Fakulteti i Shkencave Natyrore
www.fshn.edu.al

MICROBIOME

Identifying most essential genes related with obesity using complex network analysis

Eglantina Kalluçi , Xhilda Merkaj

Faculty of Natural Sciences

University of Tirana

Introduction to network medicine

❑A network can be described as a graph that shows the interconnections between a set of actors. Each actor is represented by a node and each connection between these nodes is represented by an edge. Network Analysis is a mathematical approach to study the relationships among nodes.

❑ Network medicine is an emerging area of research dealing with molecular and genetic interactions, network biomarkers of diseases and therapeutic target discovery.

❑Network- based approaches [Barabasi A., Gulbahce N., Loscalzo J., 2010] have potential biological and clinical applications, from the identification of disease genes to better drug targets.

Why networks?

- Networks provide natural description of relation between various components.
- Network Analysis introduces a new way to study the medication use in a population.
- Example of medicine networks:
  - Protein- protein network
  - Protein domain co- occurrence network
  - Metabolic networks
  - Transcription networks
  - Co- expression networks

Co- expression networks

- Gene co-expression networks can be used to associate genes of unknown function with biological processes, to prioritize candidate disease genes or to discern transcriptional regulatory programmes.

- The clinical and genomics data of this study were originally obtained from 99 obesity participants who were recruited under a natural history protocol. Blood samples were collected from fasting participants.

- In this study we have information for 13276 different genes related to 99 patients.

