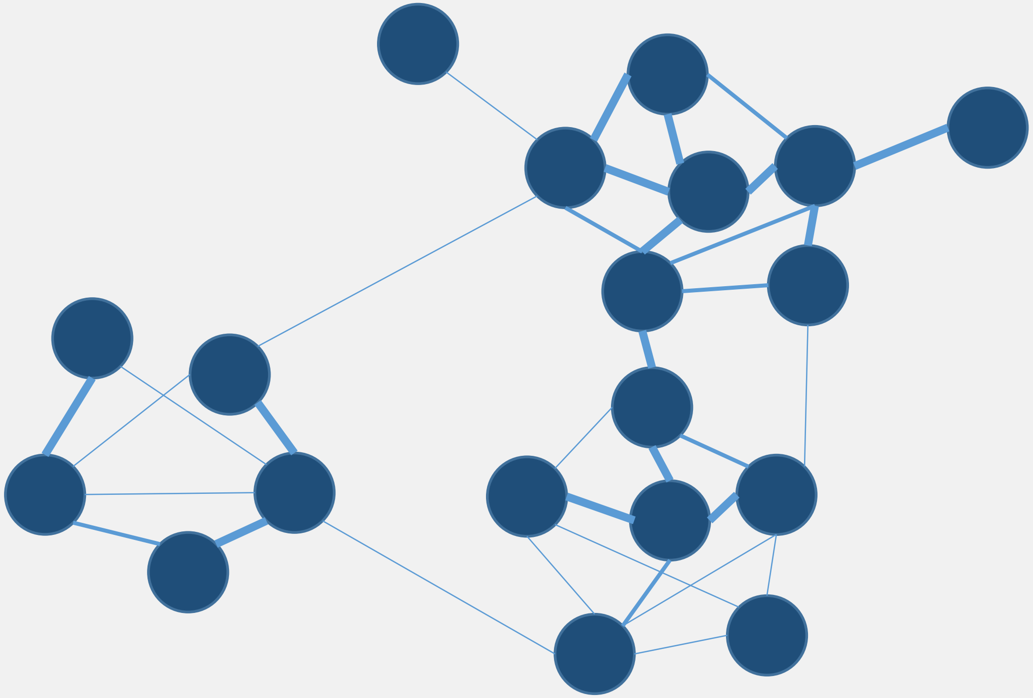
Data for WGCNA

Gene expression data (microarray or RNA-Seq)
Recommendation: at least 20 individuals

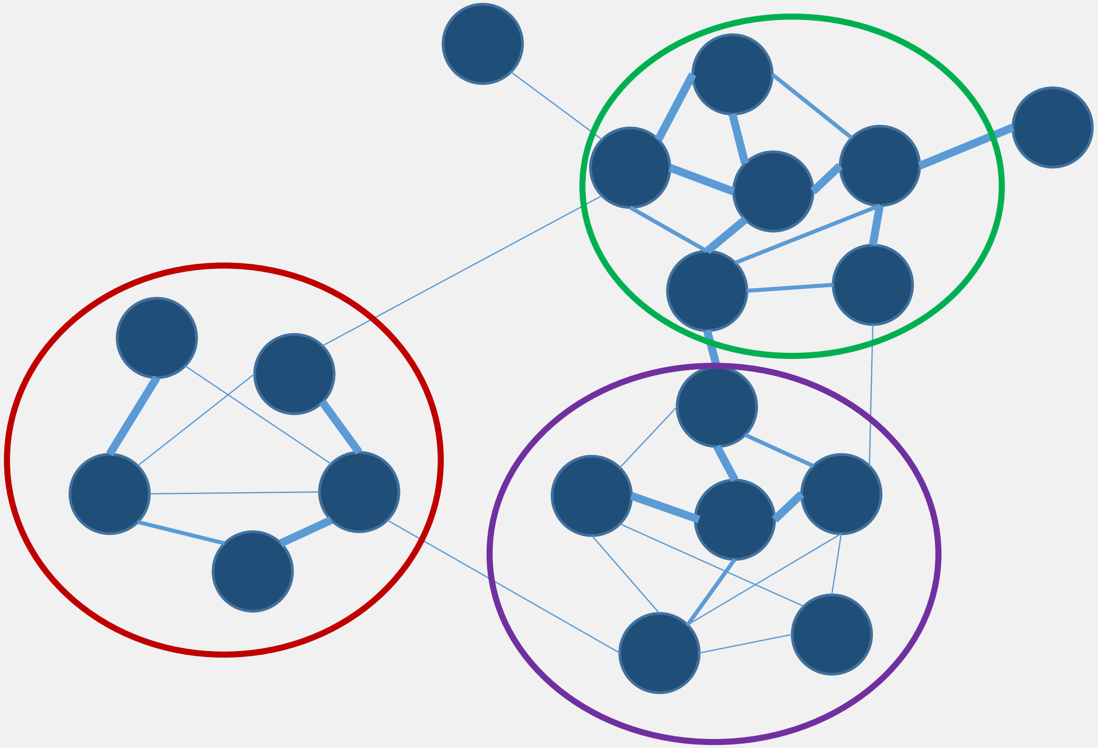
Clinical/phenotypical traits from the same individuals   (optional)
e.g. weight, insulin level, glucose level

Aims of WGCNA

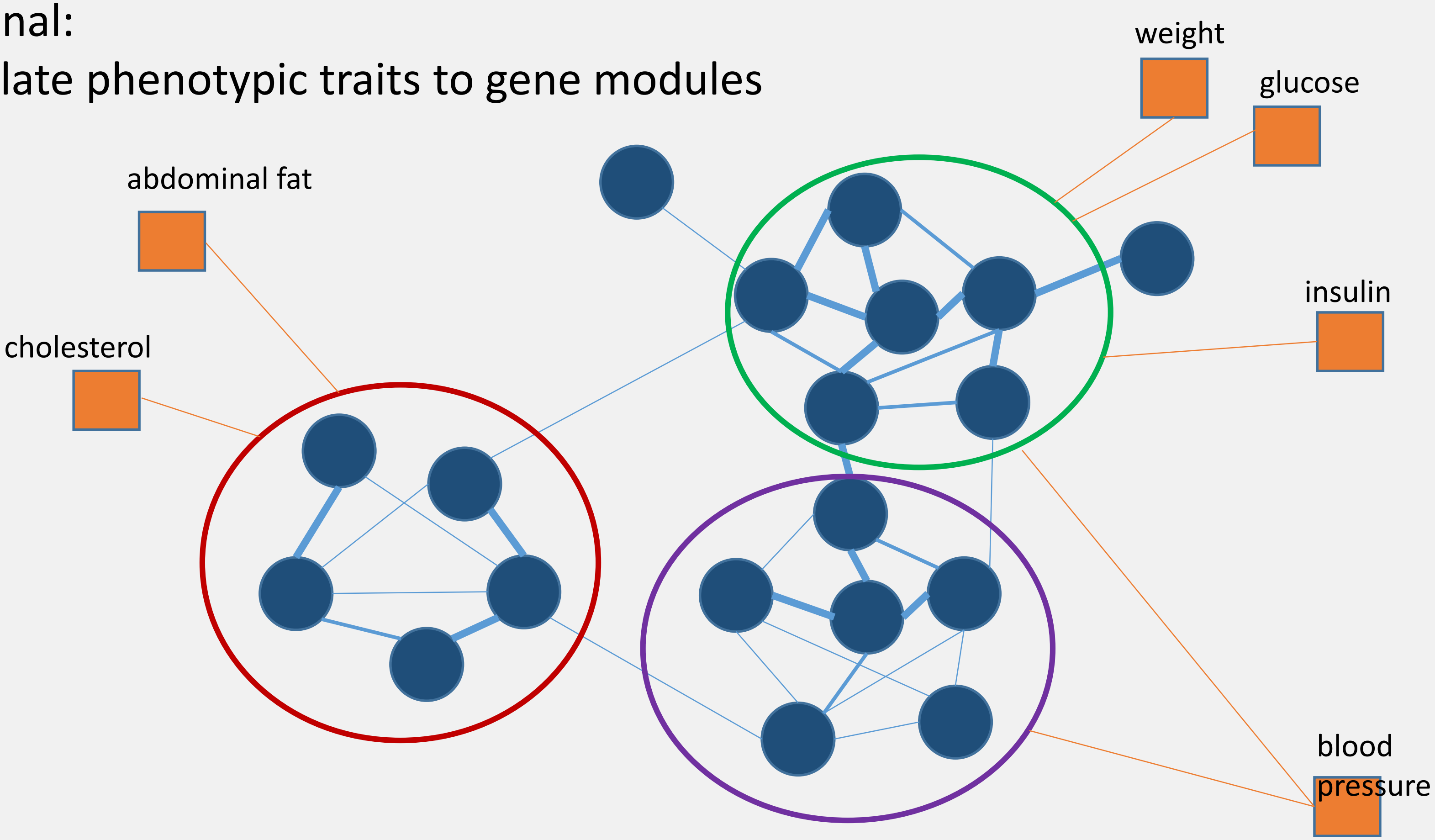
Construct a gene-gene similarity network



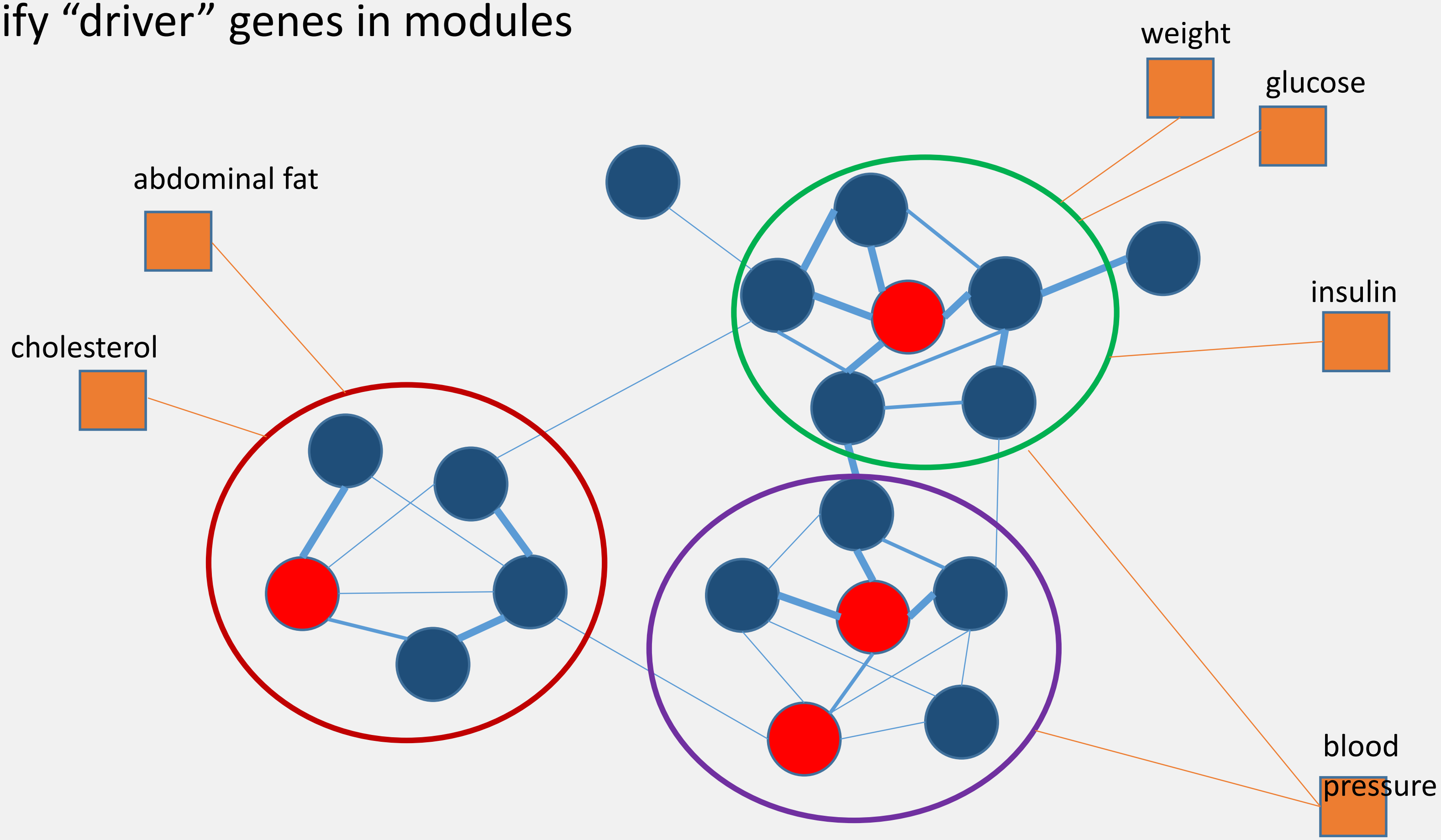
Divide network into modules
(Group genes with similar expression)



