Reproduce Part of WGCNA in Finding Weight-related Gene Modules

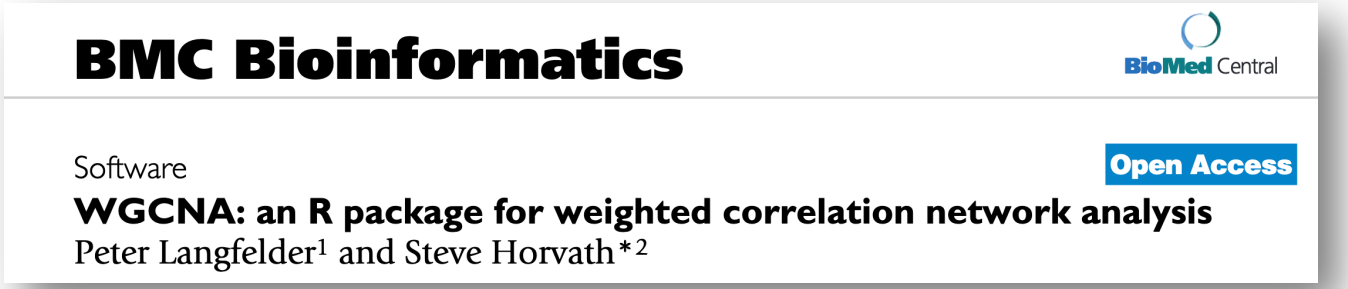
Bioinformatics Final Project

12212859 Sijie Li

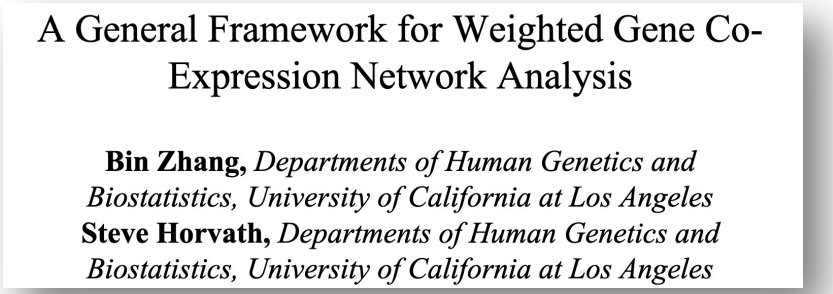
June 3th, 2025

Curious When Using WGCNA…

- **Network construction (esp. soft threshold)?**

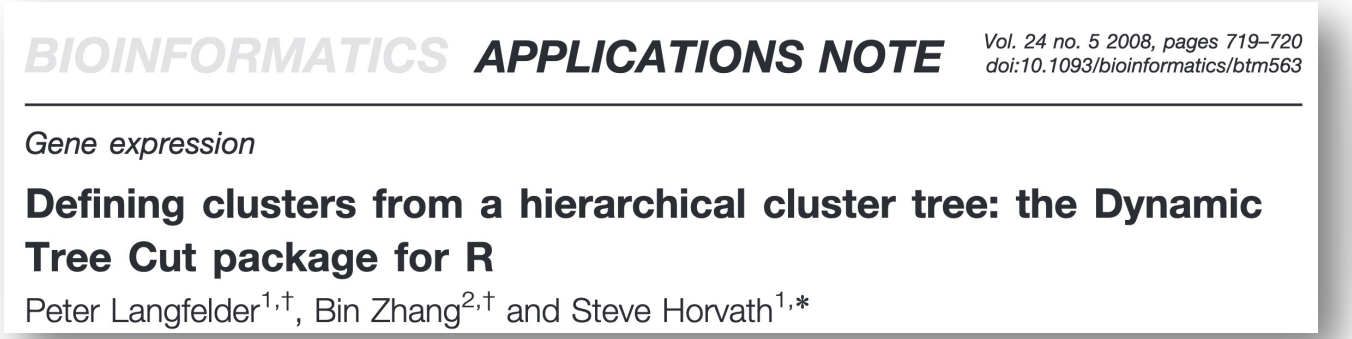


2008, Cited: 22,262

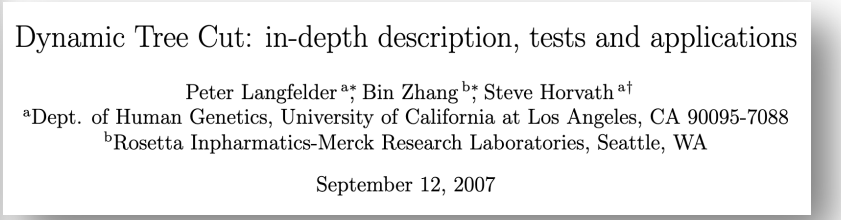


2005, Cited: 5,970

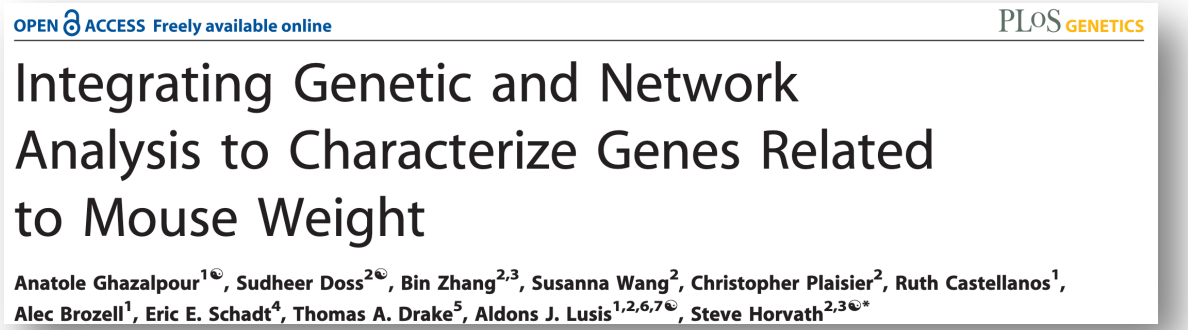
- **Dynamic Tree cut?**



2008, Cited: 2,2136



Apply to a senario


OPEN ACCESS Freely available online PLoS GENETICS

Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight

Anatole Ghazalpour[1☯], Sudheer Doss[2☯], Bin Zhang[2,3], Susanna Wang[2], Christopher Plaisier[2], Ruth Castellanos[1], Alec Brozell[1], Eric E. Schadt[4], Thomas A. Drake[5], Aldons J. Lusis[1,2,6,7☯], Steve Horvath[2,3☯*]

2006, Cited: 544

Reproduce

- **Construct Network**
- **Module Detection**
- **mQTL**

  (Module Quantitative Trait Loci)

➢ **Data** **No WT!**

- **female F2 ApoE -/- high fat Western diet** mice livers
- dye-switch microarray
- **gene expression "variation"**

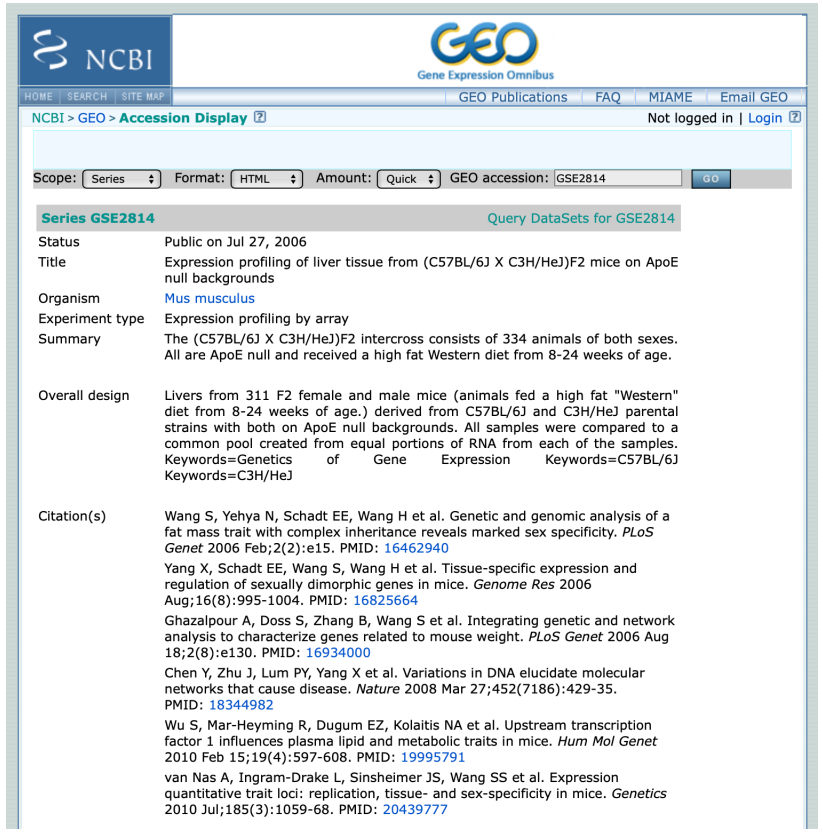
$$\log_{10}\left(\frac{\text{Expression of gene A in sample1}}{\text{average expression of gene A in 150 random selected female samples}}\right)$$