Co- expression network

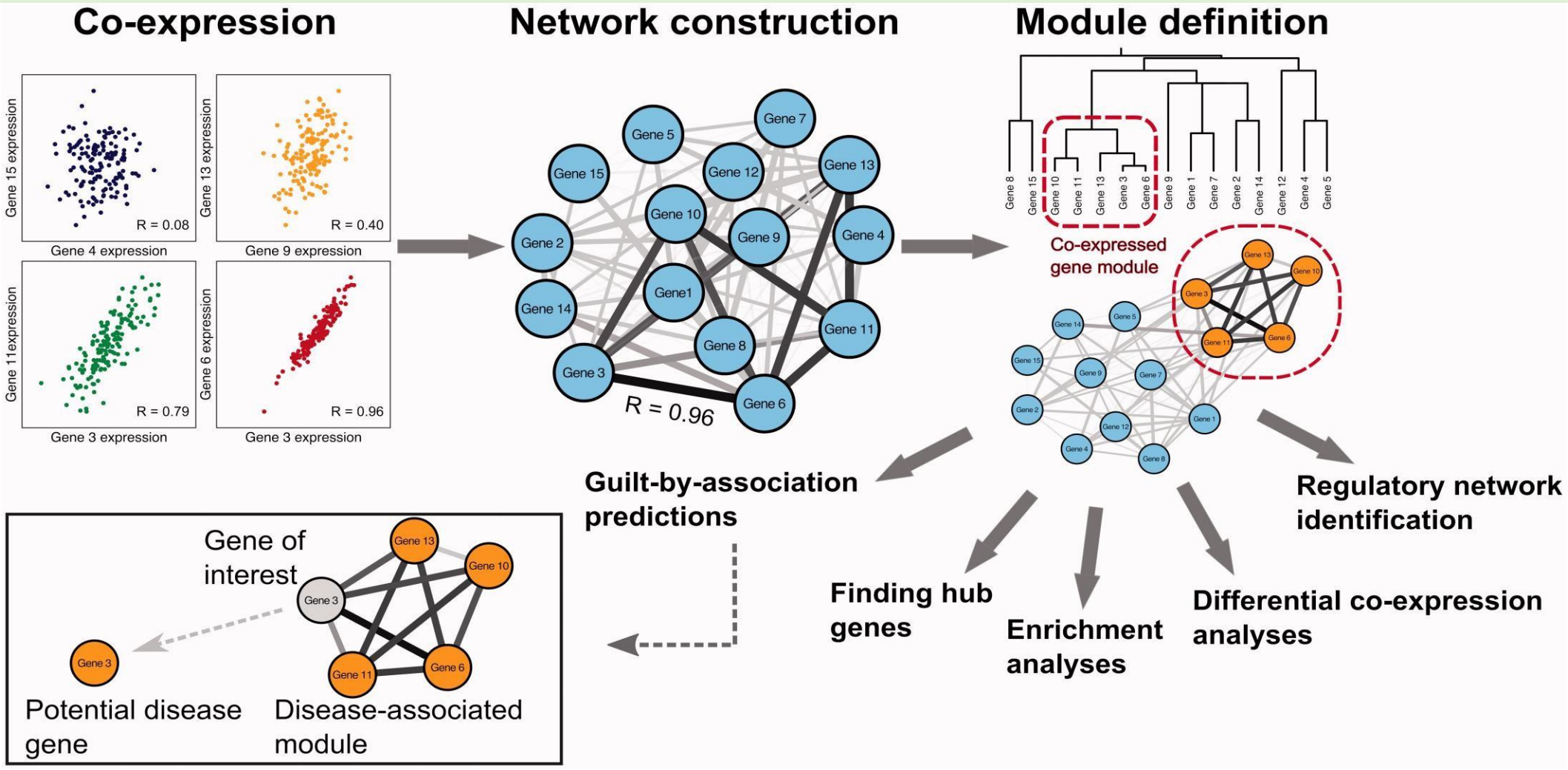


Figure 1. Construction and analysis of co- expression network

Network adjacency matrices and connectivity

- For an *unweighted* network, the adjacency matrix is given by

$$a_{ij} = \begin{cases} 1 & if\ exit\ edge\ (i,j) \\ 0 & ot\square erwise \end{cases} \quad (1)$$

- For an *weighted unsigned* network $a_{ij} = \{^{0 \le |cor(i,j)| \le 1}$ (2)

- Individual relationships between genes are defined based on correlation measures or mutual information between each pair of genes. These relationships describe the similarity between expression patterns of the gene pair across all the samples.

- An adjacency matrix is constructed where each node represents a gene and each edge represents the presence and the strength of the co-expression relationship [Wigger L. et al. 2016].

- Scaled connectivity (scaled degree) is defined as $C_i(A) = \sum_{j \neq i} \frac{a_{ij}}{n-1}, \quad 0 \le C_i(A) \le 1$ (3)

Transformations of the adjacency matrix

- The power transformation raises each adjacency to a fixed power: $Power\ (A, \beta\ ) = a_{\substack{i \\ j}}^{\beta}$
- The power transformation can be used to emphasize large adjacencies at the expense of low ones.
- The topological overlap transformation *TOM*(*A*) replaces each adjacency a_{ij} by a normalized count neighbours that are shared by the nodes *i,j*.
- For a weighted network *A*, the topological overlap measure (TOM) is defined as:

$$TOM_{ij}(A) = \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{min\left(\sum_{\substack{u \neq i \\ j}} a_{iu}, \sum_{u \neq} a_{ju}\right) + 1 - a_{ij}} \qquad (4)$$

- The topological overlap of two genes reflects their similarity in terms of the commonality of the genes they connect to.

Weighted correlation- Descisions to make

❑ ***Selecting a network type***

Unsigned network- No differentiation between positive and negative correlations.

❑ ***Choosing a correlation method***

Different measures of correlation have been used to construct networks, including Pearson's or Spearman's correlations. Alternatively, least absolute error regression or a Bayesian approach can be used to construct a co-expression network. [Guttman M. et al. 2011.]

❑ ***Picking a power term***

Selection criterion: Pick lowest possible β that leads to an approximately scale-free network topology. [Wigger L. et al. 2016]

Few nodes with many connections.

Many nodes with few connenctions.

Degree distribution of nodes follows a power law: $Y = kX^{\alpha}$ (5)

Pick a power term for gene co-expression network

- [Langfelder P., Horvath S. at el. 2014]