Optional:
Correlate phenotypic traits to gene modules

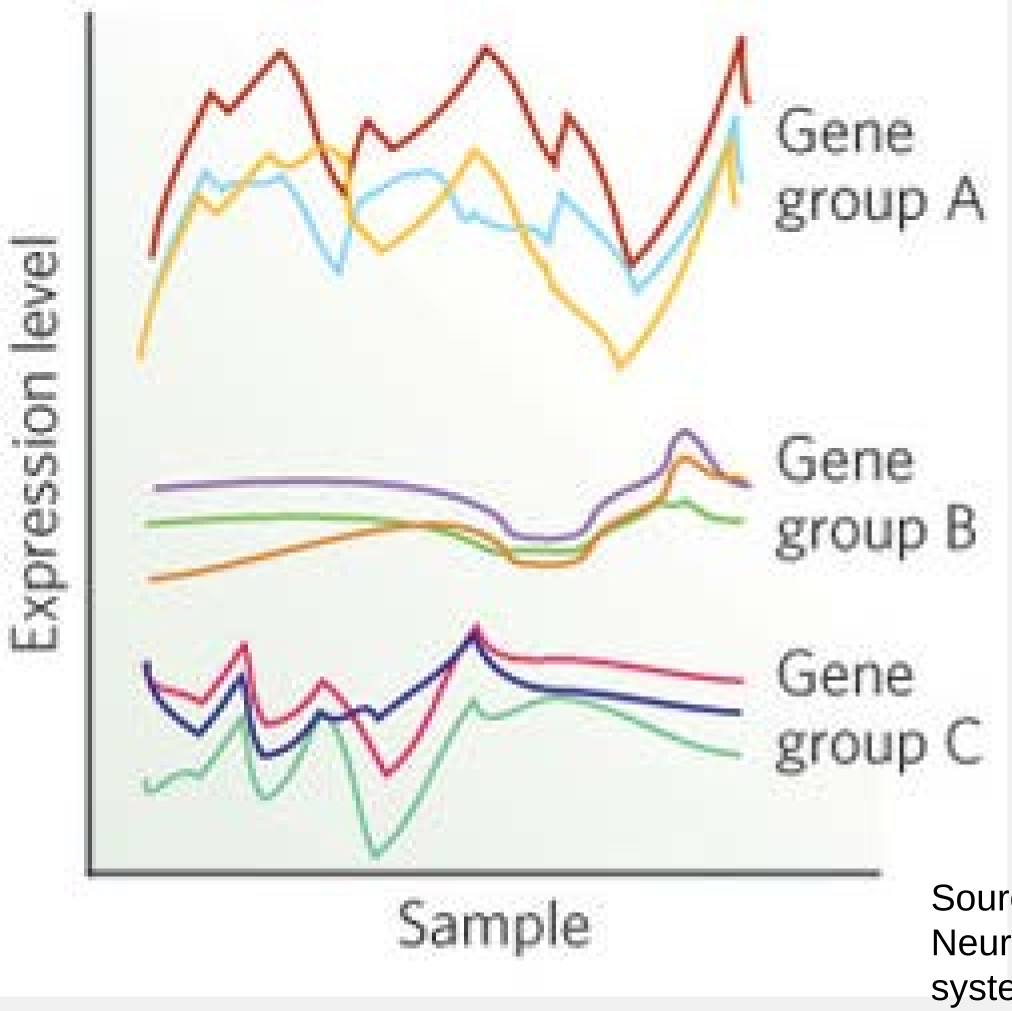


Identify “driver” genes in modules



Modules found in WGCNA:

Groups of co-expressed genes (with similar expression profiles over a large group of individuals)



Source: Daniel H. Geschwind & Genevieve Konopka. Neuroscience in the era of functional genomics and systems biology, Nature 461, 908-915

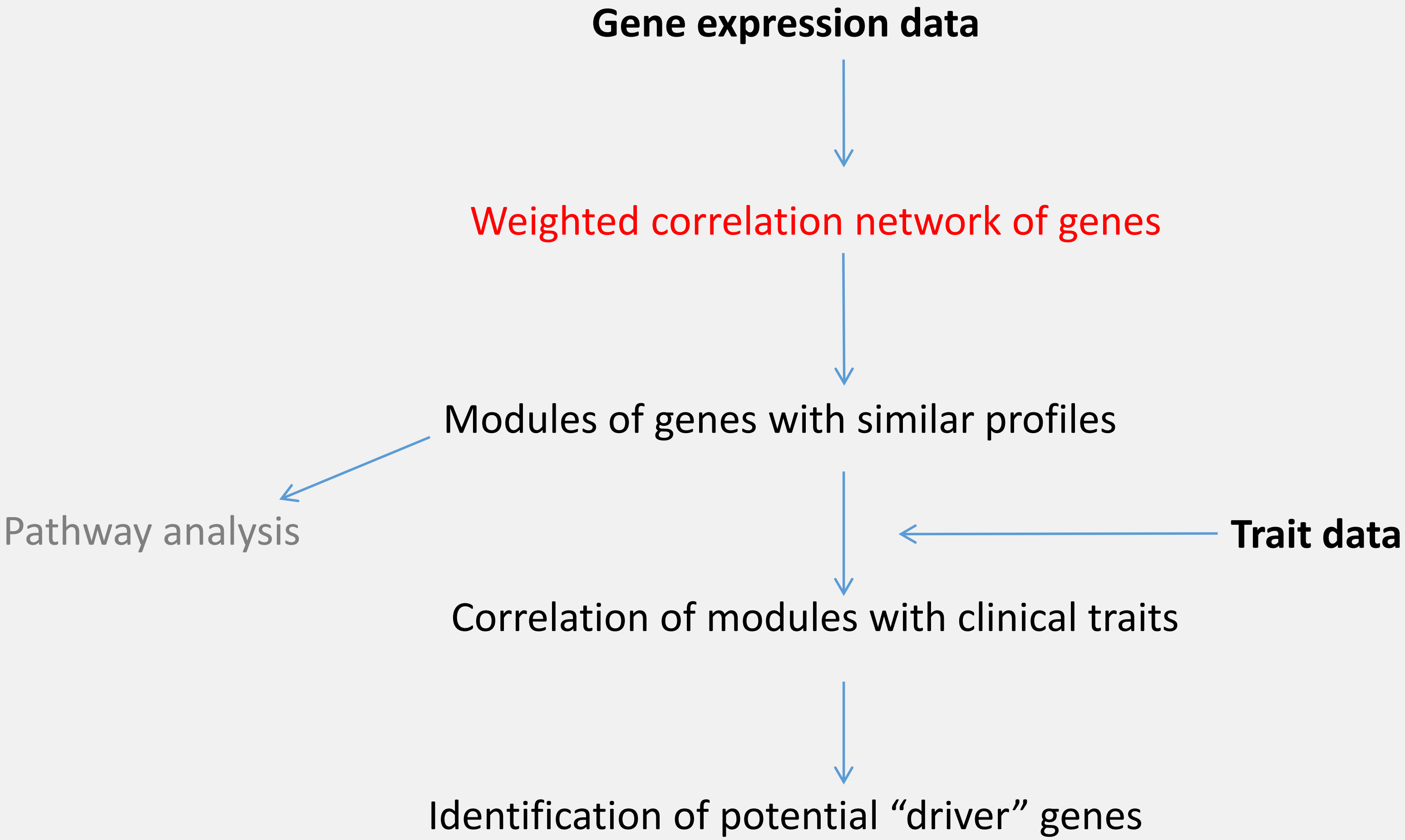
Central Hypothesis:

Genes with **similar expression patterns** are of interest because they may be

- tightly co-regulated
- functionally related
- members of the same pathway

WGCNA is a **guilt-by-association** approach:
Encourages hypotheses about genes based on their close network neighbors.

Workflow



Gene expression data

Weighted correlation network of genes

Modules of genes with similar profiles

Pathway analysis

Trait data

Correlation of modules with clinical traits

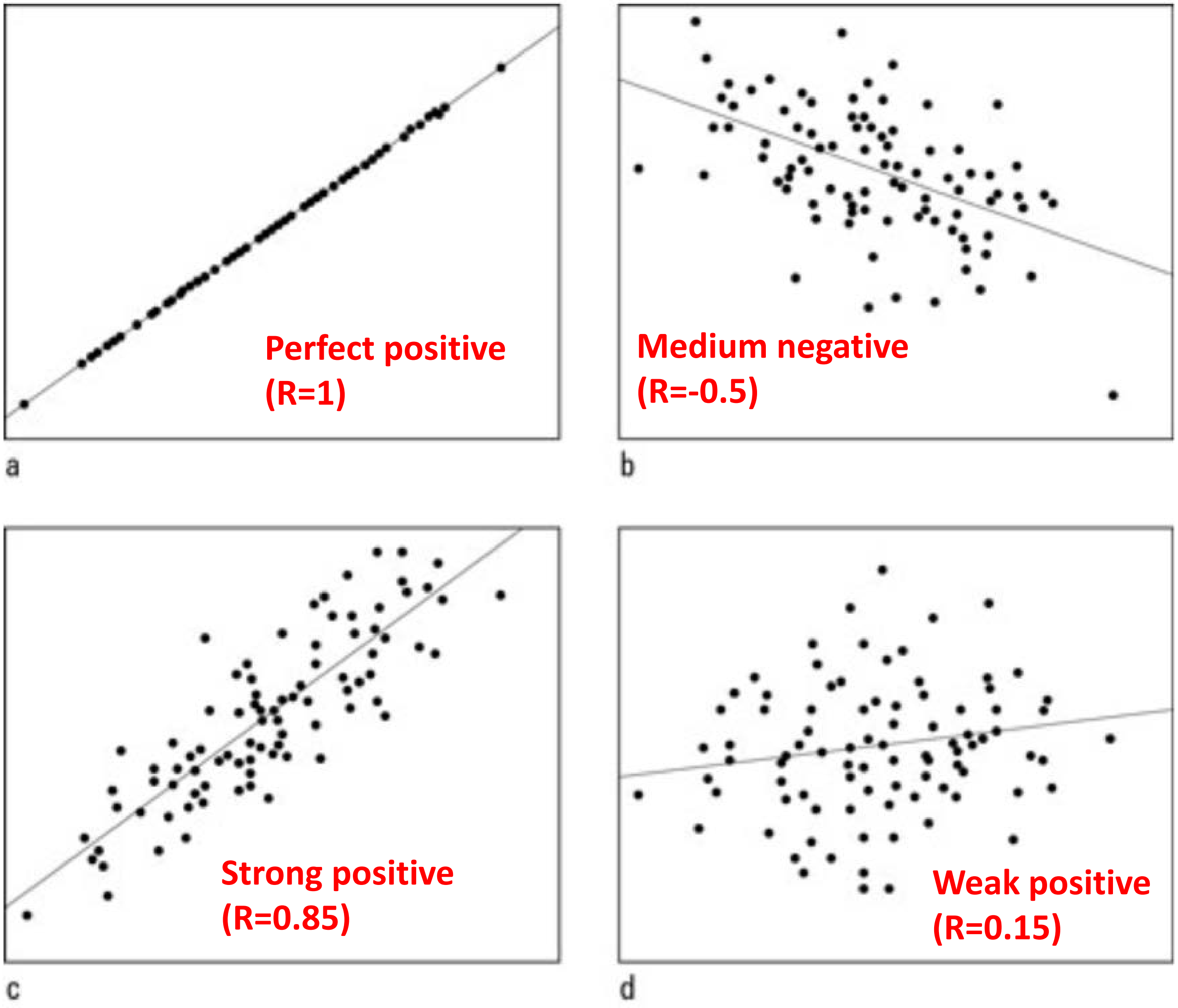
Identification of potential “driver” genes

Construct weighted correlation network

Correlation:
A statistical measure for the extent to which two variables fluctuate together.

Positive correlation: variables increase/decrease together
Negative correlation: variables increase/decrease in opposing direction

Correlation examples (Pearson correlation coefficient R)



Scatterplots with correlations of a) +1.00; b) –0.50; c) +0.85; and d) +0.15.

http://www.dummies.com/how-to/content/how-to-interpret-a-correlation-coefficient-r.html

Multiple measures for correlation exist

Implemented in WGCNA:

Pearson	(function `cor`)
Spearman, Kendall	(function `cor`)
biweight midcorrelation	(function `bicor`)

The Basis of WGCNA:
Weighted Correlation Network of Genes

Adjacencies

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

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

Effect of raising correlation to a power:

Amplifies disparity between strong and weak correlations

Example: Power term $\boldsymbol{\beta} = 4$

Strong corr.

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

Weak corr.

The Basis of WGCNA:
Weighted Correlation Network of Genes

Adjacencies

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

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

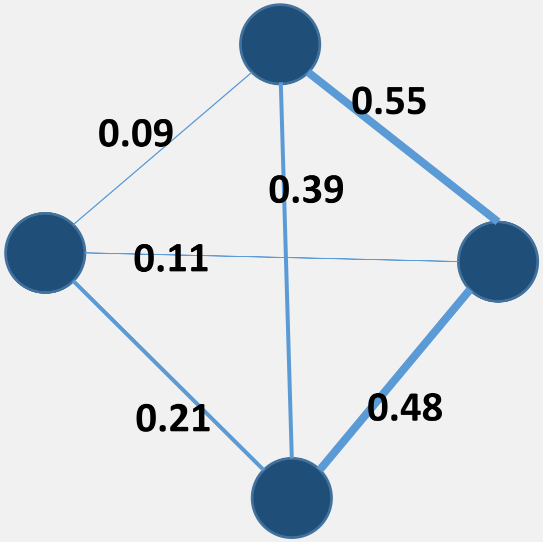
Adjacency matrix of 4 genes

$\boldsymbol{a_{i,j}}$	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

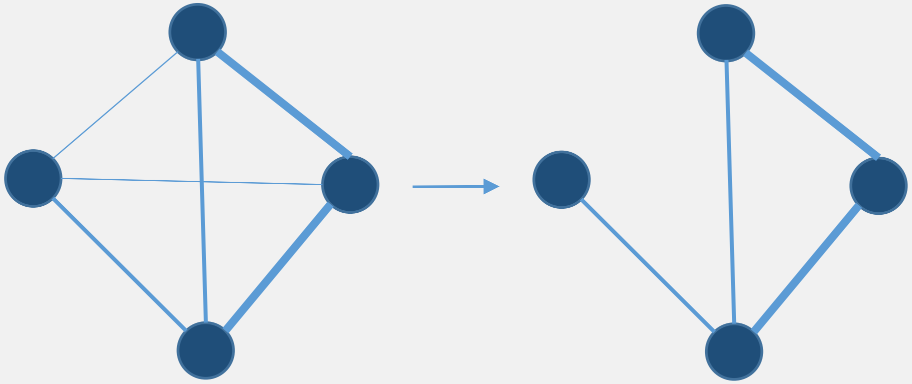
Network

Construct a fully connected network;
Genes as nodes, $\boldsymbol{a_{i,j}}$ as edge weights.