➢ **Result**

- weight-related gene module
- potential regulatory locus of the module

Data Obtain

➢GSE2814



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

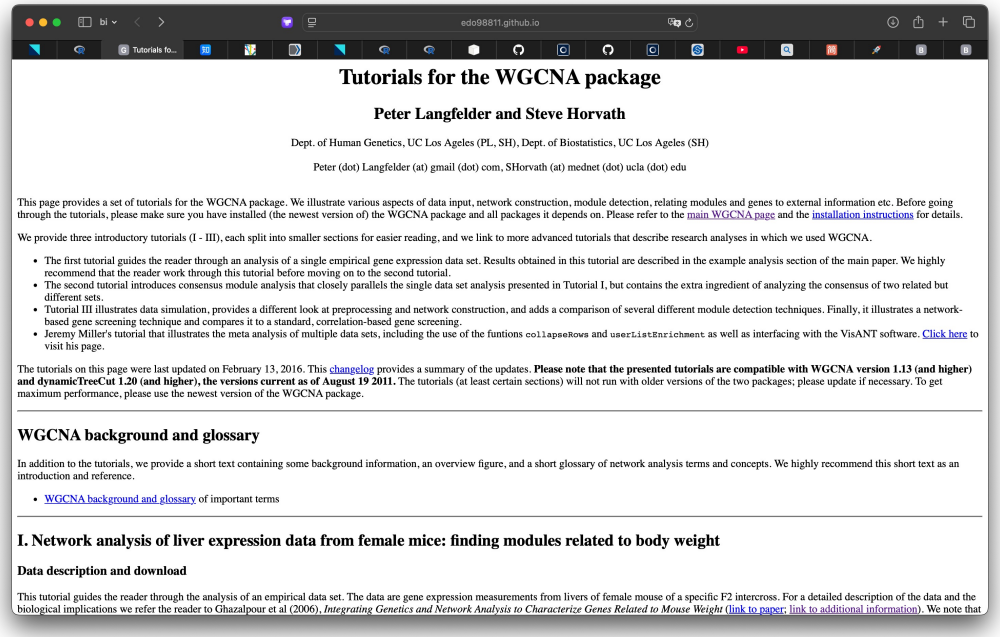
NCBI > GEO > **Accession Display** Not logged in | Login

Scope: Series Format: HTML Amount: Quick GEO accession: GSE2814 GO

Series GSE2814 Query DataSets for GSE2814

Status	Public on Jul 27, 2006
Title	Expression profiling of liver tissue from (C57BL/6J X C3H/HeJ)F2 mice on ApoE null backgrounds
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	The (C57BL/6J X C3H/HeJ)F2 intercross consists of 334 animals of both sexes. All are ApoE null and received a high fat Western diet from 8-24 weeks of age.
Overall design	Livers from 311 F2 female and male mice (animals fed a high fat "Western" diet from 8-24 weeks of age.) derived from C57BL/6J and C3H/HeJ parental strains with both on ApoE null backgrounds. All samples were compared to a common pool created from equal portions of RNA from each of the samples. Keywords=Genetics of Gene Expression Keywords=C57BL/6J Keywords=C3H/HeJ
Citation(s)	Wang S, Yehya N, Schadt EE, Wang H et al. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet* 2006 Feb;2(2):e15. PMID: 16462940 Yang X, Schadt EE, Wang S, Wang H et al. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res* 2006 Aug;16(8):995-1004. PMID: 16825664 Ghazalpour A, Doss S, Zhang B, Wang S et al. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* 2006 Aug 18;2(8):e130. PMID: 16934000 Chen Y, Zhu J, Lum PY, Yang X et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008 Mar 27;452(7186):429-35. PMID: 18344982 Wu S, Mar-Heyming R, Dugum EZ, Kolaitis NA et al. Upstream transcription factor 1 influences plasma lipid and metabolic traits in mice. *Hum Mol Genet* 2010 Feb 15;19(4):597-608. PMID: 19995791 van Nas A, Ingram-Drake L, Sinsheimer JS, Wang SS et al. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics* 2010 Jul;185(3):1059-68. PMID: 20439777

➢Tutorial dataset



edo98811.github.io

Tutorials for the WGCNA package

Peter Langfelder and Steve Horvath

Dept. of Human Genetics, UC Los Ageles (PL, SH), Dept. of Biostatistics, UC Los Ageles (SH)

Peter (dot) Langfelder (at) gmail (dot) com, SHorvath (at) mednet (dot) ucla (dot) edu

This page provides a set of tutorials for the WGCNA package. We illustrate various aspects of data input, network construction, module detection, relating modules and genes to external information etc. Before going through the tutorials, please make sure you have installed (the newest version of) the WGCNA package and all packages it depends on. Please refer to the main WGCNA page and the installation instructions for details.

We provide three introductory tutorials (I - III), each split into smaller sections for easier reading, and we link to more advanced tutorials that describe research analyses in which we used WGCNA.

- The first tutorial guides the reader through an analysis of a single empirical gene expression data set. Results obtained in this tutorial are described in the example analysis section of the main paper. We highly recommend that the reader work through this tutorial before moving on to the second tutorial.
- The second tutorial introduces consensus module analysis that closely parallels the single data set analysis presented in Tutorial I, but contains the extra ingredient of analyzing the consensus of two related but different sets.
- Tutorial III illustrates data simulation, provides a different look at preprocessing and network construction, and adds a comparison of several different module detection techniques. Finally, it illustrates a network-based gene screening technique and compares it to a standard, correlation-based gene screening.
- Jeremy Miller's tutorial that illustrates the meta analysis of multiple data sets, including the use of the funtions `collapseRows` and `userListEnrichment` as well as interfacing with the VisANT software. Click here to visit his page.

The tutorials on this page were last updated on February 13, 2016. This changelog provides a summary of the updates. **Please note that the presented tutorials are compatible with WGCNA version 1.13 (and higher) and dynamicTreeCut 1.20 (and higher), the versions current as of August 19 2011.** The tutorials (at least certain sections) will not run with older versions of the two packages; please update if necessary. To get maximum performance, please use the newest version of the WGCNA package.

WGCNA background and glossary

In addition to the tutorials, we provide a short text containing some background information, an overview figure, and a short glossary of network analysis terms and concepts. We highly recommend this short text as an introduction and reference.