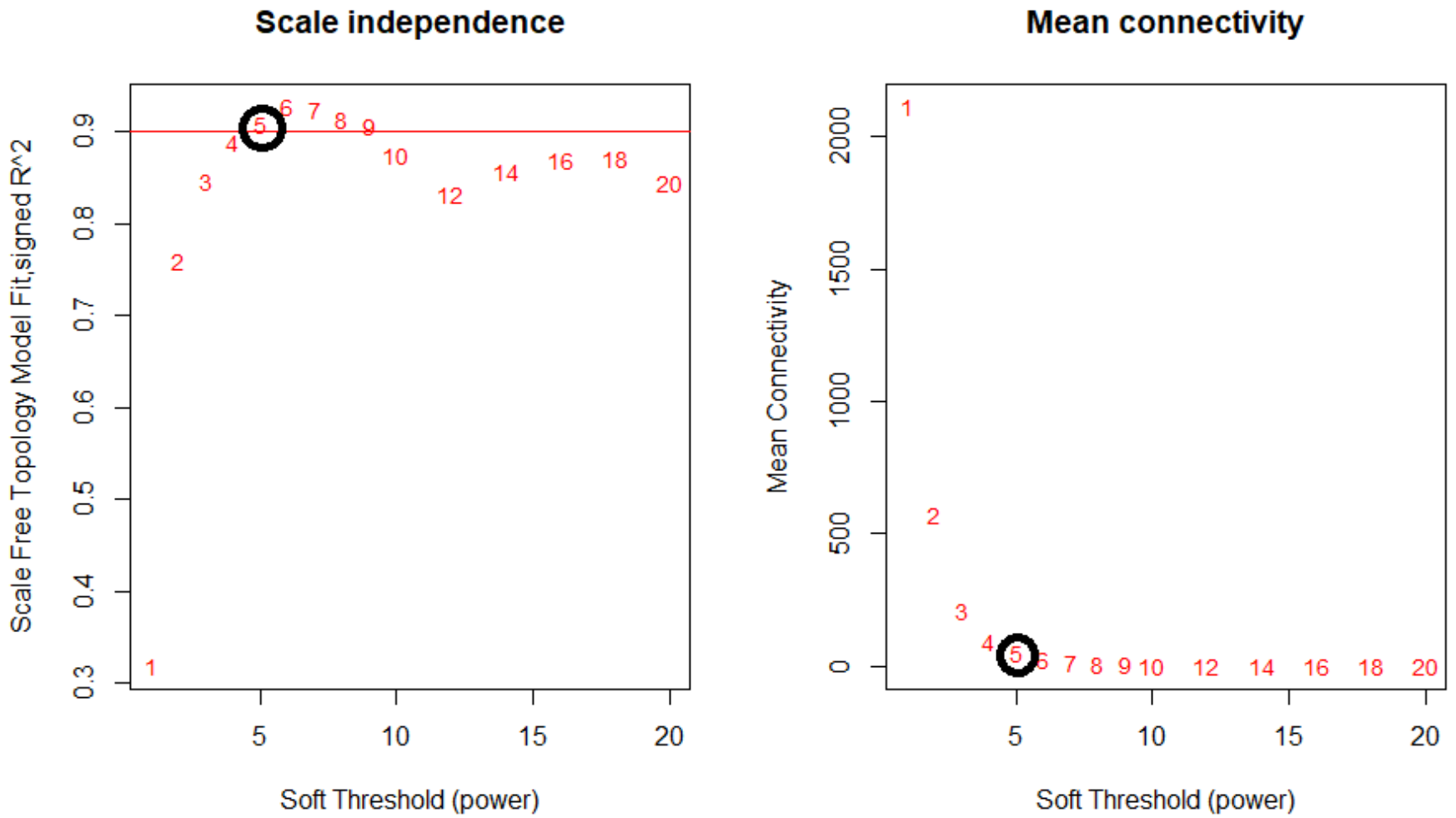


Figure 2. *Pick a power term for gene co-expression network contraction*

Dissimilarity transformation for module detection

- $Power\ (A, \beta)$ and $TOM_{ij}(A)$ satisfy the conditions of an adjacency matrix.
- The dissimilarity transformation *Dissim*(*A*) turns an adjacency matrix (which is a measure of similarity) into a measure of *dissimilarity* by subtracting it from 1.

$$Dissim_{ij}(A) = 1 - TOM(A) \qquad (5)$$

- This transformation is useful for defining module detection procedures.
- *Dissim*(*A*) does not satisfy our definition of an adjacency matrix since its diagonal elements are equal to 0.

Module detection using hierarchical clustering

- We define modules as clusters that result from using a pairwise node dissimilarity d_{ij}.

- For a gene network with adjacency matrix A, we use the topological overlap based dissimilarity:

$$d_{\substack{i\,j\\ j}} = Dissim_i\left(TOM(A)\right) = 1 - \frac{\sum_{u\neq i,j} a_{iu}a_{uj} + a_{ij}}{min\left(\sum_{\substack{u\neq i\\ j}} a_{iu}, \sum_{u\neq} a_{ju}\right) + 1 - a_{ij}} \qquad (6)$$

- This dissimilarity is used as input to average linkage hierarchical clustering.

- We use two different branch cutting techniques: the constant-height cut method and the dynamic tree cut method.

- This module detection approach has been successfully used in several studies [Langfelder P., Horvath S. at el. 2014]