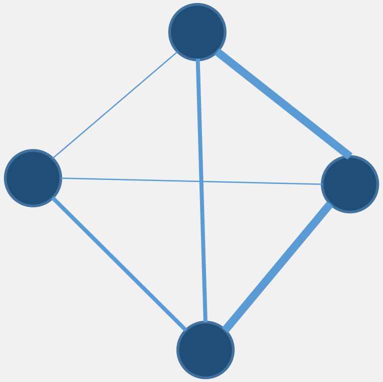
high correlation – strong connection
low correlation – weak connection



For visualizations, set a threshold on edge weight and **remove the weakest links.**



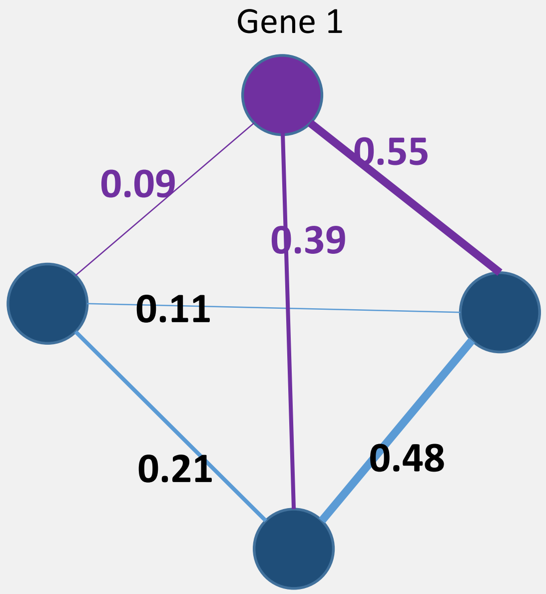
In most computations, work with all edges of the **fully connected network**.



Connectivity (degree) in a weighted network

Connectivity of a gene:

Sum of the weights of all edges connecting to this gene



Example:
Connectivity of gene 1:
0.55 + 0.39 + 0.09 = 1.03

Weighted Correlation: Decisions to make

Selecting a network type

unsigned, signed etc.

Choosing a correlation method

Pearson, Spearman, biweight midcorr. etc.

Picking a Power term

1, …, 20 etc.

Selecting a network type

Unsigned Network (Default)

No differentiation between positive and negative correlations.

Use this if negative correlation are of interest

Signed Hybrid Network

Only positive correlations are taken into account. Negative correlations are set to 0.

Use this if negative correlations are NOT of interest.

Signed Network

[not covered in course, use as an alternative to signed hybrid network]

Choosing a correlation method

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

Fastest, but sensitive to outliers:

Pearson correlation `cor(x)`

"standard" measure of linear correlation

Less sensitive to outliers but much slower:

Biweight mid-correlation `bicor(x)`

robust, recommended by the authors for most situations

[needs modification for correlations involving binary/categorical variables]

Spearman correlation `cor(x, method="spearman")`

rank-based, works even if relationship is not linear

less sensitive to gene expression differences

[can be used as-is for correlations involving binary/categorical variables]

Default correlation method in WGCNA: cor (Pearson).

Caveat: use it only if there are no outliers, or for exercises/tutorials.

Picking a power term

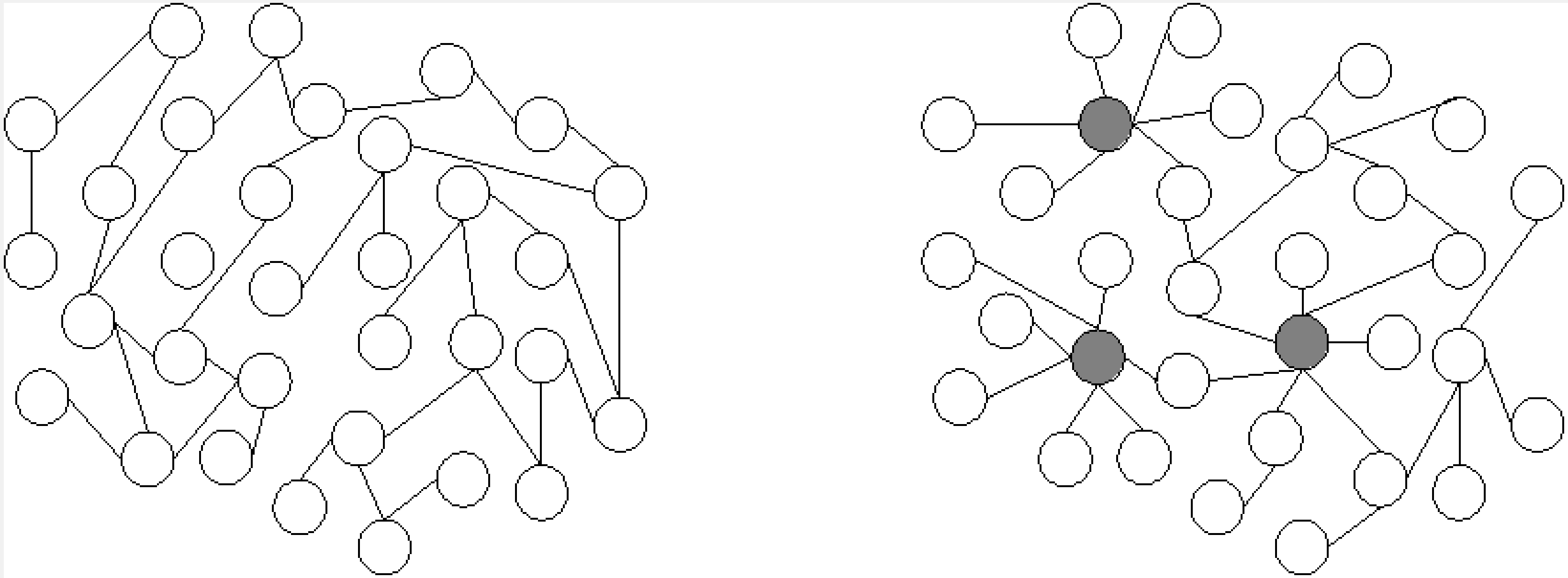
$$a_{i,j} = |cor\ (i\ , j\ )|^{\beta}$$

Selection criterion: Pick lowest possible β that leads to an approximately scale-free network topology

- few nodes with many connections (”hubs”)
- many nodes with few connections

Degree distribution follows a power law:

the probability for a node of having k connections is k^{γ}



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

Source: Carlos Castillo: Effective Web Crawling,
PhD Thesis, University of Chile, 2004.
(obtained from Wikimedia Commons)

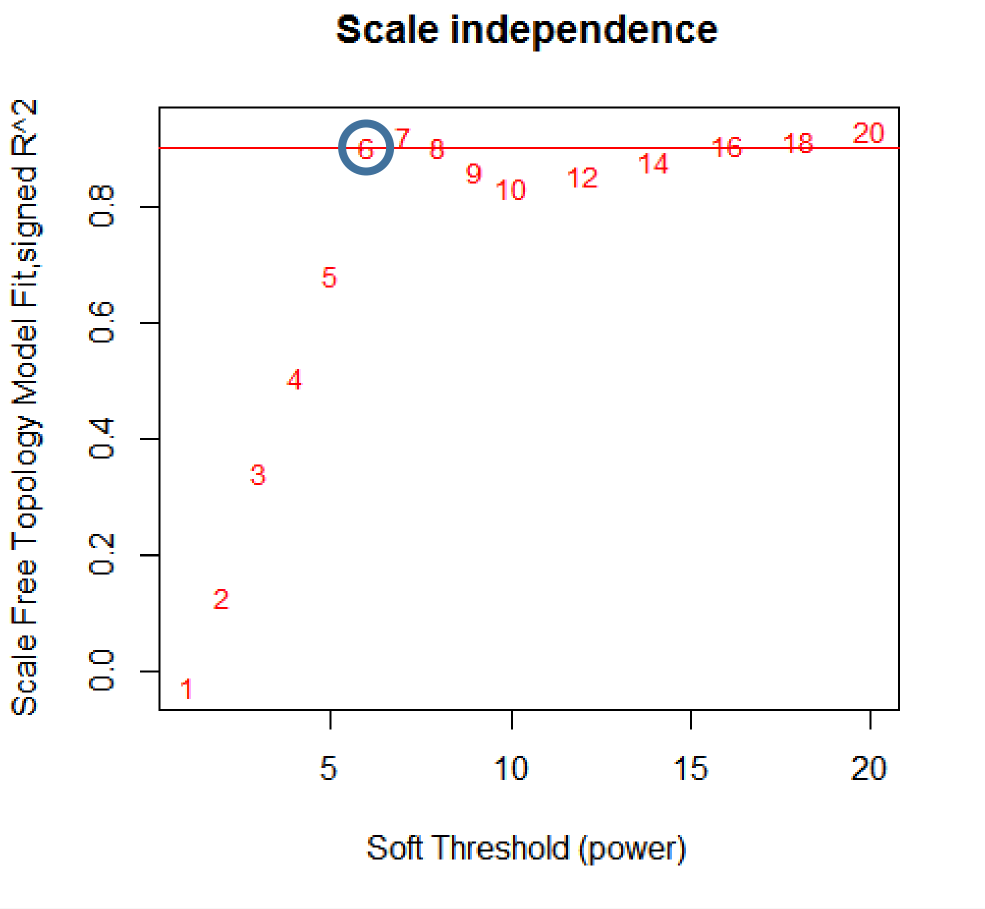
Why scale-free network topology?

Background:
Barábasi et al. found many types of network in many domains to be approximately scale-free, including metabolic and protein interaction

Aim in WGCNA:
Build a biologically “realistic” network.

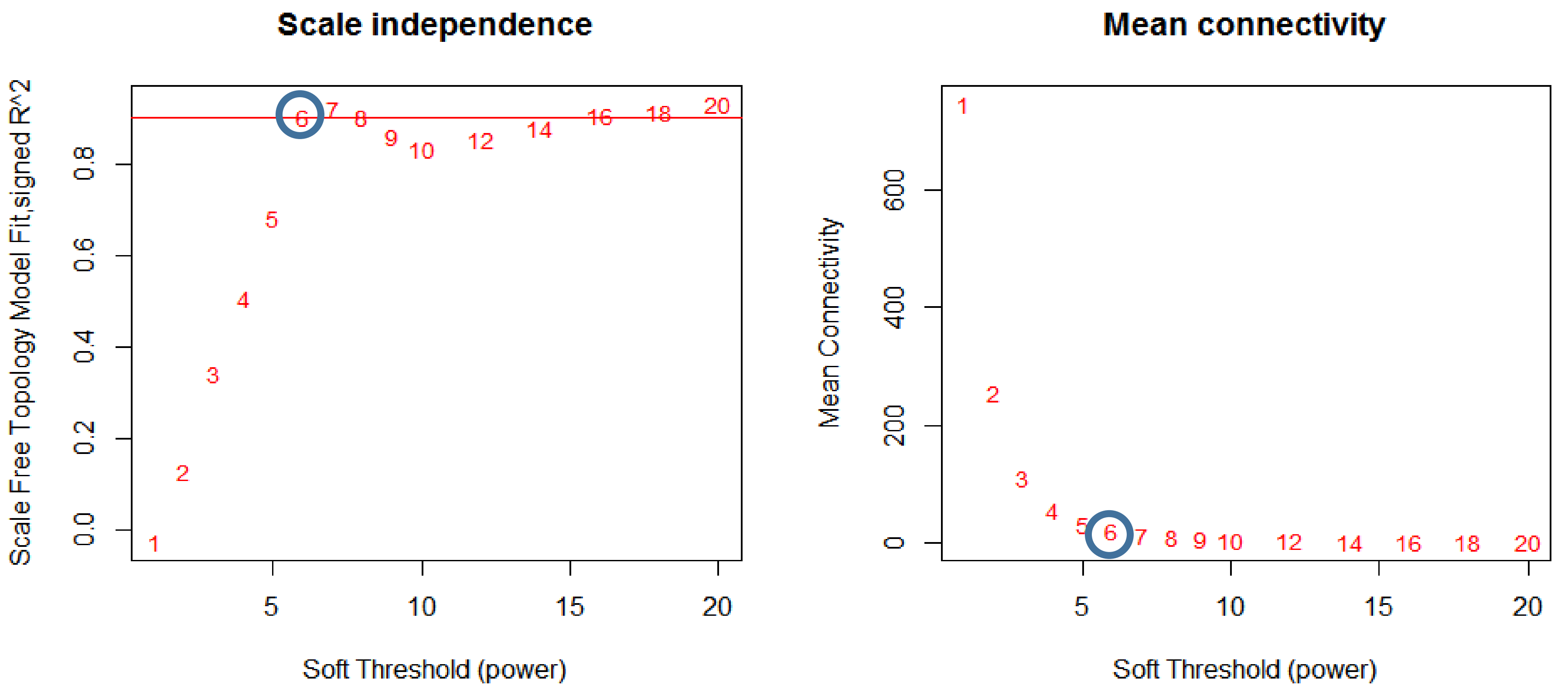
Barabási, Albert-László; Bonabeau, Eric (May 2003). "Scale-Free Networks" (PDF). Scientific American. **288** (5): 50–9.