- WGCNA background and glossary of important terms

I. Network analysis of liver expression data from female mice: finding modules related to body weight

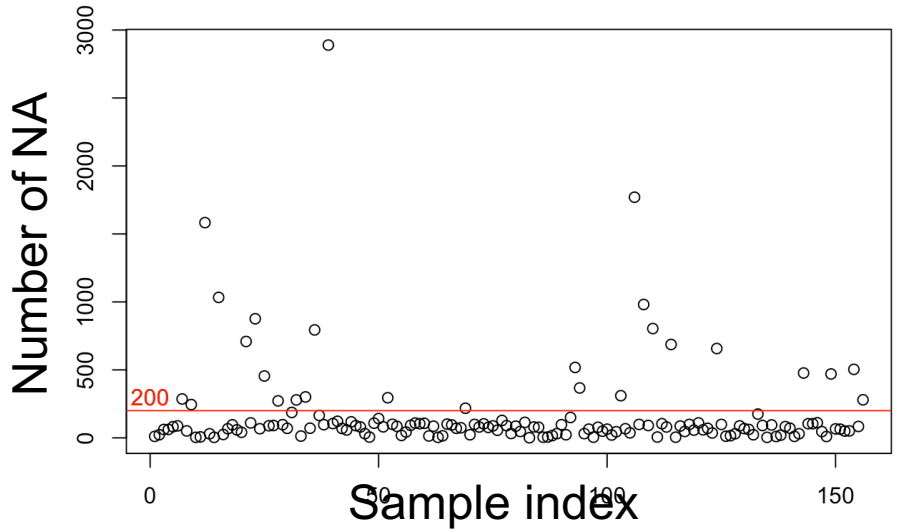
Data description and download

This tutorial guides the reader through the analysis of an empirical data set. The data are gene expression measurements from livers of female mouse of a specific F2 intercross. For a detailed description of the data and the biological implications we refer the reader to Ghazalpour et al (2006), *Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight* (link to paper; link to additional information). We note that

- **3600** genes
- **135** samples

Process Data--Sample Reduction

1. Trim annotation → **23,388** genes × **311** sample matrix
2. Extract **female** sample
3. Filter NA, remove sample with
   - > 200 NA expression
   - any NA trait

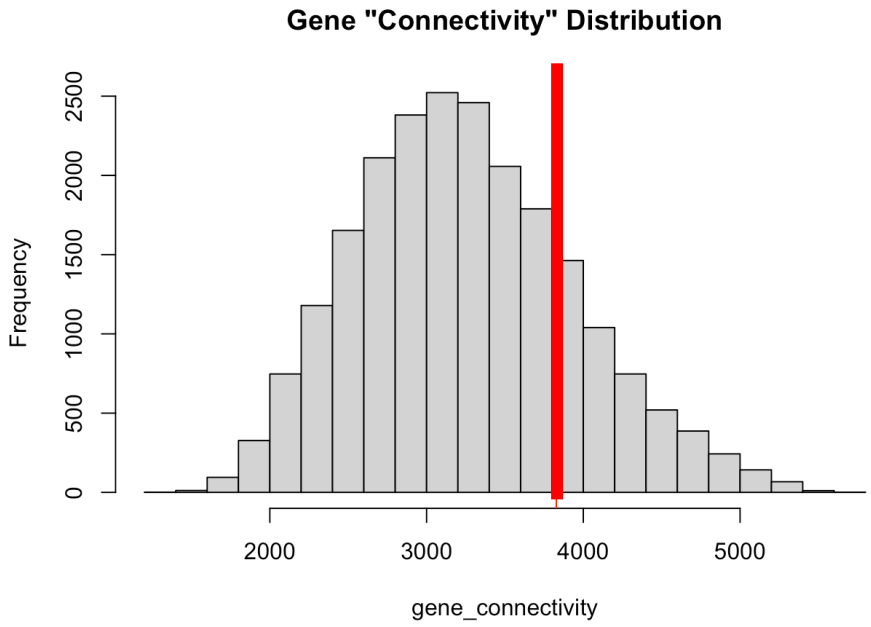


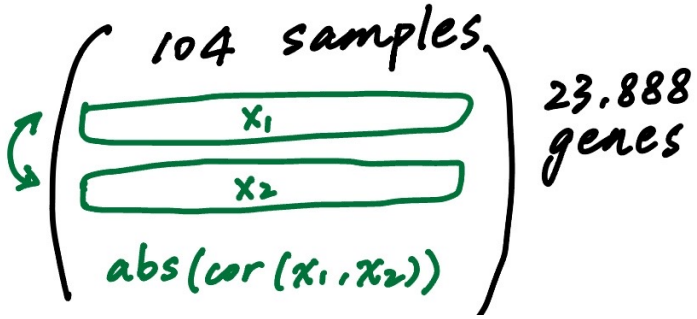
and comparison of the male network will be reported elsewhere. Only those mice with complete phenotype, genotype, and array data were used. This gave a final experimental sample of 135 female mice used

104 samples left

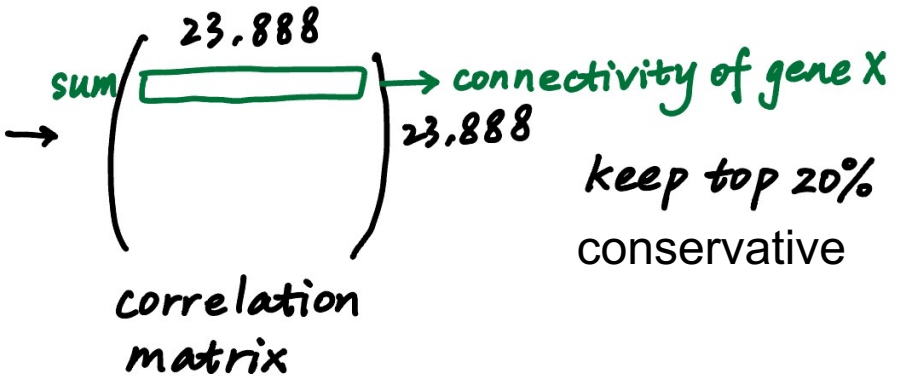
Process Data--Gene Reduction

4. Remove genes with NA records
5. Select genes with high connectivity





Absolute Pearson Correlation

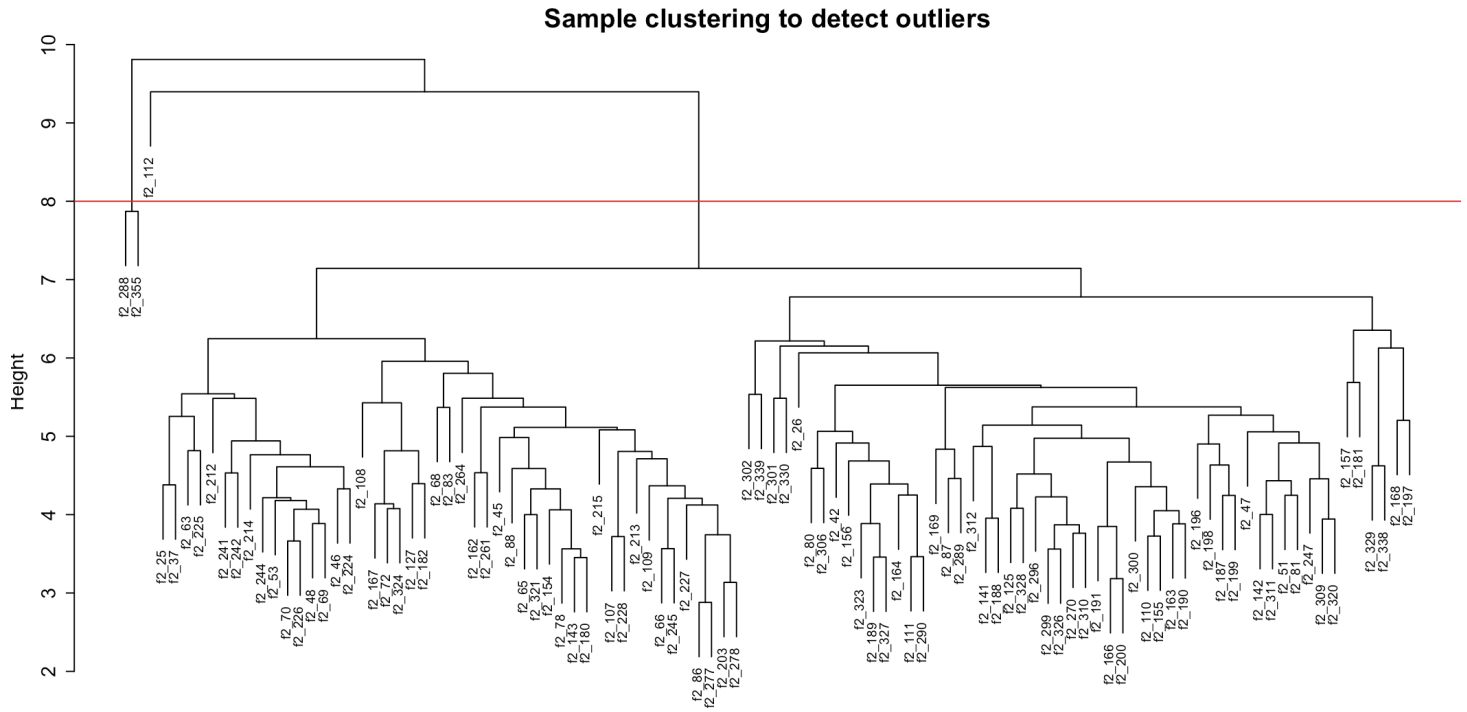


keep top 20%

conservative

23,388 × 311 → 4,319 × 104 VS 3,600 × 135

Check by Sample Clustering



Error source 1: Data

- Do not remove outliers
- Differ in sample filtration
- Differ in gene selection