Module detection using hierarchical clustering

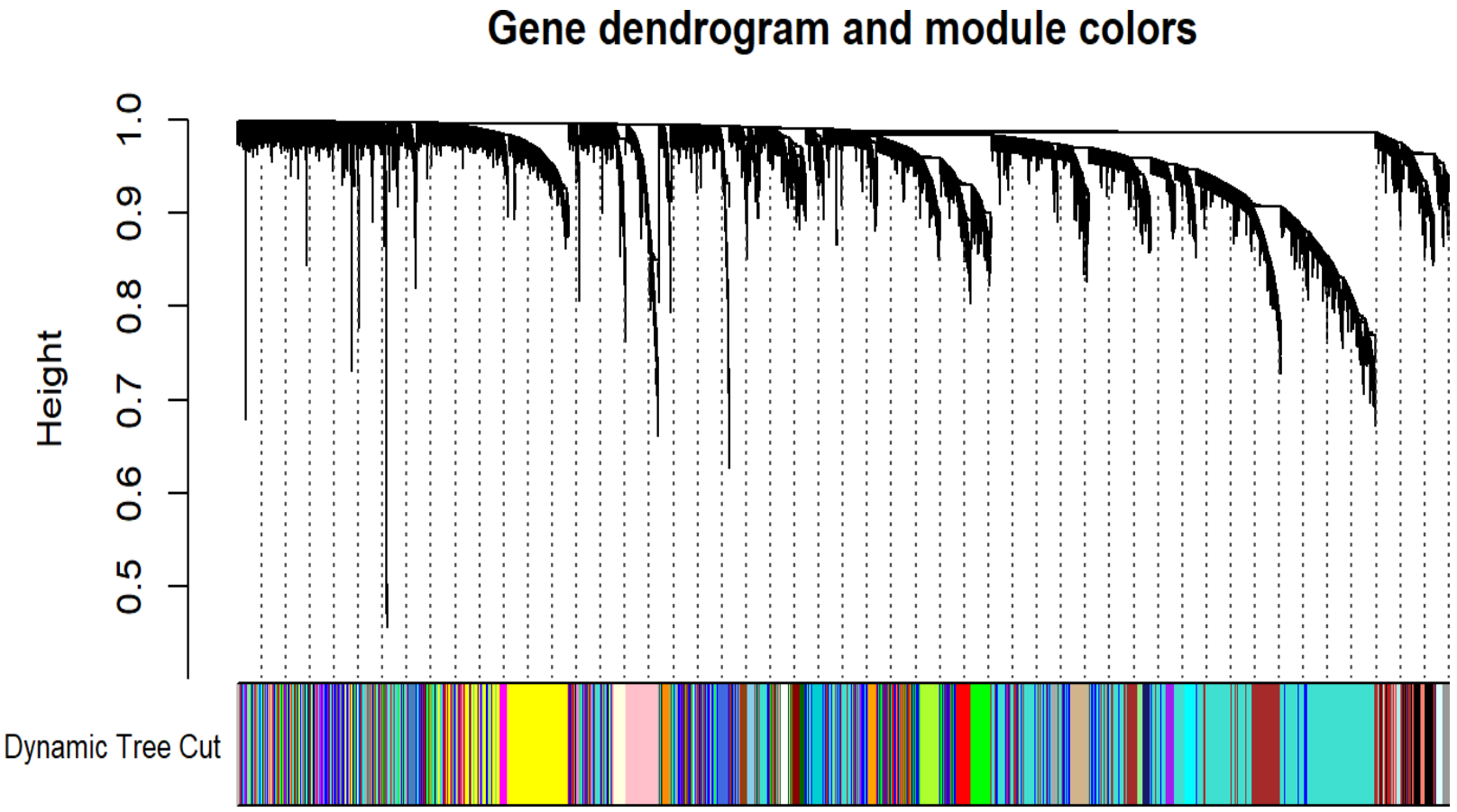


Figure 3. *Gene clustering on TOM-based dissimilarity*

- Branches in the cluster tree (dendrogram) are referred to as modules.
- *Dynamic Tree Cut algorithm has discovered 34 clusters.*
- *There are 79 genes that aren't assigned to any of the clusters.*
- *The biggest cluster has 3455 nodes and the smallest one has 38 nodes.*

Module eigengenes

- To define the module eigengene of a module, we use the Singular Value Decomposition (SVD) of the module expression matrix. [Langfelder P., Horvath S., at el. 2007]

- The gene expression matrix of the I-th module is denoted by

$$X^{(I)} = \left(a_i^{(I)}\right) \qquad (7)$$

where the index $i = 1, 2, \ldots, n_I$ corresponds to module genes and the index $j = 1, 2, \ldots m$ corresponds to samples.

- Gene expression profiles are standartised to mean 0 and variance 1.

- The singular value decomposition of $X^{(I)}$ is denoted by

$$X^{(I)} = UDV^T \qquad (8)$$

where the columns of the orthogonal matrices *U* and *V* are the left- and right-singular vectors, respectively.

Relating genes within a module to the module eigengene

❑ We assume that the singular values $|d_i^{(I)}|$ are arranged in non-increasing order. Adapting terminology from [Oldham M., Horvath S. at el. 2006, Fuller T. at el. 2007, Alter O., Brown P. at el. 2000], we refer to the first column of $V^{(I)}$ as the Module Eigengene: $E^I = v_1^{(I)}$

❑Although our approach emphasizes modules (represented by eigengenes) as the basic building blocks of eigengene networks, it is important to have a measure of how closely related a particular actual gene is to the eigengenes within the co-expression networks.