Pick a power term: Visual Aid in WGCNA



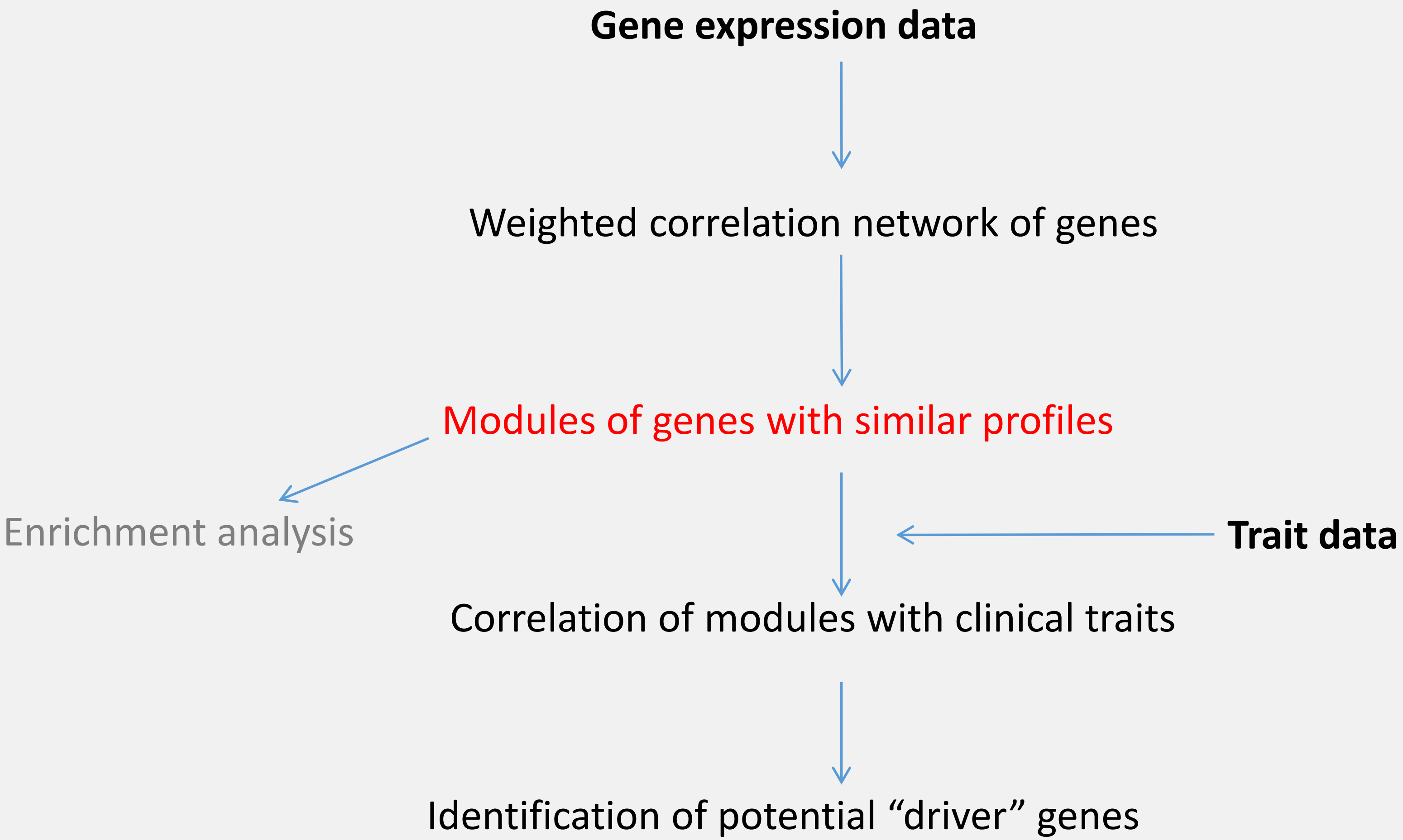
Choose power 6. Lowest possible power term where topology approximately fits a scale free network (on or above red horizontal line).

Pick a power term: Visual Aid in WGCNA



- Left plot: Choose power 6. Lowest possible power term where topology approximately fits a scale free network (on or above red horizontal line).
- Right plot: mean connectivity drops as power goes up. Must not drop too low.

Workflow



Gene expression data

Weighted correlation network of genes

Modules of genes with similar profiles

Enrichment analysis

Trait data

Correlation of modules with clinical traits

Identification of potential “driver” genes

Detect modules of co-expressed genes

4 steps to get from network to modules

1. Compute dissimilarity between genes
   (topological overlap measure dissimilarity)
2. Perform hierarchical clustering of genes
   (obtain tree structure)
3. Divide clustered genes into modules
   (cut tree branches)
4. Merge very similar modules
   (use module “eigengenes”)

Step 1: Compute dissimilarity between genes

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**
  (overlap of their network neighbors is large)
- A high TOM(i,j) implies that genes have similar expression patterns

Calculating TOM

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

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

Originally defined for unweighted networks

TOM similarity between two nodes:

1. Count number of shared neighbors
   (“agreement” of the set of neighboring nodes)

2. Normalize to [0,1]
TOM(i,j) = 0: no overlap of network neighbors
TOM(I,j) = 1: identical set of network neighbors

Generalized to the case of weighted networks in Zhang and Horvath (2005), first WGCNA paper

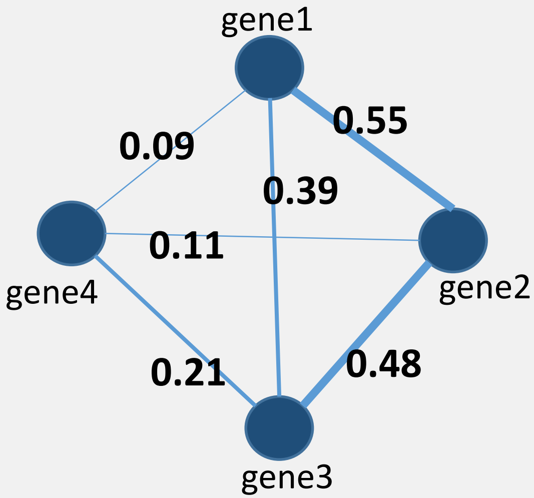
All nodes are neighbors; counting them is not informative.
Compute agreement of the set of neighboring nodes based on edge strengths.

TOM as a **similarity** measure can be transformed into a **dissimilarity** measure distTOM = 1- TOM.

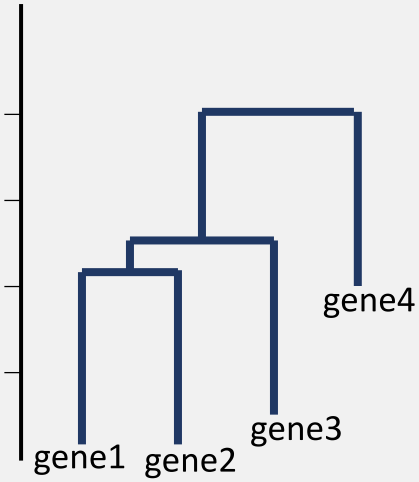
Step 2: Perform hierarchical clustering of genes

Compute gene dendrogram

Weighted correlation network
from gene expression data



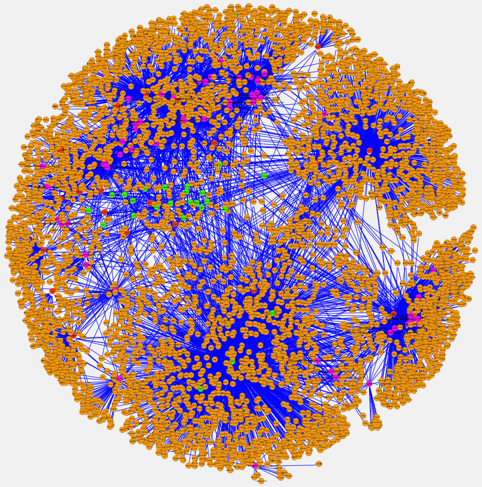
Clustering dendrogram



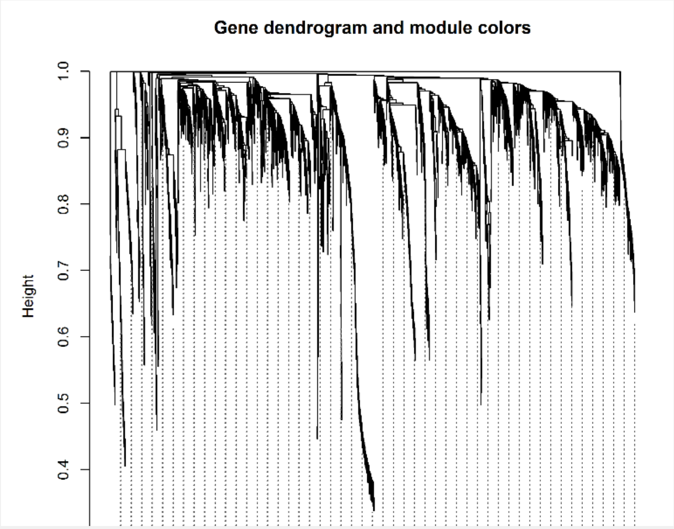
(dis)similarity between genes:
Topological Overlap Measure TOM

Compute gene dendrogram

Weighted correlation network
from gene expression data



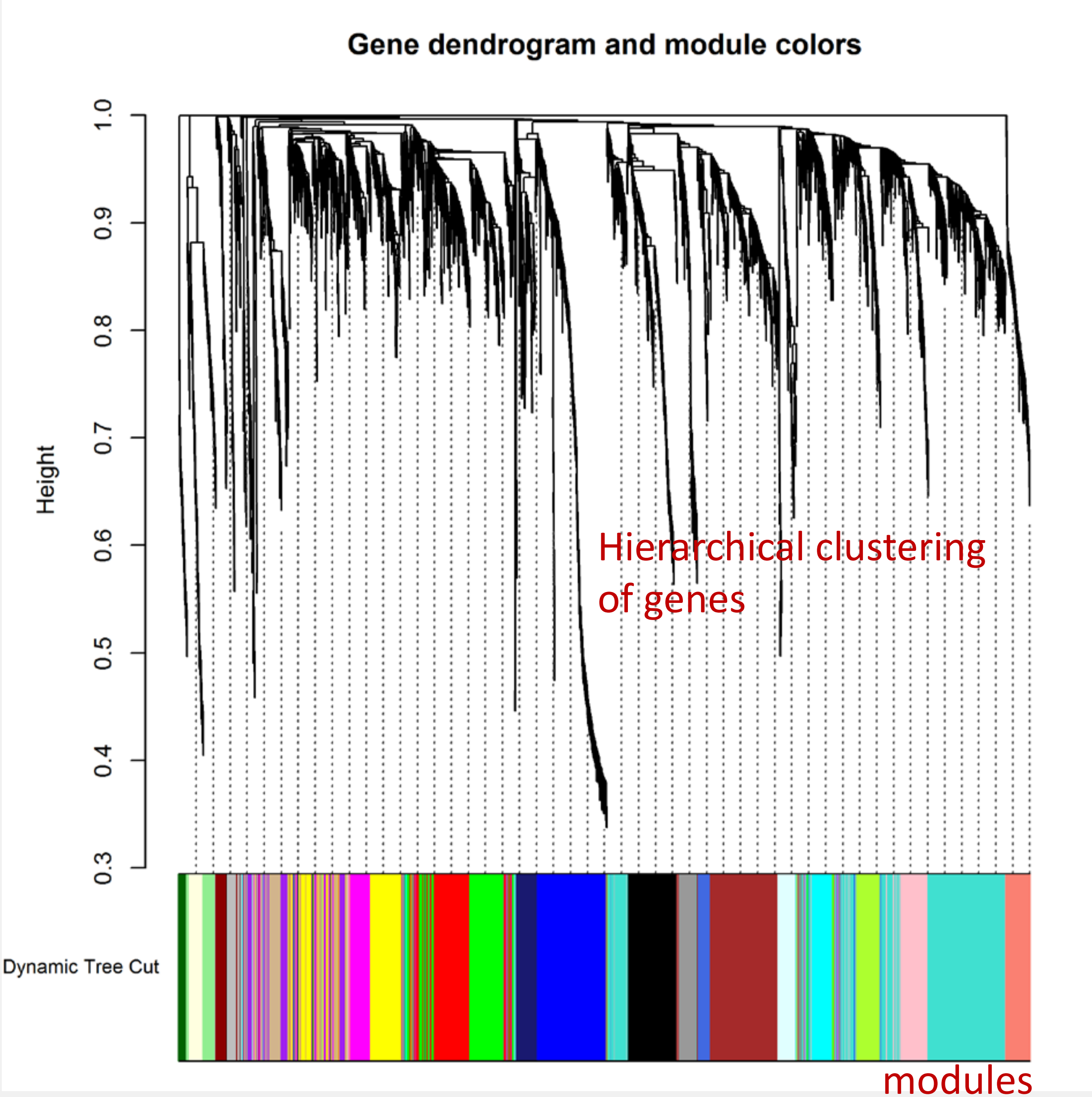
Clustering dendrogram



(dis)similarity between genes:
Topological Overlap Measure TOM

Step 3: Divide clustered genes into modules

Gene dendrogram and detected modules



Hierarchical clustering of genes

Dynamic tree cut algorithm groups genes into modules

modules

corFnc="pearson"; power=6; min. module size=30

Step 4: Merge very similar modules

Module eigengenes

A module eigengene is

a 1-dimensional data vector,
summarizing the expression data of the genes that form a module