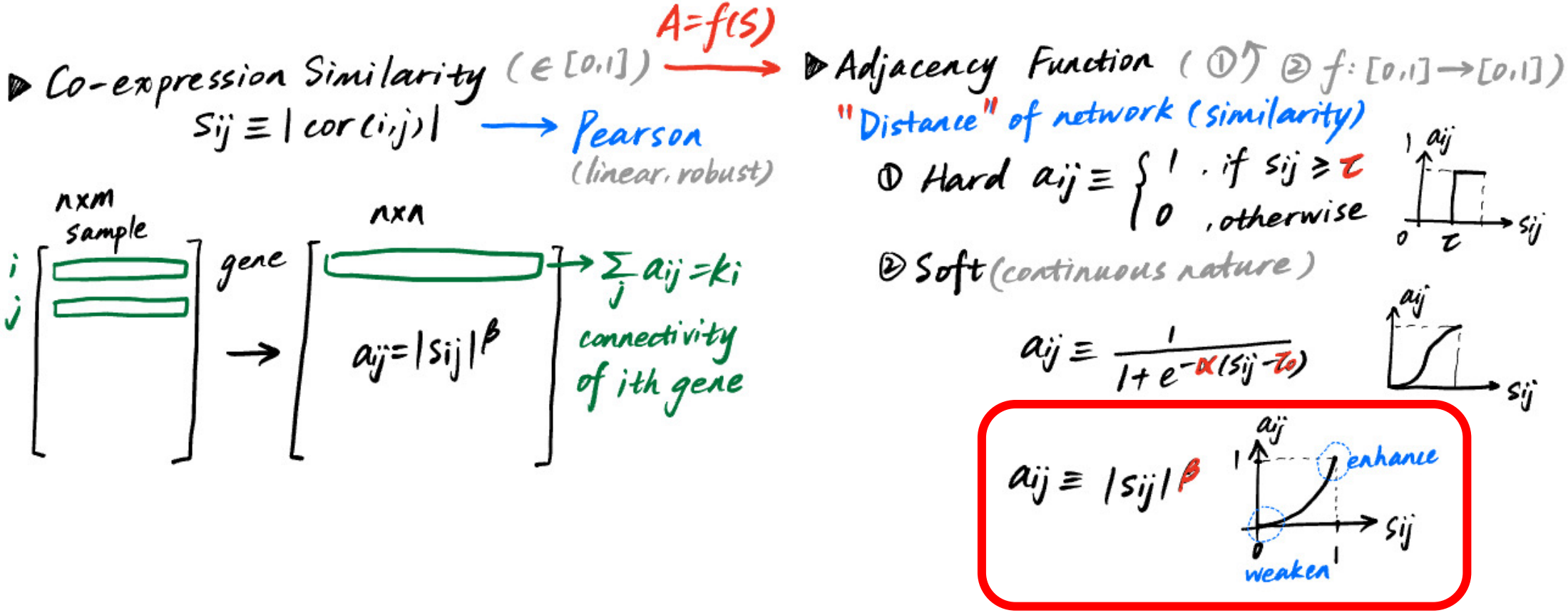
💡Use WGCNA to analyze my data to evaluate the effect of different data.

Cluster the downstream analyze data

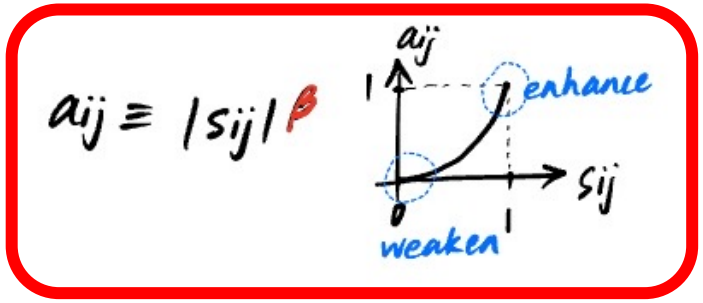
23,388 × 311 → 4,319 × 104 VS 3,600 × 135

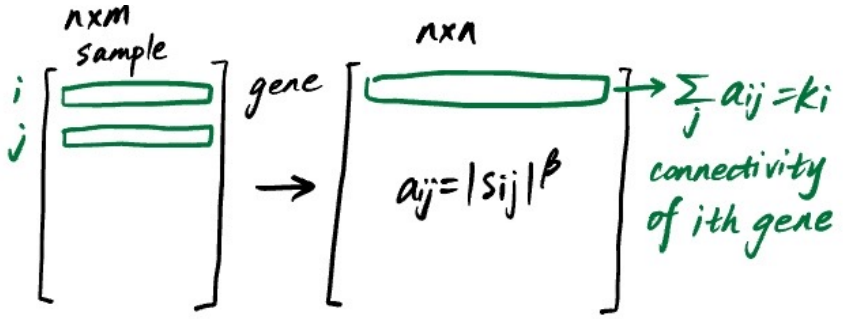
Construct Network--Overview

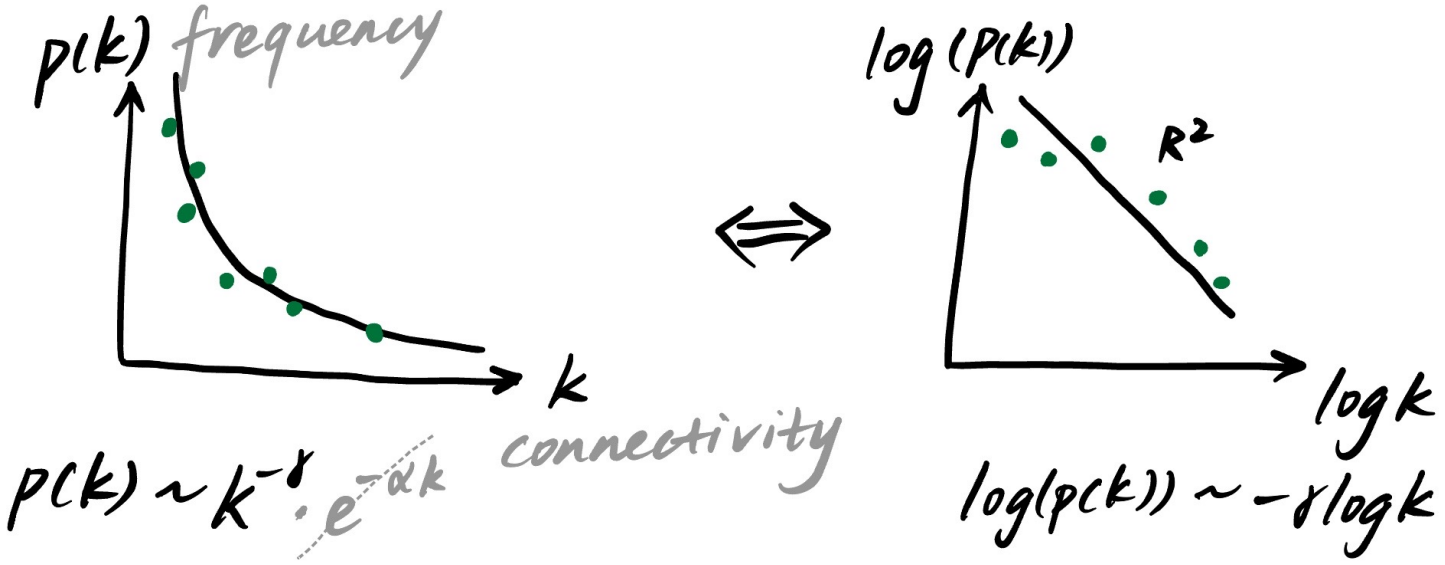
- Node (biological objects)
- Edge (**relationship** between biological quantitives of the objects)