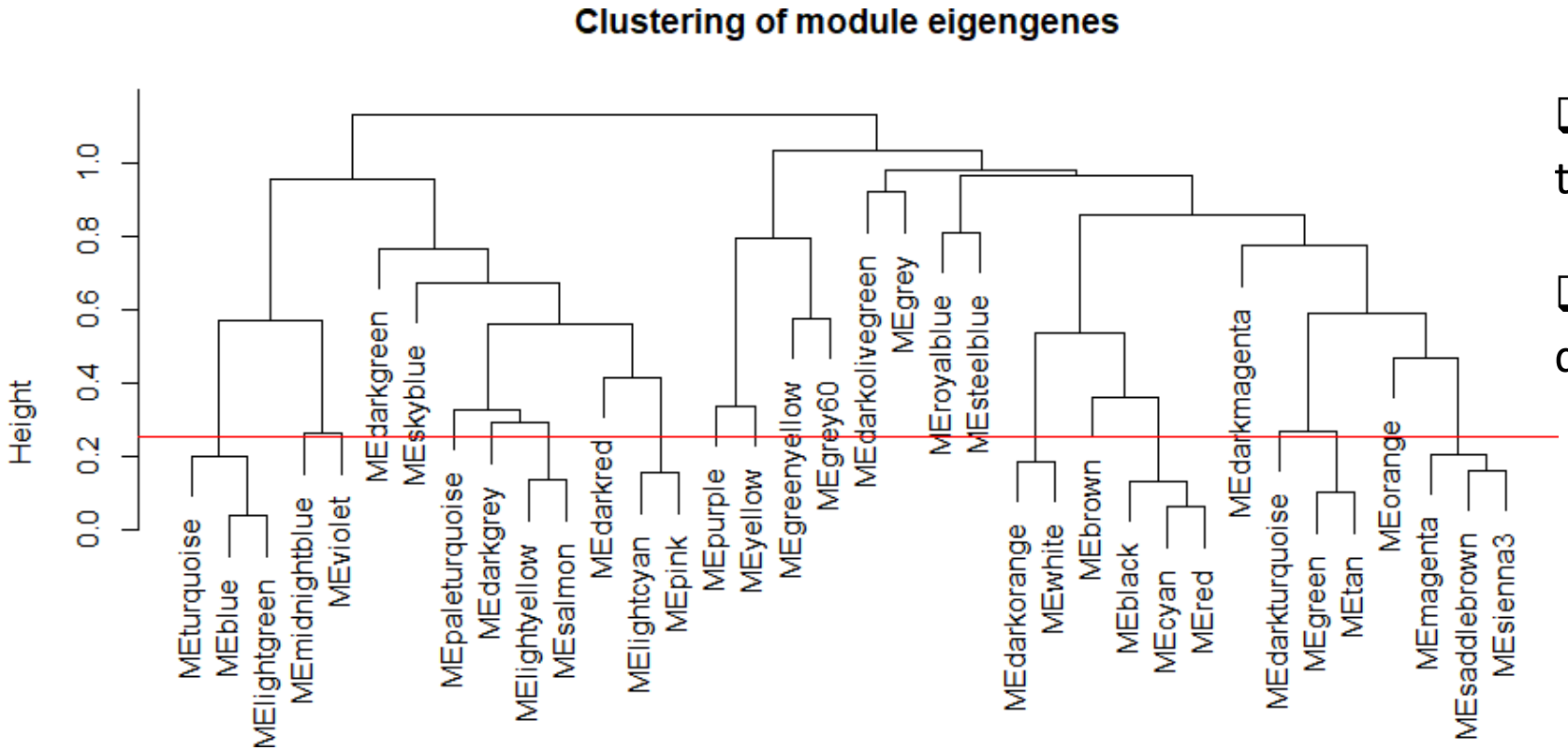
❑A natural measure is the eigengene-based connectivity $k_E(i)$, defined as the correlation between the expression profile of the studied gene x_i and the eigengene E_I.

$$k_{E_I}(j) = cor(x_j, E_I) \qquad (7)$$

❑ The closer $k_E(i)$ is to 1 or -1, the stronger the evidence that the j-th gene is part of the I-th module.