How it is computed:

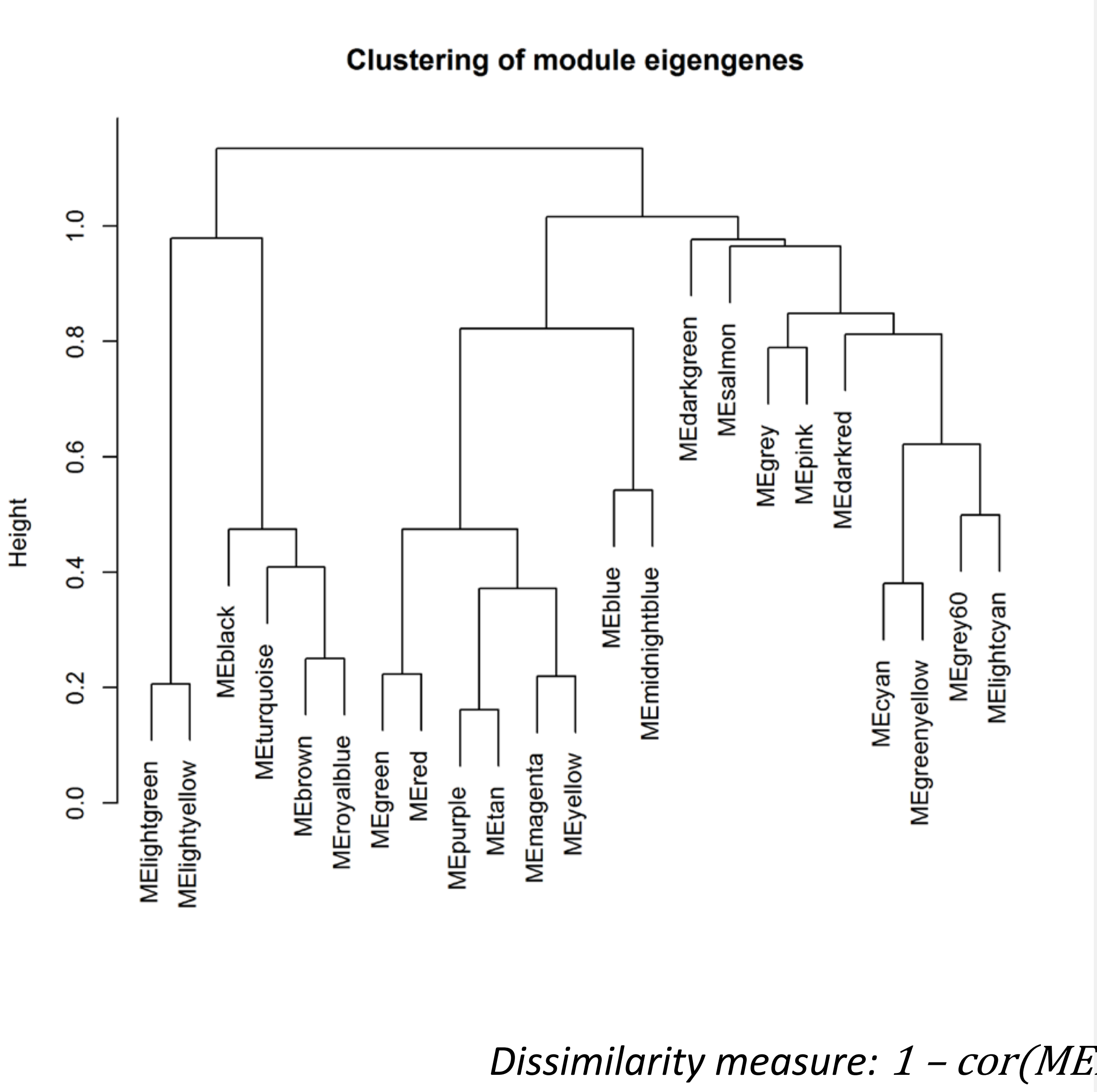
1st principal component of the expression data

What it is used for:

to represent the module in mathematical operations

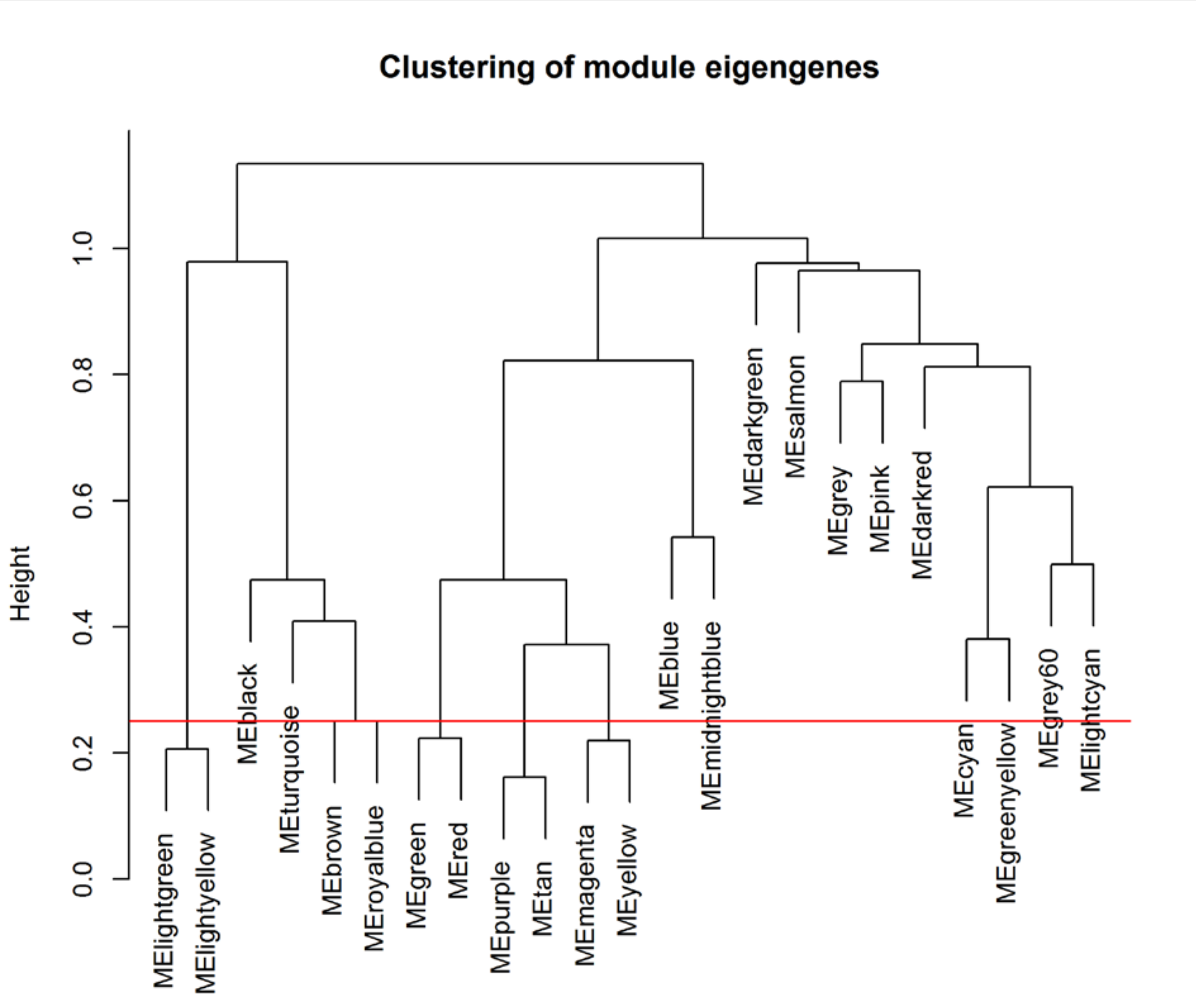
- modules can be correlated with one another
- modules can be **clustered**
- modules can be correlated with external traits

Clustering of module eigengenes



Dissimilarity measure: 1 – cor(MEigengenes)

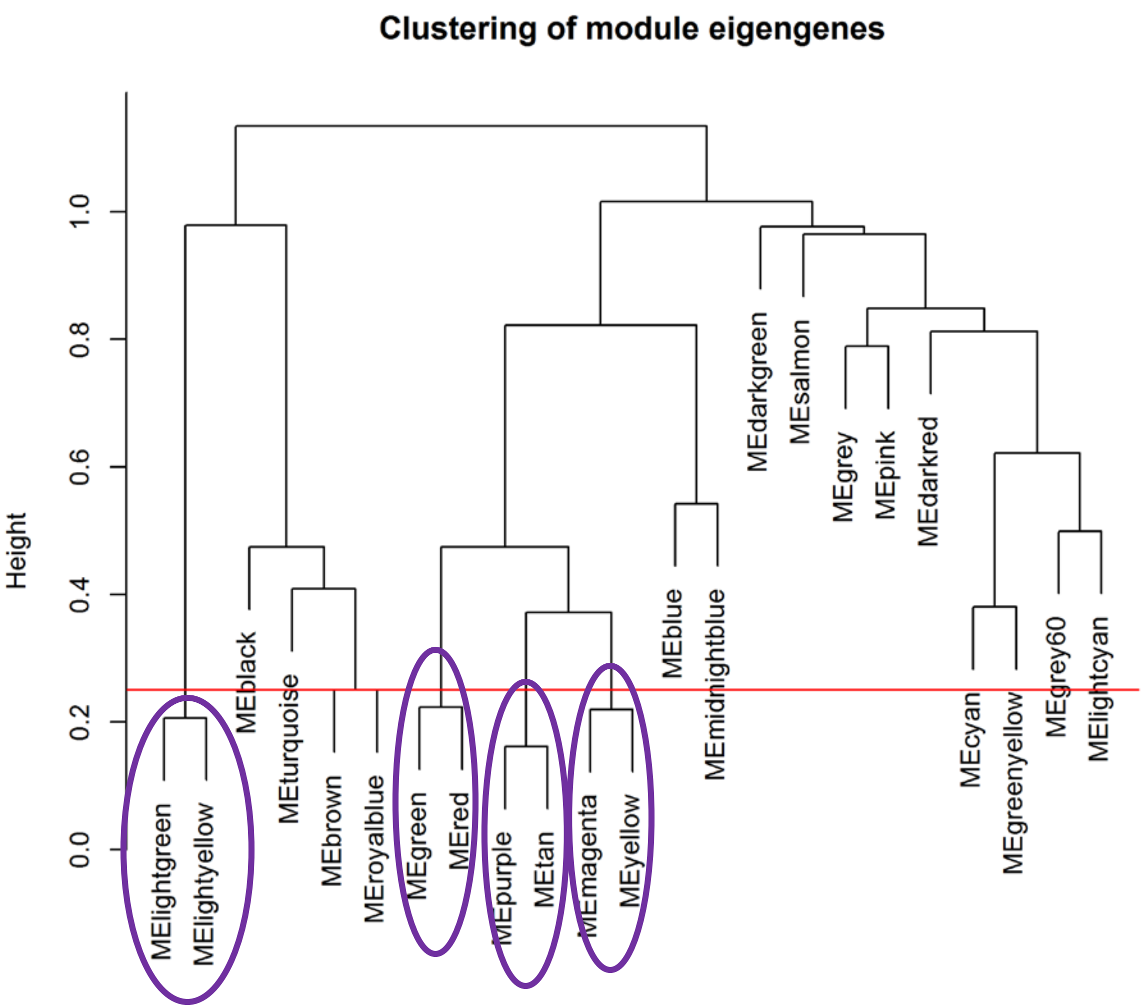
Clustering of module eigengenes



Cutoff for merging

Dissimilarity measure: 1 – cor(MEigengenes)