Network Property





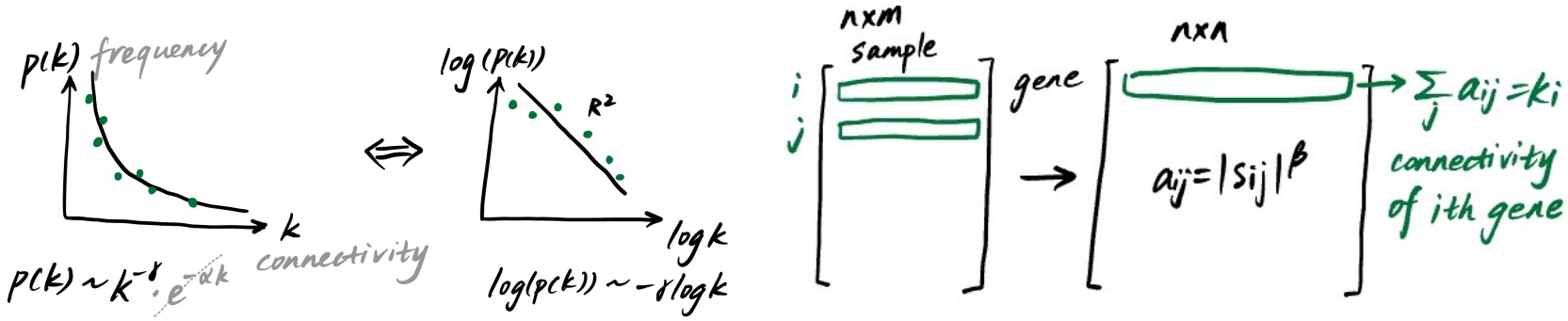


Scale-free

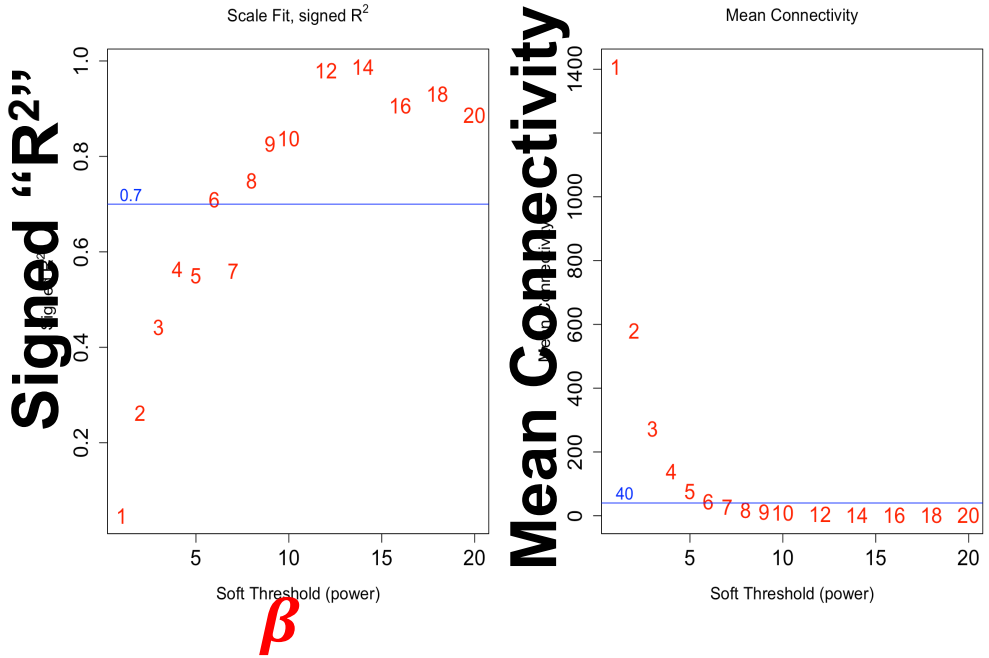
Property of many biological networks

- ✓ R^2 > threshold
- ✓ High mean connectivity
- ✓ Signed "R^2" (slope)

Soft Threshold β



Mean Connectivity
$\sum_i k_i/n$

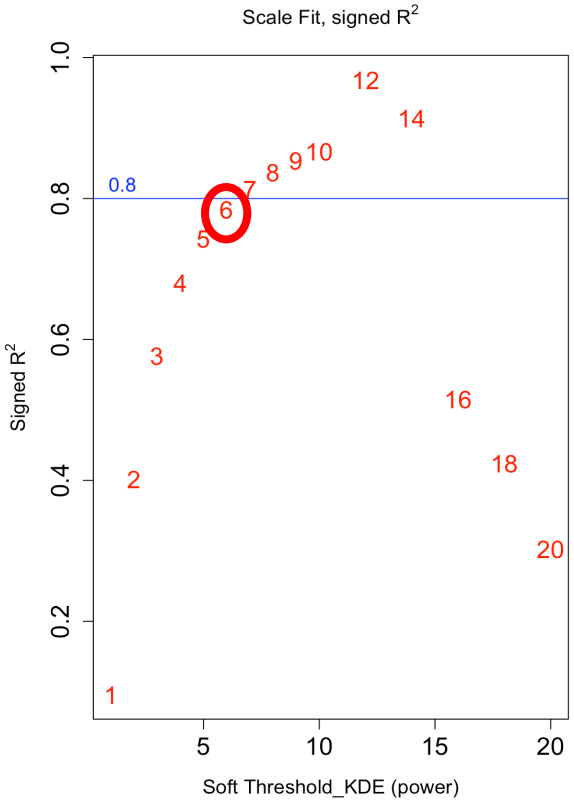


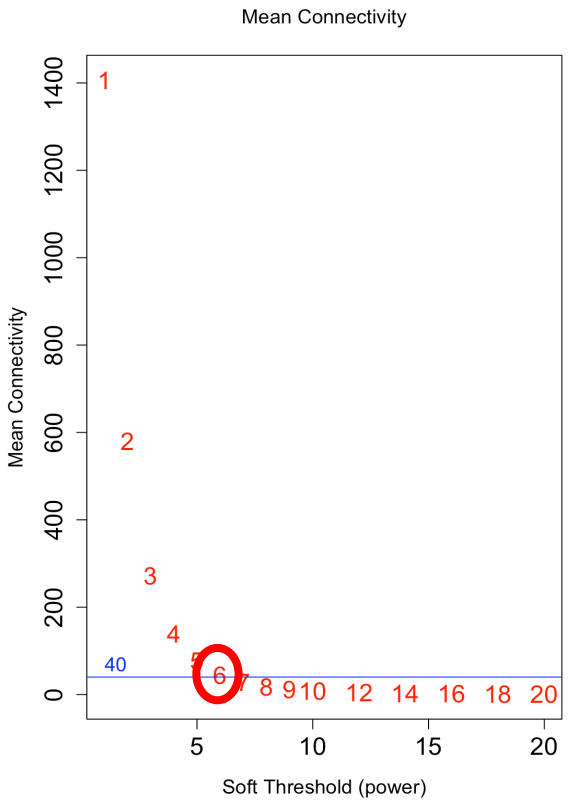
p(k) for continuous k?

- Bin, size?
- Kernel Density Estimation (KDE)

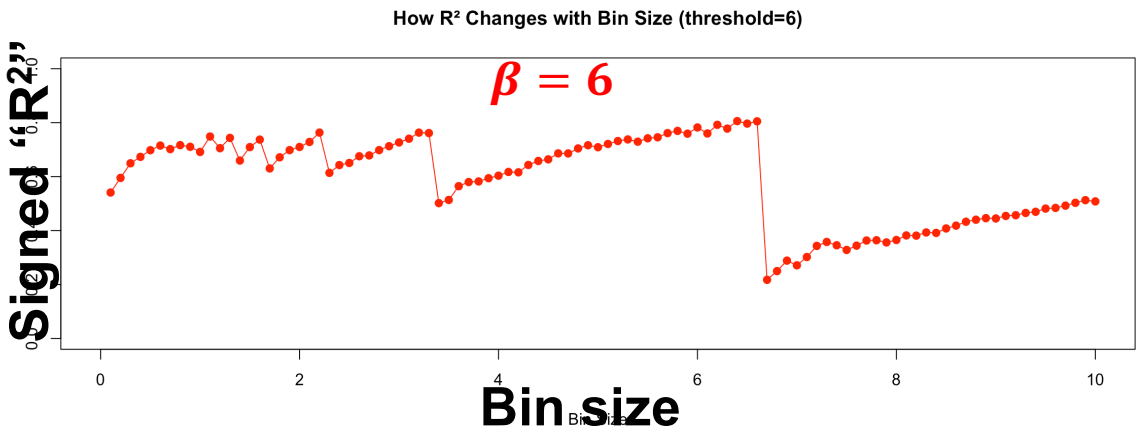
Soft Threshold $\beta = 6$ Also indicated by the paper