Merging of module eigengenes

❑ ***Constant height cut [***Langfelder P., Zhang B., at el. 2008***]***

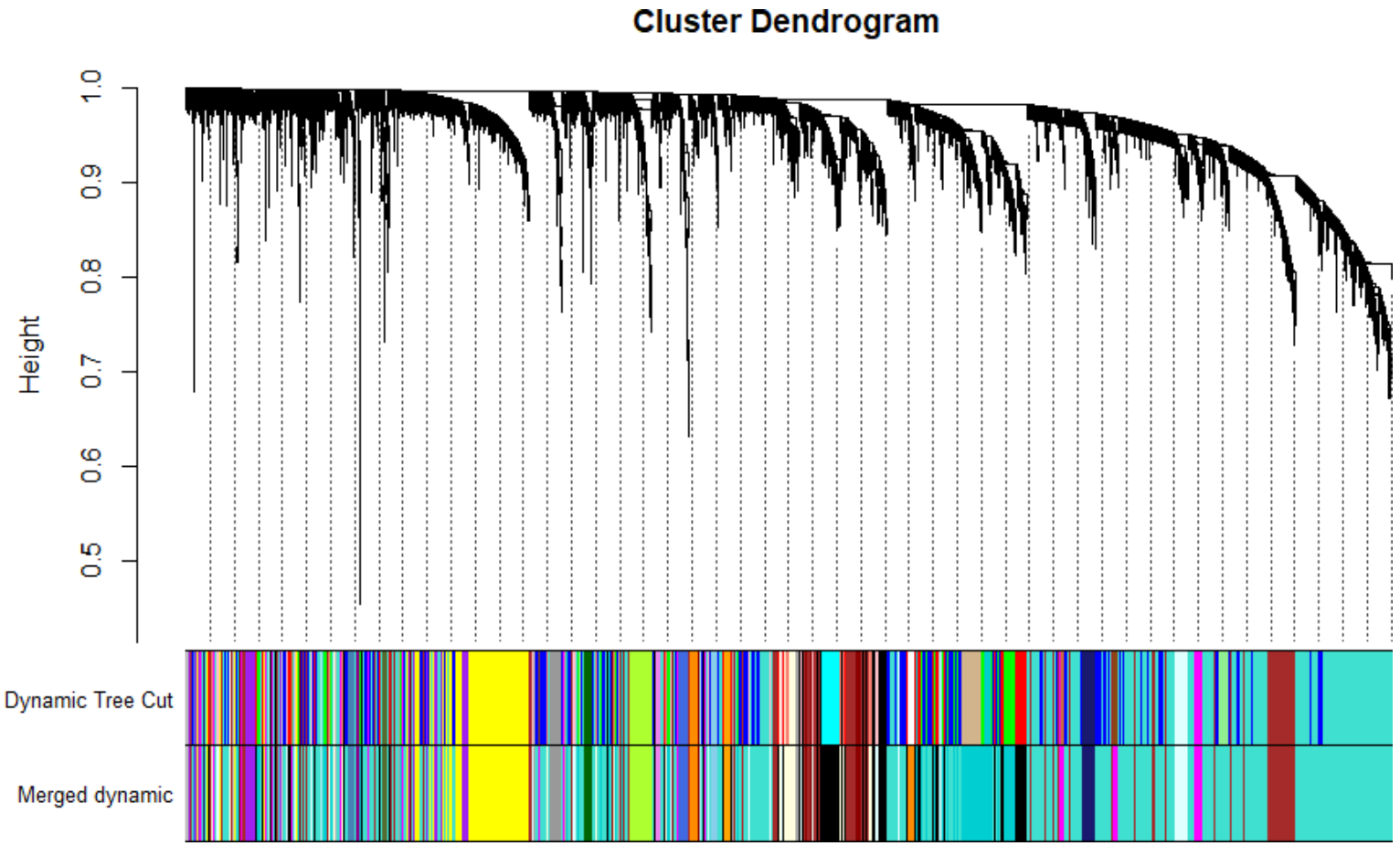


❑ Height cut is 0.25 corresponding to correlation 0.75 to merge.

❑Merge modules whose dissimilarity is below the cutoff

Figure 4. *Clustering dendrogram of genes, with dissimilarity based on topological overlap.*

Gene dendogram and detected modules, before and after merging



- The Dynamic Tree Cut may identify modules whose expression profiles are very similar.
- To quantify co-expression similarity of entire modules, we calculate their eigengenes and cluster them on their correlation.

Figure 5: *Clustering dendrogram of genes, with dissimilarity based on topological overlap, together with assigned merged module colours and the original module colours.*

Gene modules

Module	*Nr. of genes*	*Module*	*Nr. of genes*
0	79	17	142
1	3455	18	139
2	1785	19	130
3	1544	20	124
4	1019	21	118
5	770	22	114
6	594	23	113
7	389	24	94
8	375	25	90
9	284	26	89
10	236	27	85
11	219	28	82
12	212	29	77
13	197	30	76
14	180	31	46
15	174	32	41
16	166	33	38

Table 1. *Modules and genes*

- ❑ 34 modules
- ❑ 79 genes that aren't assigned to any of the modules.
- ❑The biggest module has 3455 genes and the smallest one has 38 genes.

Merged Modules

Module	*Nr. of genes*
0	79
2	2177
3	4999
4	1019
5	989
7	1226
8	375
9	284
10	236
17	621
19	130
20	124
23	203
24	94
26	89
27	317
28	82
29	77
30	76
32	41
33	38

Table 2. *Merged modules and genes*

- ❑ 21 modules
- ❑79 genes that aren't assigned to any of the modules.
- ❑The biggest module has 4999 genes and the smallest one has 38 genes.