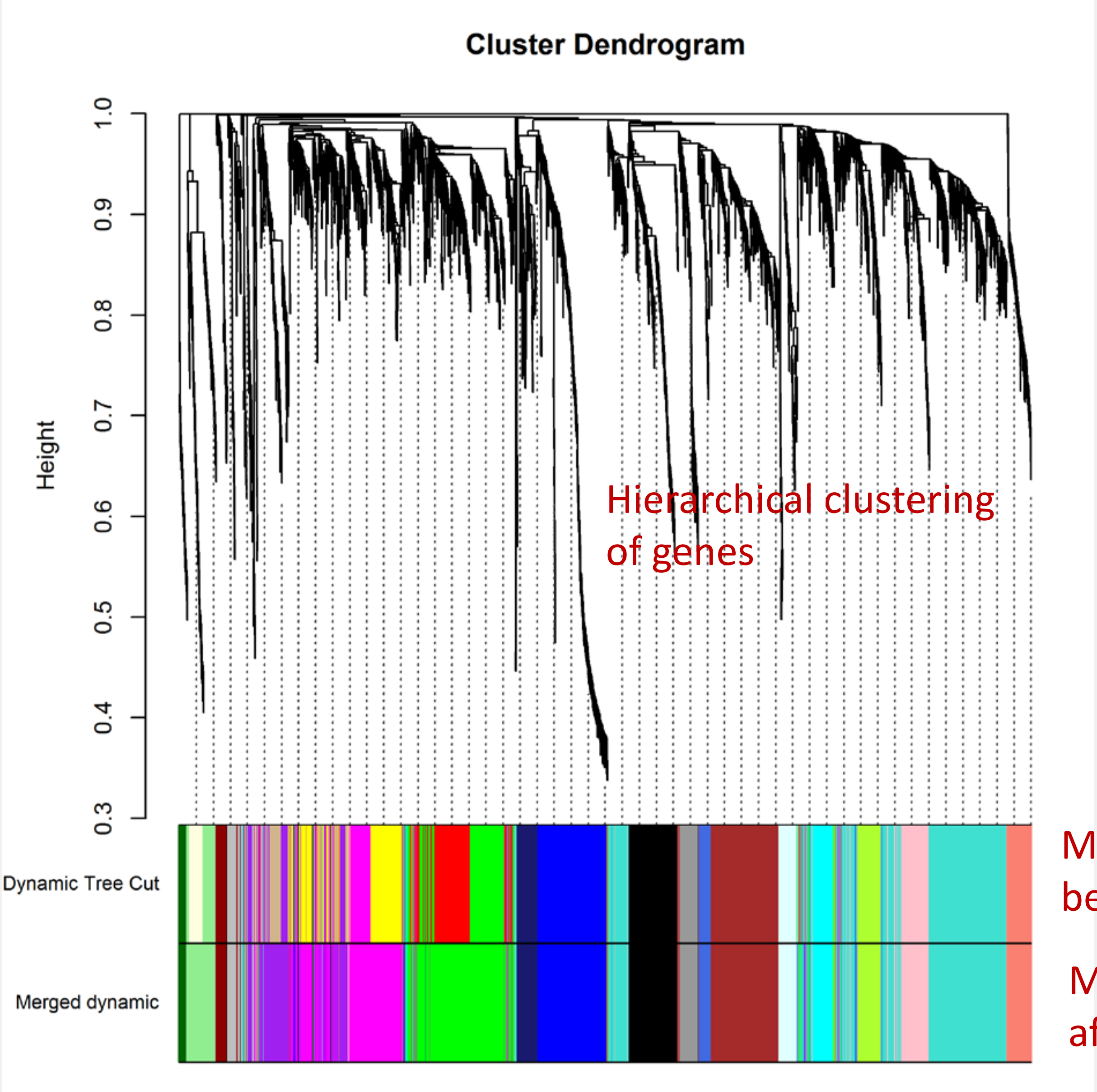
Clustering of module eigengenes



Cutoff for merging

Merge modules whose dissimilarity is below the cutoff

Gene dendrogram and detected modules, before and after merging



Hierarchical clustering of genes

Modules, before merging

Modules, after merging

corFnc="pearson"; power=6; min. module size=30

Module Detection: Decisions to make

Dynamic Tree cut

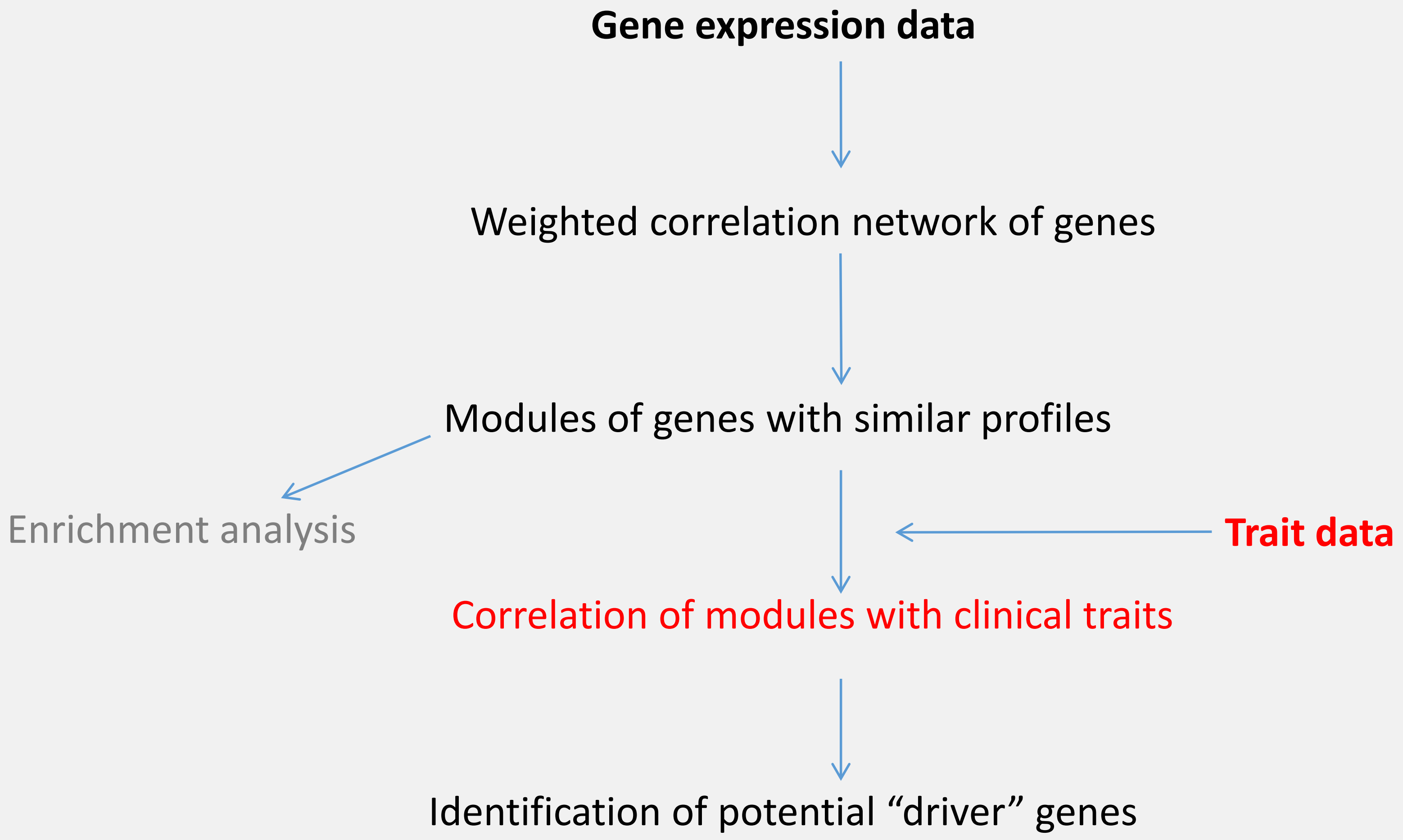
Minimal module size
typically 20 or 30

Module Merging

Cutoff for module eigengene dendrogram
typically between 0.15 and 0.25
check if clusters look ok on dendrogram

Merge once or several times?
usually once, but merge step can be repeated
- *if some modules are very similar*
- *if we want larger modules*

Workflow



Gene expression data

Weighted correlation network of genes

Modules of genes with similar profiles

Enrichment analysis

Trait data

Correlation of modules with clinical traits

Identification of potential “driver” genes

Correlate modules to external traits

Example: trait variables from Mouse Liver Data Set

weight_g
length_cm
ab_fat
other_fat
total_fat
Trigly
Total_Chol
HDL_Chol
UC
FFA
Glucose
LDL_plus_VLDL
MCP_1_phys

Insulin_ug_l
Glucose_Insulin
Leptin_pg_ml
Adiponectin
Aortic lesions
Aneurysm
Aortic_cal_M
Aortic_cal_L
CoronaryArtery_Cal
Myocardial_cal
BMD_all_limbs
BMD_femurs_only

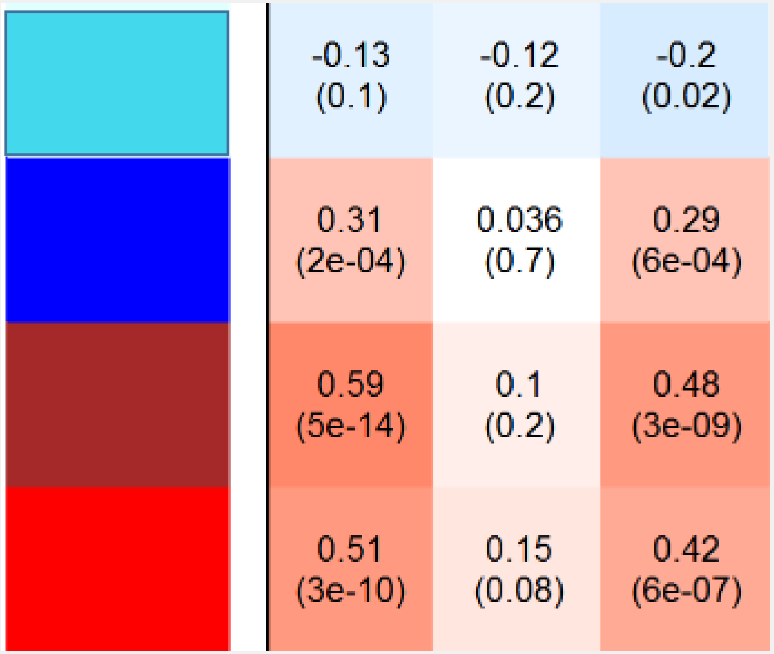
Compute correlations:
each module eigengene to each trait variable

cor(MEs, traitDat)

Table of module – trait correlations

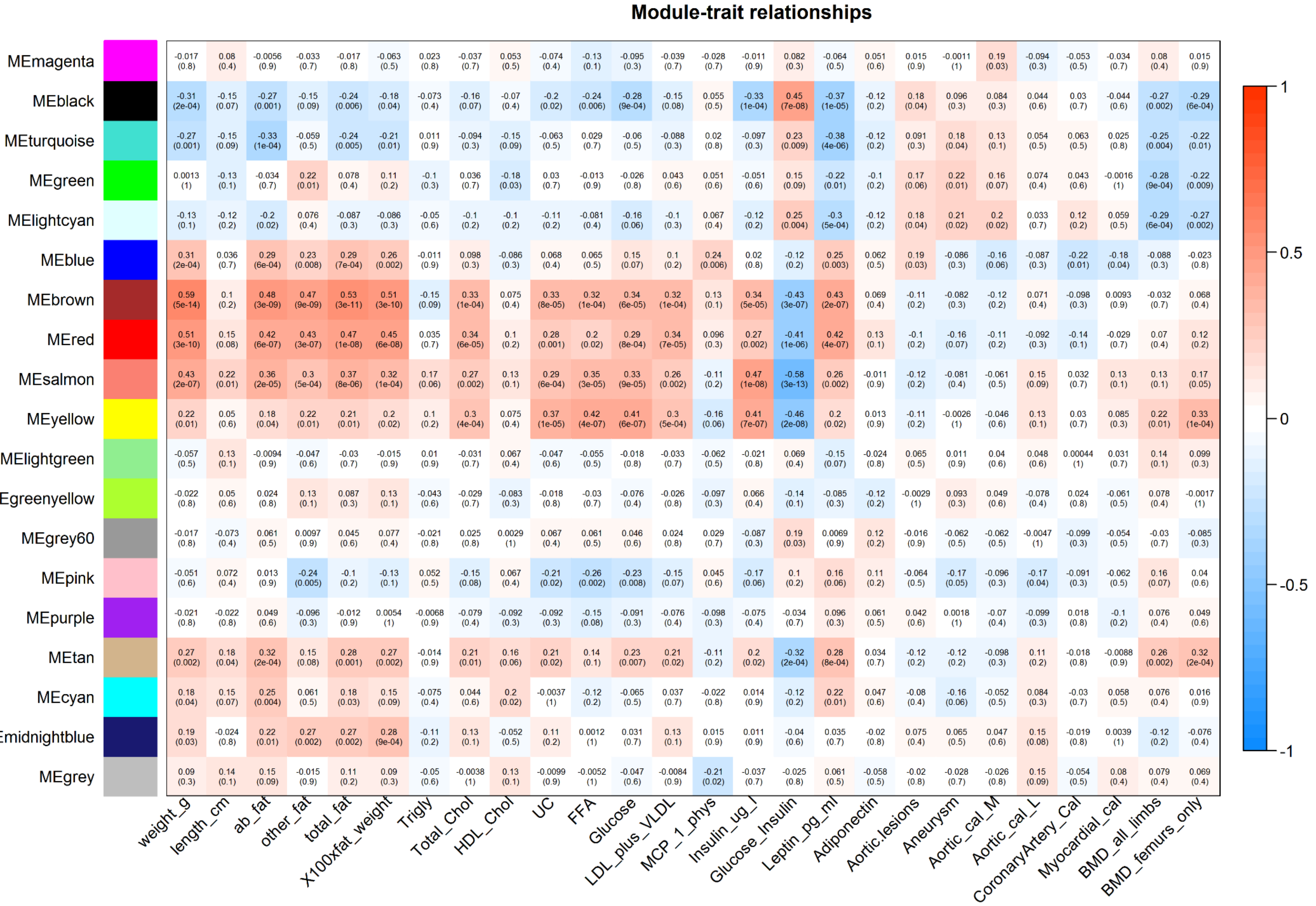
- Identify modules highly correlated to traits of interest
- Identify traits highly correlated to multiple modules

Table excerpt:



	Weight (g)	Length (cm)	Ab_fat
MEcyan	-0.13 (0.1)	-0.12 (0.2)	-0.2 (0.02)
MEblue	0.31 (2e-04)	0.036 (0.7)	0.29 (6e-04)
MEbrown	0.59 (5e-14)	0.1 (0.2)	0.48 (3e-09)
MEred	0.51 (3e-10)	0.15 (0.08)	0.42 (6e-07)

Mouse Liver Data, WGCNA



Module-trait relationships

	weight_g	length_cm	ab_fat	other_fat	total_fat	X100xfat_weight	Trigly	Total_Chol	HDL_Chol	UC	FFA	Glucose	LDL_plus_VLDL	MCP_1_phys	Insulin_ug_l	Glucose_Insulin	Leptin_pg_ml	Adiponectin	Aortic.lesions	Aneurysm	Aortic_cal_M	Aortic_cal_L	CoronaryArtery_Cal	Myocardial_cal	BMD_all_limbs	BMD_femurs_only
MEmagenta	-0.017 (0.8)	0.08 (0.4)	-0.0056 (0.9)	-0.033 (0.7)	-0.017 (0.8)	-0.063 (0.5)	0.023 (0.8)	-0.037 (0.7)	0.053 (0.5)	-0.074 (0.4)	-0.13 (0.1)	-0.095 (0.3)	-0.039 (0.7)	-0.028 (0.7)	-0.011 (0.9)	0.082 (0.3)	-0.064 (0.5)	0.051 (0.6)	0.015 (0.9)	-0.0011 (1)	0.19 (0.03)	-0.094 (0.3)	-0.053 (0.5)	-0.034 (0.7)	0.08 (0.4)	0.015 (0.9)
MEblack	-0.31 (2e-04)	-0.15 (0.07)	-0.27 (0.001)	-0.15 (0.09)	-0.24 (0.006)	-0.18 (0.04)	-0.073 (0.4)	-0.16 (0.07)	-0.07 (0.4)	-0.2 (0.02)	-0.24 (0.006)	-0.28 (9e-04)	-0.15 (0.08)	0.055 (0.5)	-0.33 (1e-04)	0.45 (7e-08)	-0.37 (1e-05)	-0.12 (0.2)	0.18 (0.04)	0.096 (0.3)	0.084 (0.3)	0.044 (0.6)	0.03 (0.7)	-0.044 (0.6)	-0.27 (0.002)	-0.29 (6e-04)
MEturquoise	-0.27 (0.001)	-0.15 (0.09)	-0.33 (1e-04)	-0.059 (0.5)	-0.24 (0.005)	-0.21 (0.01)	0.011 (0.9)	-0.094 (0.3)	-0.15 (0.09)	-0.063 (0.5)	0.029 (0.7)	-0.06 (0.5)	-0.088 (0.3)	0.02 (0.8)	-0.097 (0.3)	0.23 (0.009)	-0.38 (4e-06)	-0.12 (0.2)	0.091 (0.3)	0.18 (0.04)	0.13 (0.1)	0.054 (0.5)	0.063 (0.5)	0.025 (0.8)	-0.25 (0.004)	-0.22 (0.01)
MEgreen	0.0013 (1)	-0.13 (0.1)	-0.034 (0.7)	0.22 (0.01)	0.078 (0.4)	0.11 (0.2)	-0.1 (0.3)	0.036 (0.7)	-0.18 (0.03)	0.03 (0.7)	-0.013 (0.9)	-0.026 (0.8)	0.043 (0.6)	0.051 (0.6)	-0.051 (0.6)	0.15 (0.09)	-0.22 (0.01)	-0.1 (0.2)	0.17 (0.06)	0.22 (0.01)	0.16 (0.07)	0.074 (0.4)	0.043 (0.6)	-0.0016 (1)	-0.28 (9e-04)	-0.22 (0.009)
MElightcyan	-0.13 (0.1)	-0.12 (0.2)	-0.2 (0.02)	0.076 (0.4)	-0.087 (0.3)	-0.086 (0.3)	-0.05 (0.6)	-0.1 (0.2)	-0.1 (0.2)	-0.11 (0.2)	-0.081 (0.4)	-0.16 (0.06)	-0.1 (0.3)	0.067 (0.4)	-0.12 (0.2)	0.25 (0.004)	-0.3 (5e-04)	-0.12 (0.2)	0.18 (0.04)	0.21 (0.02)	0.2 (0.02)	0.033 (0.7)	0.12 (0.2)	0.059 (0.5)	-0.29 (6e-04)	-0.27 (0.002)
MEblue	0.31 (2e-04)	0.036 (0.7)	0.29 (6e-04)	0.23 (0.008)	0.29 (7e-04)	0.26 (0.002)	-0.011 (0.9)	0.098 (0.3)	-0.086 (0.3)	0.068 (0.4)	0.065 (0.5)	0.15 (0.07)	0.1 (0.2)	0.24 (0.006)	0.02 (0.8)	-0.12 (0.2)	0.25 (0.003)	0.062 (0.5)	0.19 (0.03)	-0.086 (0.3)	-0.16 (0.06)	-0.087 (0.3)	-0.22 (0.01)	-0.18 (0.04)	-0.088 (0.3)	-0.023 (0.8)
MEbrown	0.59 (5e-14)	0.1 (0.2)	0.48 (3e-09)	0.47 (9e-09)	0.53 (3e-11)	0.51 (3e-10)	-0.15 (0.09)	0.33 (1e-04)	0.075 (0.4)	0.33 (8e-05)	0.32 (1e-04)	0.34 (6e-05)	0.32 (1e-04)	0.13 (0.1)	0.34 (5e-05)	-0.43 (3e-07)	0.43 (2e-07)	0.069 (0.4)	-0.11 (0.2)	-0.082 (0.3)	-0.12 (0.2)	0.071 (0.4)	-0.098 (0.3)	0.0093 (0.9)	-0.032 (0.7)	0.068 (0.4)
MEred	0.51 (3e-10)	0.15 (0.08)	0.42 (6e-07)	0.43 (3e-07)	0.47 (1e-08)	0.45 (6e-08)	0.035 (0.7)	0.34 (6e-05)	0.1 (0.2)	0.28 (0.001)	0.2 (0.02)	0.29 (8e-04)	0.34 (7e-05)	0.096 (0.3)	0.27 (0.002)	-0.41 (1e-06)	0.42 (4e-07)	0.13 (0.1)	-0.1 (0.2)	-0.16 (0.07)	-0.11 (0.2)	-0.092 (0.3)	-0.14 (0.1)	-0.029 (0.7)	0.07 (0.4)	0.12 (0.2)
MEsalmon	0.43 (2e-07)	0.22 (0.01)	0.36 (2e-05)	0.3 (5e-04)	0.37 (8e-06)	0.32 (1e-04)	0.17 (0.06)	0.27 (0.002)	0.13 (0.1)	0.29 (6e-04)	0.35 (3e-05)	0.33 (9e-05)	0.26 (0.002)	-0.11 (0.2)	0.47 (1e-08)	-0.58 (3e-13)	0.26 (0.002)	-0.011 (0.9)	-0.12 (0.2)	-0.081 (0.4)	-0.061 (0.5)	0.15 (0.09)	0.032 (0.7)	0.13 (0.1)	0.13 (0.1)	0.17 (0.05)
MEyellow	0.22 (0.01)	0.05 (0.6)	0.18 (0.04)	0.22 (0.01)	0.21 (0.01)	0.2 (0.02)	0.1 (0.2)	0.3 (4e-04)	0.075 (0.4)	0.37 (1e-05)	0.42 (4e-07)	0.41 (6e-07)	0.3 (5e-04)	-0.16 (0.06)	0.41 (7e-07)	-0.46 (2e-08)	0.2 (0.02)	0.013 (0.9)	-0.11 (0.2)	-0.0026 (1)	-0.046 (0.6)	0.13 (0.1)	0.03 (0.7)	0.085 (0.3)	0.22 (0.01)	0.33 (1e-04)
MElightgreen	-0.057 (0.5)	0.13 (0.1)	-0.0094 (0.9)	-0.047 (0.6)	-0.03 (0.7)	-0.015 (0.9)	0.01 (0.9)	-0.031 (0.7)	0.067 (0.4)	-0.047 (0.6)	-0.055 (0.5)	-0.018 (0.8)	-0.033 (0.7)	-0.062 (0.5)	-0.021 (0.8)	0.069 (0.4)	-0.15 (0.07)	-0.024 (0.8)	0.065 (0.5)	0.011 (0.9)	0.04 (0.6)	0.048 (0.6)	0.00044 (1)	0.031 (0.7)	0.14 (0.1)	0.099 (0.3)
Egreenyellow	-0.022 (0.8)	0.05 (0.6)	0.024 (0.8)	0.13 (0.1)	0.087 (0.3)	0.13 (0.1)	-0.043 (0.6)	-0.029 (0.7)	-0.083 (0.3)	-0.018 (0.8)	-0.03 (0.7)	-0.076 (0.4)	-0.026 (0.8)	-0.097 (0.3)	0.066 (0.4)	-0.14 (0.1)	-0.085 (0.3)	-0.12 (0.2)	-0.0029 (1)	0.093 (0.3)	0.049 (0.6)	-0.078 (0.4)	0.024 (0.8)	-0.061 (0.5)	0.078 (0.4)	-0.0017 (1)
MEgrey60	-0.017 (0.8)	-0.073 (0.4)	0.061 (0.5)	0.0097 (0.9)	0.045 (0.6)	0.077 (0.4)	-0.021 (0.8)	0.025 (0.8)	0.0029 (1)	0.067 (0.4)	0.061 (0.5)	0.046 (0.6)	0.024 (0.8)	0.029 (0.7)	-0.087 (0.3)	0.19 (0.03)	0.0069 (0.9)	0.12 (0.2)	-0.016 (0.9)	-0.062 (0.5)	-0.062 (0.5)	-0.0047 (1)	-0.099 (0.3)	-0.054 (0.5)	-0.03 (0.7)	-0.085 (0.3)
MEpink	-0.051 (0.6)	0.072 (0.4)	0.013 (0.9)	-0.24 (0.005)	-0.1 (0.2)	-0.13 (0.1)	0.052 (0.5)	-0.15 (0.08)	0.067 (0.4)	-0.21 (0.02)	-0.26 (0.002)	-0.23 (0.008)	-0.15 (0.07)	0.045 (0.6)	-0.17 (0.06)	0.1 (0.2)	0.16 (0.06)	0.11 (0.2)	-0.064 (0.5)	-0.17 (0.05)	-0.096 (0.3)	-0.17 (0.04)	-0.091 (0.3)	-0.062 (0.5)	0.16 (0.07)	0.04 (0.6)
MEpurple	-0.021 (0.8)	-0.022 (0.8)	0.049 (0.6)	-0.096 (0.3)	-0.012 (0.9)	0.0054 (1)	-0.0068 (0.9)	-0.079 (0.4)	-0.092 (0.3)	-0.092 (0.3)	-0.15 (0.08)	-0.091 (0.3)	-0.076 (0.4)	-0.098 (0.3)	-0.075 (0.4)	-0.034 (0.7)	0.096 (0.3)	0.061 (0.5)	0.042 (0.6)	0.0018 (1)	-0.07 (0.4)	-0.099 (0.3)	0.018 (0.8)	-0.1 (0.2)	0.076 (0.4)	0.049 (0.6)
MEtan	0.27 (0.002)	0.18 (0.04)	0.32 (2e-04)	0.15 (0.08)	0.28 (0.001)	0.27 (0.002)	-0.014 (0.9)	0.21 (0.01)	0.16 (0.06)	0.21 (0.02)	0.14 (0.1)	0.23 (0.007)	0.21 (0.02)	-0.11 (0.2)	0.2 (0.02)	-0.32 (2e-04)	0.28 (8e-04)	0.034 (0.7)	-0.12 (0.2)	-0.12 (0.2)	-0.098 (0.3)	0.11 (0.2)	-0.018 (0.8)	-0.0088 (0.9)	0.26 (0.002)	0.32 (2e-04)
MEcyan	0.18 (0.04)	0.15 (0.07)	0.25 (0.004)	0.061 (0.5)	0.18 (0.03)	0.15 (0.09)	-0.075 (0.4)	0.044 (0.6)	0.2 (0.02)	-0.0037 (1)	-0.12 (0.2)	-0.065 (0.5)	0.037 (0.7)	-0.022 (0.8)	0.014 (0.9)	-0.12 (0.2)	0.22 (0.01)	0.047 (0.6)	-0.08 (0.4)	-0.16 (0.06)	-0.052 (0.5)	0.084 (0.3)	-0.03 (0.7)	0.058 (0.5)	0.076 (0.4)	0.016 (0.9)
Emidnightblue	0.19 (0.03)	-0.024 (0.8)	0.22 (0.01)	0.27 (0.002)	0.27 (0.002)	0.28 (9e-04)	-0.11 (0.2)	0.13 (0.1)	-0.052 (0.5)	0.11 (0.2)	0.0012 (1)	0.031 (0.7)	0.13 (0.1)	0.015 (0.9)	0.011 (0.9)	-0.04 (0.6)	0.035 (0.7)	-0.02 (0.8)	0.075 (0.4)	0.065 (0.5)	0.047 (0.6)	0.15 (0.08)	-0.019 (0.8)	0.0039 (1)	-0.12 (0.2)	-0.076 (0.4)
MEgrey	0.09 (0.3)	0.14 (0.1)	0.15 (0.09)	-0.015 (0.9)	0.11 (0.2)	0.09 (0.3)	-0.05 (0.6)	-0.0038 (1)	0.13 (0.1)	-0.0099 (0.9)	-0.0052 (1)	-0.047 (0.6)	-0.0084 (0.9)	-0.21 (0.02)	-0.037 (0.7)	-0.025 (0.8)	0.061 (0.5)	-0.058 (0.5)	-0.02 (0.8)	-0.028 (0.7)	-0.026 (0.8)	0.15 (0.09)	-0.054 (0.5)	0.08 (0.4)	0.079 (0.4)	0.069 (0.4)

What to do if there are binary traits?