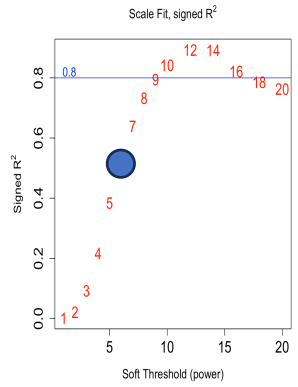
➢ KDE



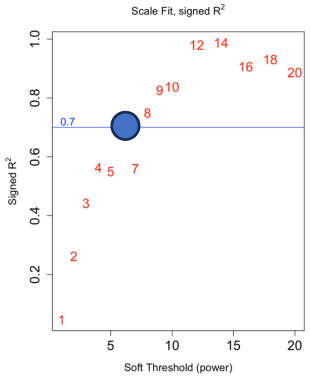


➢ Bin—size matters, a lot!

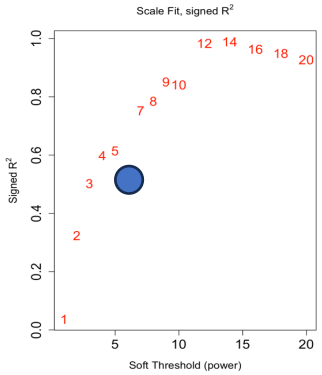




Size=0.1



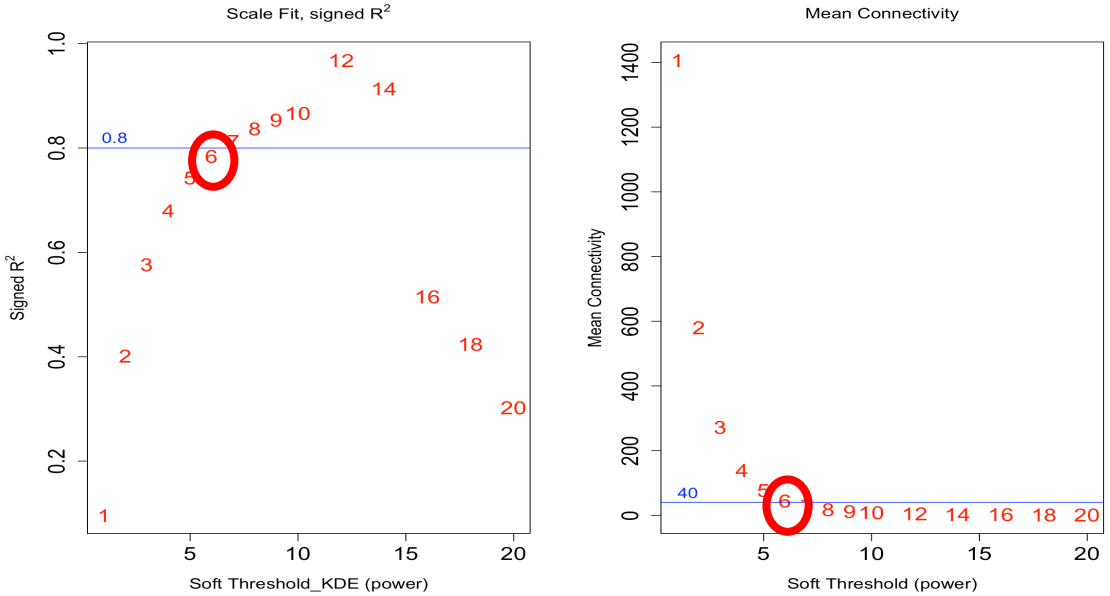
Size=5



Size=10

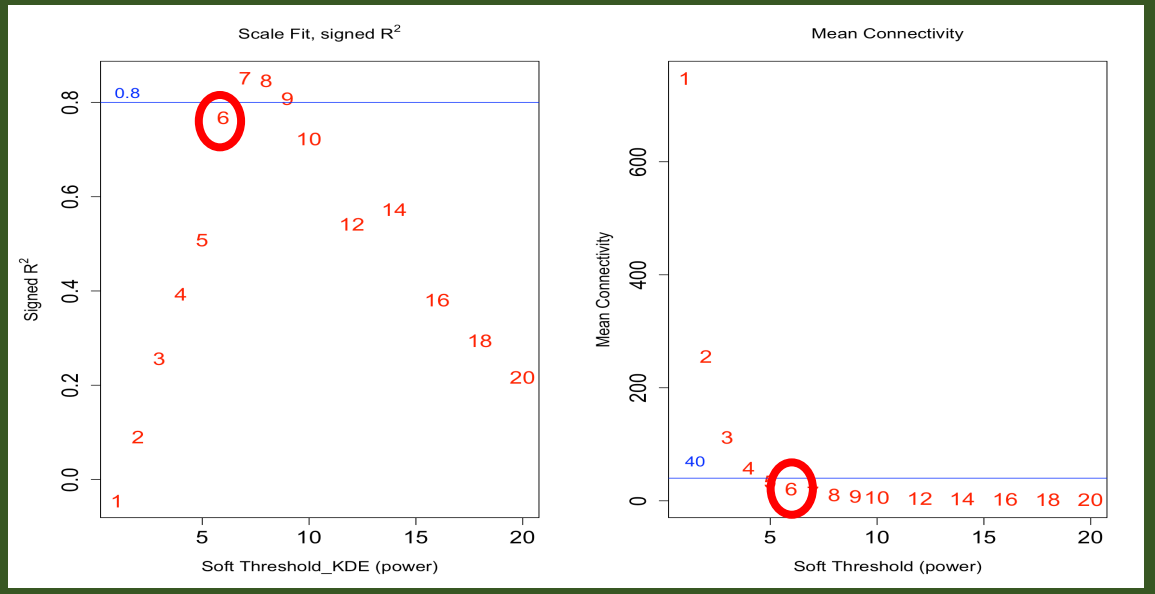
Soft Threshold $\beta = 6$

➢ KDE

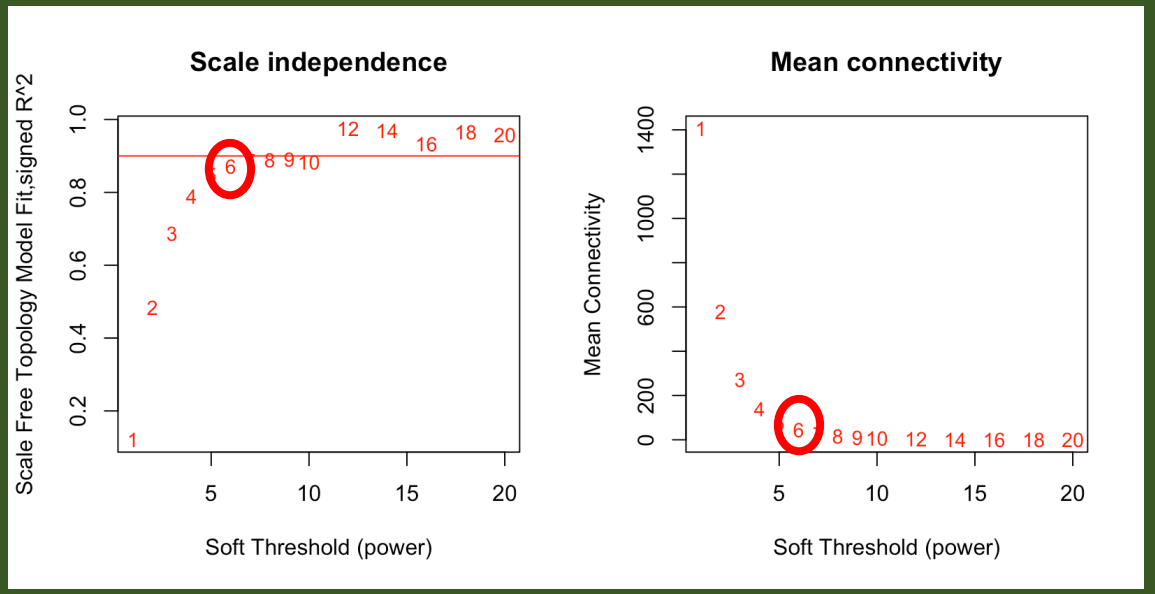


🤔 Though different, 6 can be pick out by all of it.
And **my data seems to work,** for the same soft threshold recommended by WGCNA