Merged Modules

Module 2	SEMA6B	USP32	OCIAD1	TARBP1
	IQCF3	SEMA5A	TNIP1	LOR
	CDH1	EDAR	NFE2	LMAN1
	SH3GL3	HLF	KRT2	TFE3
	KCNK9	RBM20	MAST2	C8orf46

Module 4 (Genes related to obesity)	PCSK1	DPYSL5	KLF17	BCHE
	TENC1	LEP	VAX2	ADAM6
	ROBO4	AL832737	RASL10A	APOE
	FTO	SLC30A3	LEPR	ISLR2
	INSIG2	MC4R	PPARG	PABL

Module 3	NBL1	CDK5R1	CLCA1	GDPD5
	RBMXL1	BTG1	CACHD1	DHX35
	IQCF5	AP2B1	DCAF11	FAM3A
	FAM101A	LINC00896	ZNF786	UBA5
	MGAT4A	ORC6	SFMBT2	ZFAND6

Module 7	PYGL	FANCE	CECR6	ATG16L2
	TBX15	TEX14	FAM49A	NFAT5
	DYRK1A	AIFM2	TNFSF14	BC105019
	SMCO4	XRCC4	PPM1D	AGO4
	PUS3	MAEA	GREB1	AICDA

Table 3. *Merged modules and genes*

Potential driver genes in merged modules

Module	***Potential drive genes***	***Gene Name***	***Diseases related with the gene***
2	GAB2	Associated Binding Protein 2	Leukemia
3	TNFAIP8	TNF Alpha Induced Protein 8	
Obesity module	***FTO***	Fat mass and obesity associated	Obesity
5	RAVER1	Ribonucleoprotein, PTB Binding 1	
7	TXLNG2P	Taxilin Gamma Pseudogene, Y-Linked	
8	TBX21	T-Box Transcription Factor 21	Asthma
9	MYH9	Myosin Heavy Chain 9	
10	DEDD2	Death Effector Domain Containing	

Hub genes are those genes which are highly connected within a module.

Table 4. *Potential driver genes in merged modules*

Obesity module

Betweeness	*Degree*	*Eigenvector*	*Pagerank*
FTO	FTO	FTO	FTO
PJA2	LEPR	MC4R	MC4R
CDC25B	USP12	WDTC1	WDTC1
ADIPOQ	FNDC3B	DEDD2	DEDD2
PPARG	RNF14	PPARG	PPARG
ZMYM2	PJA2	ZMYM2	ZMYM2
PCSK1	LEP	PCSK1	PCSK1
LEP	CDC25B	LEP	LEP
LEPR	ADIPOQ	LEPR	LEPR
USP12	PPARG	USP12	USP12
FNDC3B	ZMYM2	FNDC3B	FNDC3B
RNF14	PCSK1	RNF14	RNF14
INSIG2	NEDD9	NEDD9	NEDD9
MC4R	POC1B	POC1B	POC1B
WDTC1	BLOC1S6	BLOC1S6	BLOC1S6

***Table 5**. Most central genes in obesity module*

Eigengene-based connectivity for gene FTO

Module	$k_{E_I}(j)$
Module 2	0.4722639
Module 3	0.385879
Module 5	0.4309691
Obesity Module	0.8428542
Module 7	0.3270049

***Table 6**. Eigengene-based connectivity for gene FTO*

Obesity module

Betweeness	*Degree*	*Eigenvector*	*Pagerank*
49682.98	325	0.8332058	0.001014406
52707.62	331	0.8411068	0.00101559
53369.47	334	0.8457187	0.001019887
57272.49	335	0.8546221	0.001037688
59393.29	336	0.854761	0.001049783
64382.96	336	0.8589078	0.001050863
65115.81	338	0.8602408	0.001051442
65708.64	339	0.8637388	0.001053788
65900.6	341	0.8644066	0.001064638
71165.68	343	0.8656035	0.001067329
71995.89	344	0.8695161	0.001074412
76783.27	345	0.874374	0.001083625
76855.16	351	0.8903137	0.001091227
80357.86	361	0.9055079	0.001104277
83774.1	367	0.9152965	0.001147849

***Table 7**. Centralities for most central nodes in obesity module*

	Betweeness	*Degree*	*Eigenvector*	*Pagerank*
Betweeness	1	0.958263	0.964777	0.931858
Degree	0.958263	1	0.992599	0.942887
Eigenvector	0.964777	0.992599	1	0.962425
Pagerank	0.931858	0.942887	0.962425	1

***Table 8**. Correlation of centralities measures*

Conclusions

❑A network visualizes the relationships of a dataset in one graph. This unique ability of data representation is combined with many measures that are helpful for many research disciplines.