Examples of binary traits:

Sex (male/female)

Disease status (healthy/diseased)

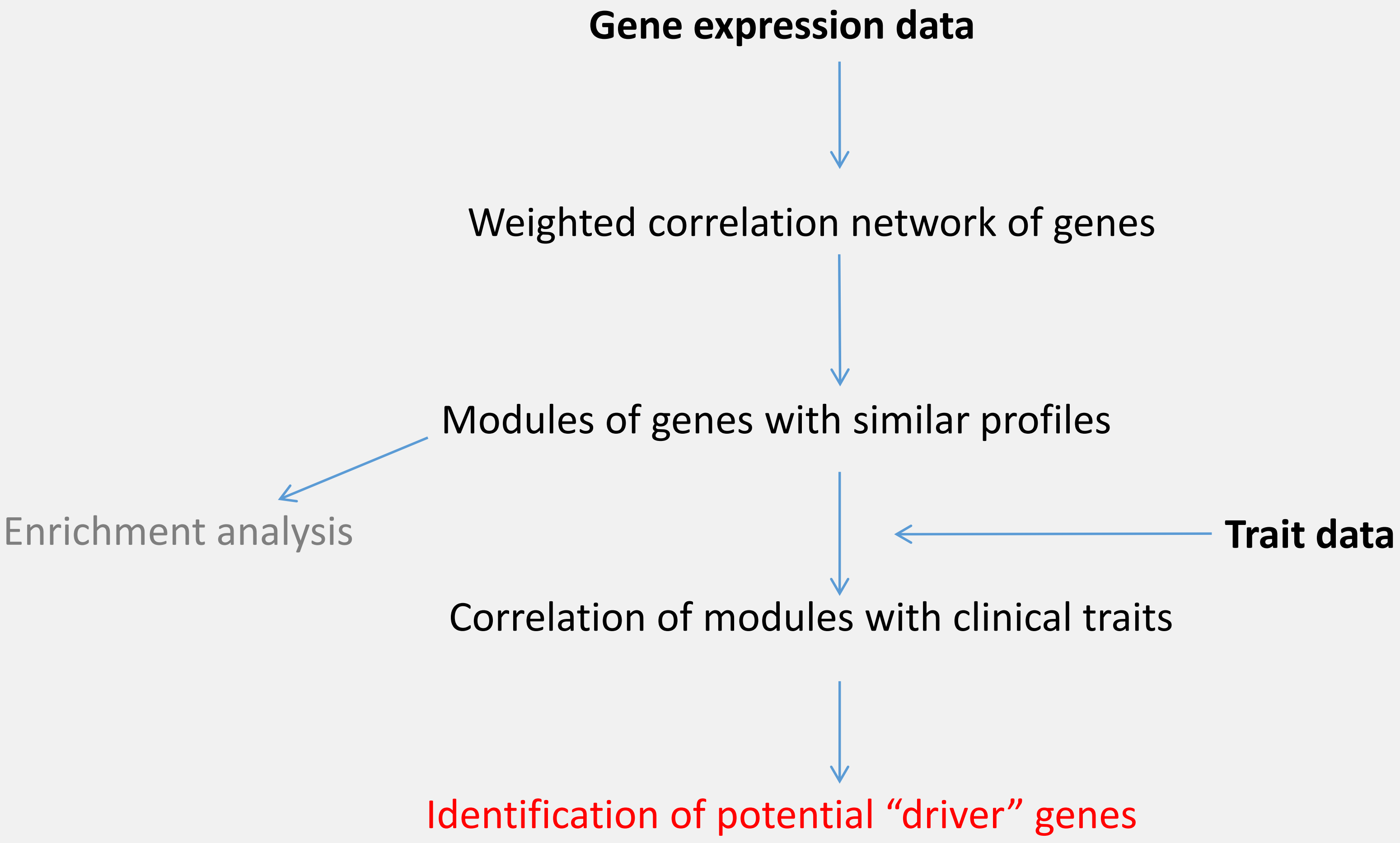
Pearson and **Spearman** correlation (function `cor`): will work

Biweight midcorrelation (function `bicor`) needs modification

- restrict number of values that will be treated as outliers
- turn off robust treatment for the binary variable

```
bicor(moduleEigengenes, datTraits, use=“p",
robustY=FALSE, maxPOutliers=0.1)
```

Workflow



Gene expression data

Weighted correlation network of genes

Modules of genes with similar profiles

Enrichment analysis

Trait data

Correlation of modules with clinical traits

Identification of potential “driver” genes

Identify potential “driver” genes

Potential driver genes

Aim: Identify key genes that

- May **influence** the expression or function of **other genes**
- May be **causal** factors for a **trait** of interest

Caveat:

- WGCNA **cannot** show whether gene-gene or gene-trait relationships are **causal**.
- WGCNA **can** help find **candidate** genes for further study.

Potential driver genes

Strategy:

Identify those genes within a module that are

1) Highly connected within the module (hub genes)

   AND

2) Most strongly correlated with a clinical/phenotypical trait of interest

How to detect hub genes inside a module?

The straightforward way:

gene(s) with highest intramodular connectivity

(=sum of in-module edge weights)

Alternative way proposed in WGCNA:

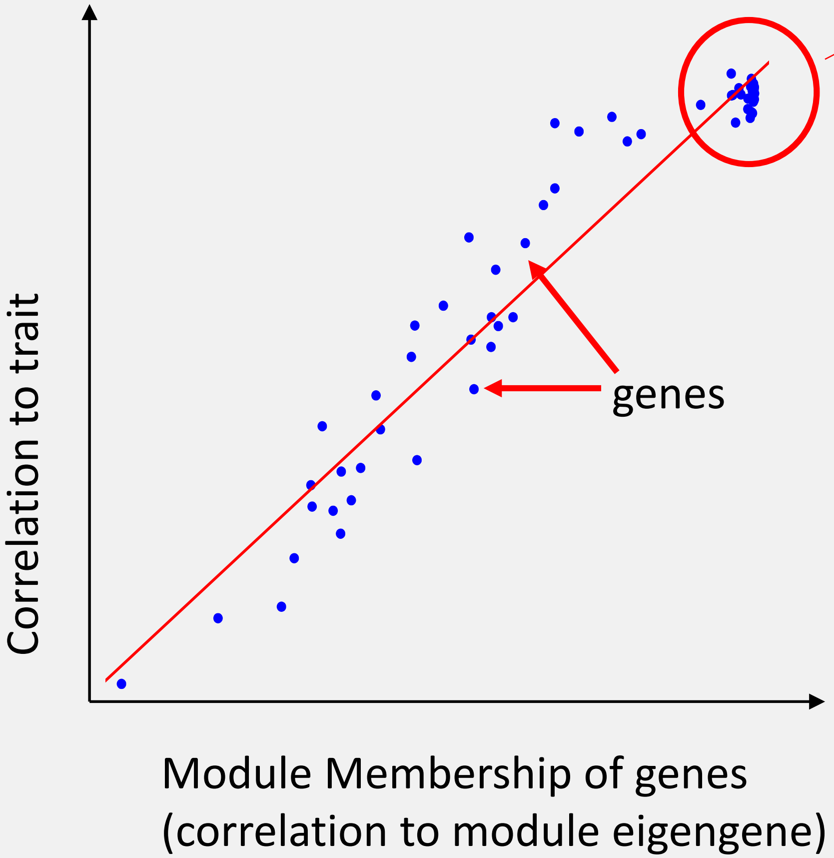
gene(s) with highest module membership

Module Membership of Genes

Module membership: **Correlation** of a **gene** to a **module eigengene**

- Genes with high module membership are good representatives of the overall expression profile in the module
- Genes with high module membership tend to be "hub" genes in the module (high intramodule connectivity)
- A gene can have high membership in several modules (not just the one to which it is assigned)

Potential driver genes



“driver” genes:
most correlated with
both module and trait

(picked by visual
inspection)

Evaluate module quality

Check module sizes

Potential issues:

Very large modules

- may make biological sense, but difficult to handle

Many similar modules

- level of merging not sufficient

Quality checks on modules

Are modules better than random groupings of genes?

Connectivity

mean intra-module connectivity

mean ratio of intra-module / total connectivity

Trait correlations

strong correlation between module eigengenes and traits of interest

strong correlation between gene module membership and gene-trait corr.

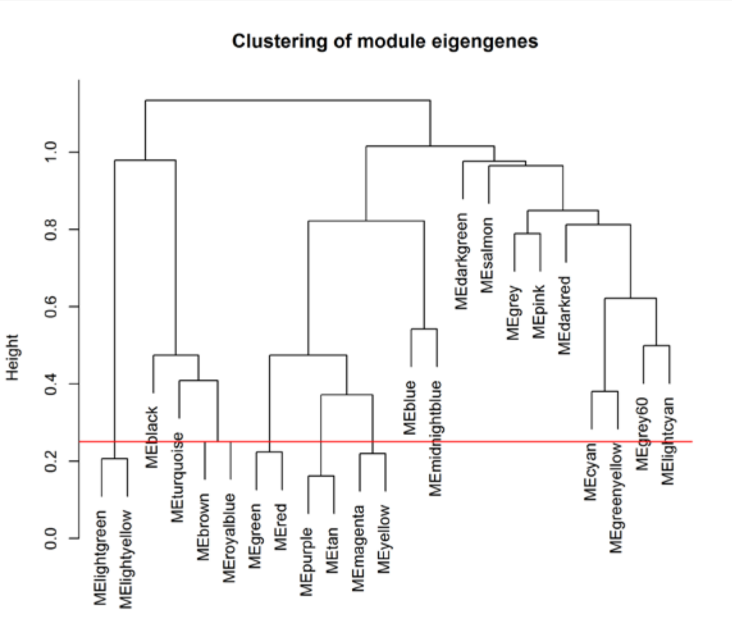
Functional enrichment

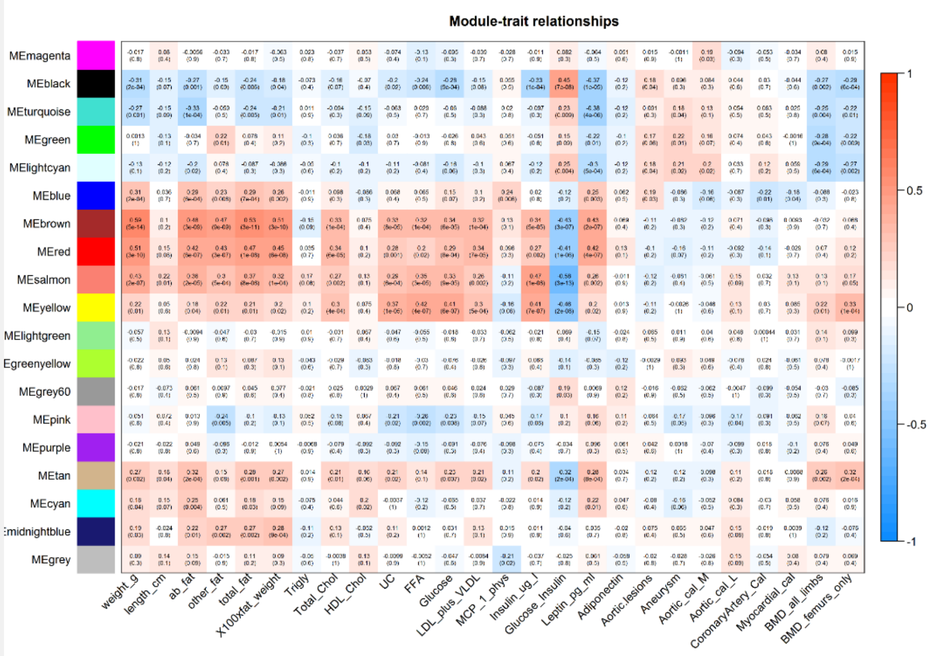
many functionally related genes in the same module →

Next up:
Presentation on
GO enrichment
analysis

Most important initial checks:

Look at plots!





Quality checks on modules can be simple or complex

Assessing modules by intramodular connectivity

- Simple ranking of modules
  (highest to lowest mean connectivity)

- Statistical analysis
  obtain p-values, e.g. via bootstrap
  (is module internally more connected than would be expected by chance?)

Quality checks on modules can be simple or complex

Assessing modules by functional enrichment of GO terms

multiple methods exists

with or without taking GO hierarchy into account (next presentation)

Two purposes of module evaluation:

1) Identify modules of interest for further analysis

2) Assess overall quality of modules

if not satisfactory, re-do network construction and/or module detection with changed parameters

Resources

WGCNA theory papers and applied papers, including PowerPoint presentations

https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/

Examples of topics

- Constructing a gene coexpression network (2005)
- Understanding the Topological Overlap Measure (2007)
- Dynamic Tree Cut (2007)
- Eigenegene networks (2007)
- *When Is Hub Gene Selection Better than Standard Meta-Analysis (2013)*

WGCNA R package Website, including many tutorials

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