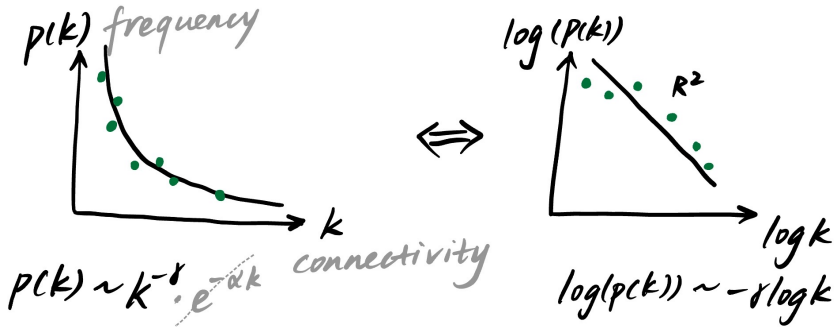
Tutorial Data, My Method (KDE)



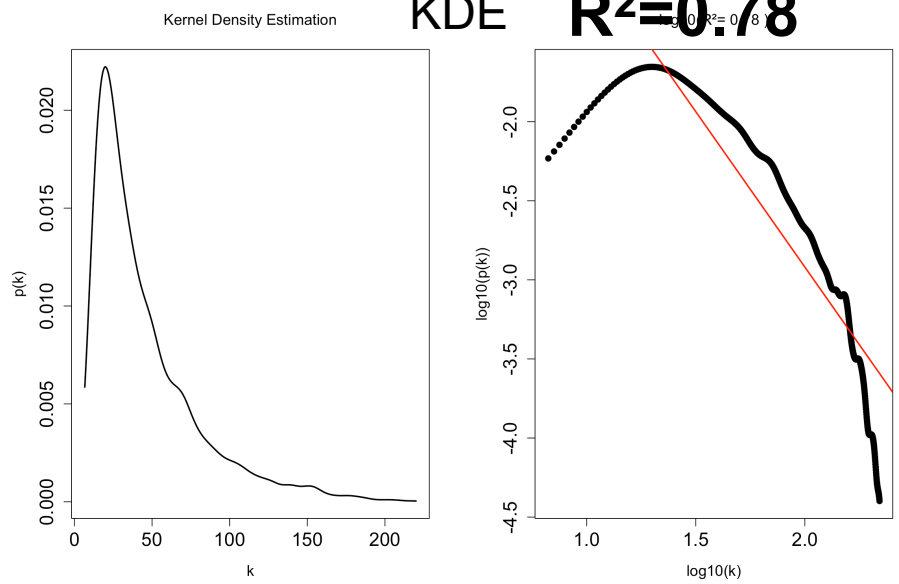
WGCNA Package, My Data (recommend 6)



Validation: KDE $\beta = 6$



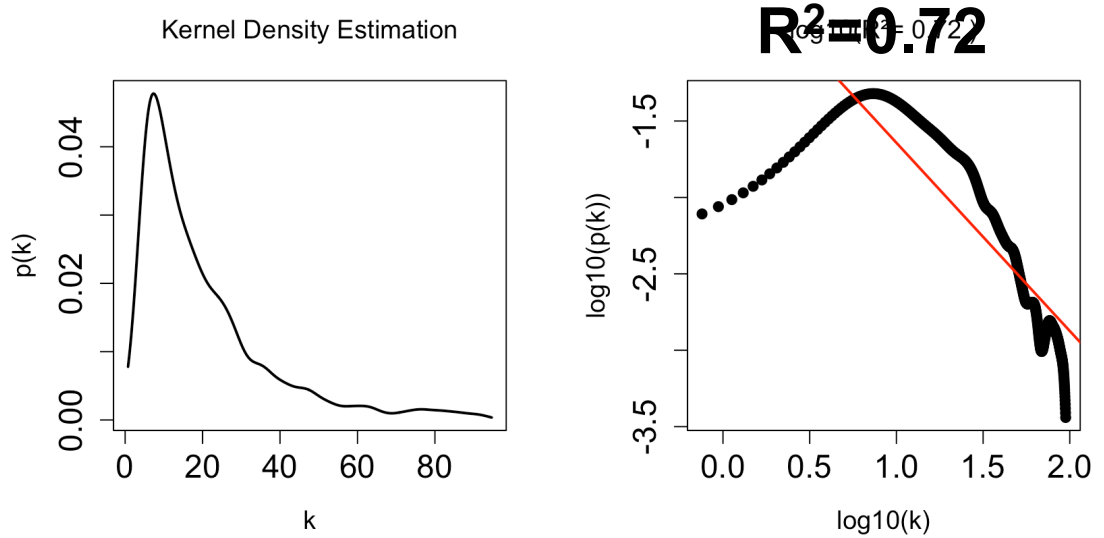
KDE R^2=0.78



🤔 Both Method and Data matters
But do not vary a lot (I think)

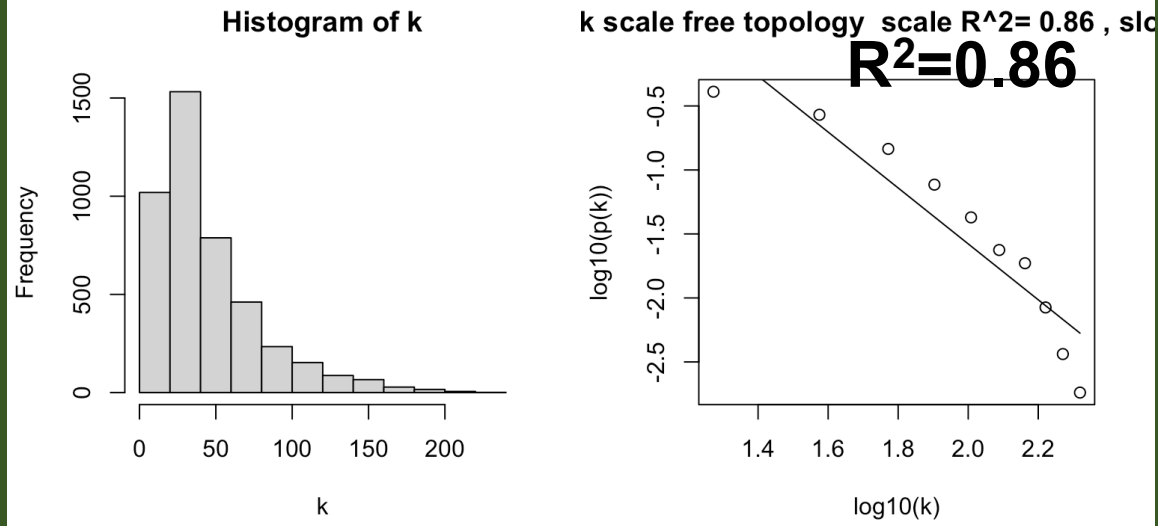
Tutorial Data, My Method (KDE)