❑ Co-expression networks can be very powerful tools for identification of diseases genes.

❑Average linkage hierarchical clustering is used to detect modules in co-expression network giving as a parameter the topological overlap based dissimilarity.

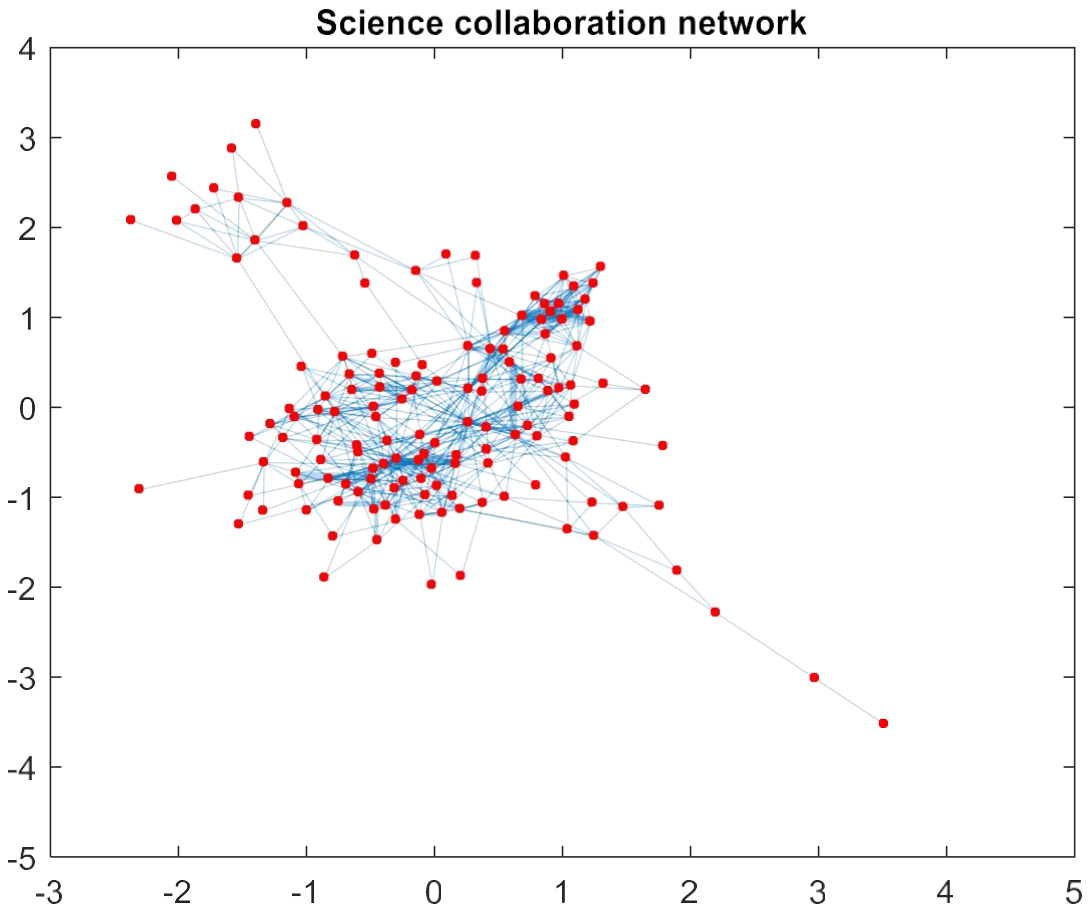
❑ Dynamic Tree Cut algorithm had detected 34 different modules in our gene co-expression network.

❑ Module related to obesity has 1019 genes and the potential driver gene for this module was FTO.

❑Eigengene connectivity is used to measure the strength of the relationship of a particular gene to a given eigengene. In the obesity module, it was found that gene FTO is strongly connected with this module.

❑One great challenge for us would be to associate each module to a diseases and to find for each of the diseases the potential driver gene.

“We are all part of some networks”



References

1 S.N.Dorogovtsev, J.F.F. Mendes (2004), Evolution of Networks
2 Reka Albert, Albert-Lazlo Barabasi (2002), Statistical mechanics of complex networks, Reviews of Modern Physics, 74.
3 M.E.J. Newman (2003), The Structure of Complex Networks, SIAM Review 45, 2, 167-256.
4 Thomas Fuhrmann, On the Topology of Overlay-Networks
5 Martin Rosvall (2003), Complex Networks and Dynamics of an Information Network Model.
6 Ulrike Muckstein (2002), Statistical Mechanics of Complex Networks, Institute for Theoretical Chemistry and Sructural Biology, University Vienna.
7 W.W.Powell, D.R.White, K.W.Koput, J.O.Smith (2002), Networks Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences
8Barabási, AL., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. Nat Rev Genet 12, 56–68 (2011) https://doi.org/10.1038/nrg2918