R^2=0.72

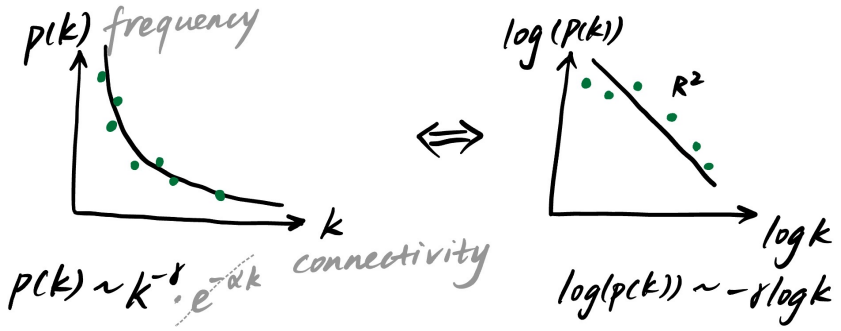


WGCNA Package, My Data

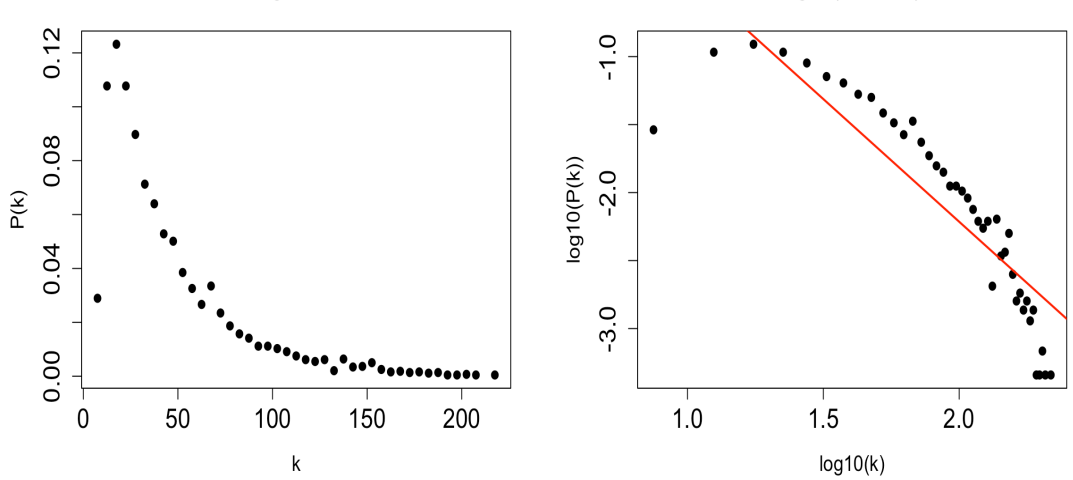
R^2=0.86



Validation: bin $\beta = 6$

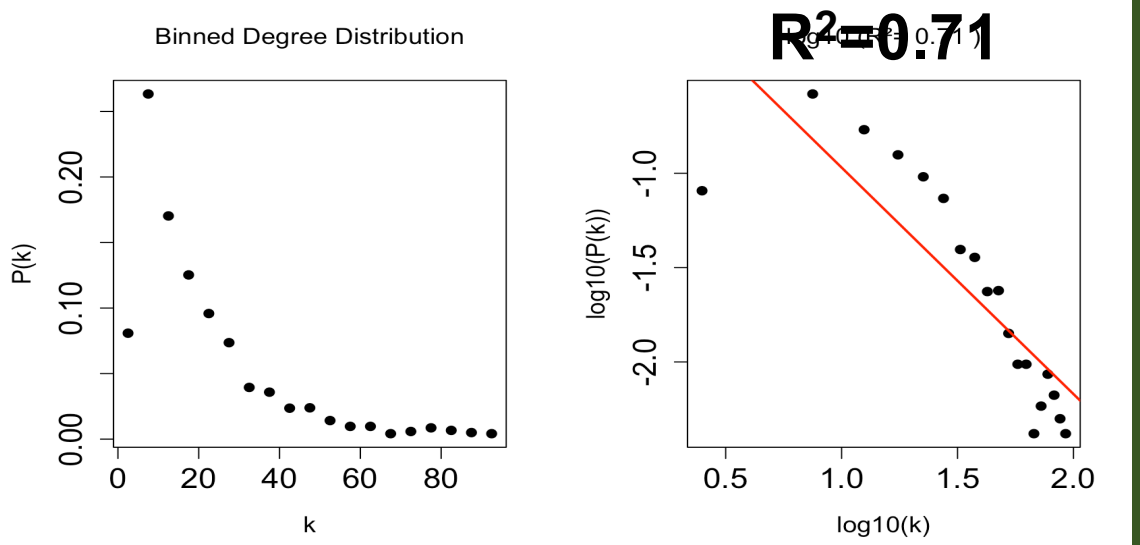


Bin-size=5 R^2=0.76

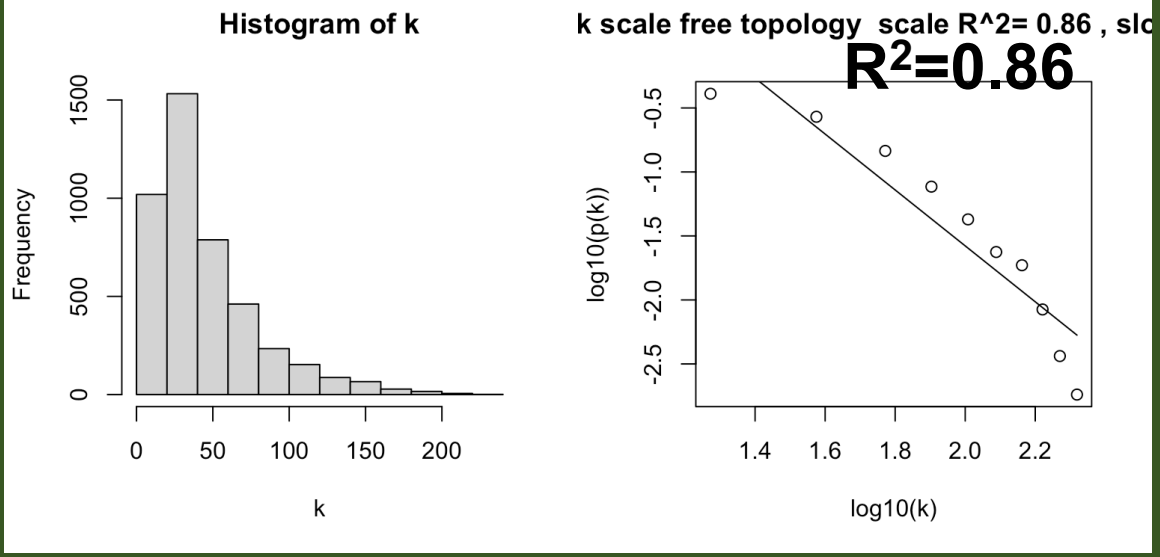


🤔 Both Method and Data matters
But do not vary a lot (I think)

Tutorial Data, My Method (bin-size=5)



WGCNA Package, My Data



Detect Module--overview

- Module **definition**: clusters of densely interconnected genes
- **Measure** interconnectivity: **dTOM**
- Detect **method**: **clustering**

❖ **Topological Overlap Matrix (TOM)**

consider **topological structure** in similarity measurement

Direct connect

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \quad (4)$$

Normalization s.t. $\omega_{ij} \in [0, 1]$

Common Neighbor

where $l_{ij} = \sum_u a_{iu} a_{uj}$, and $k_i = \sum_u a_{iu}$ is the node connectivity, see equation

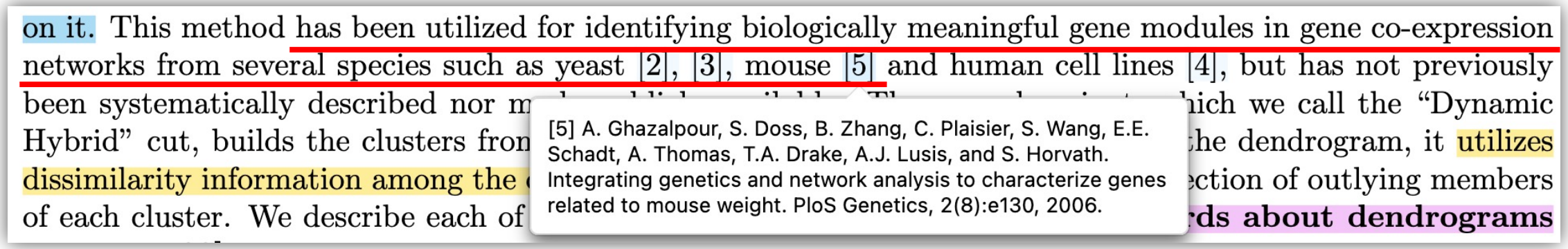
❖ **Dissimilarity TOM (dTOM)** $d_{ij}^{\omega} = 1 - \omega_{ij}.$

Larger value, more similar

Tree Cut After Clustering

❌ Fix tree height
❌ Fix cluster number

"The first variant, called 'Dynamic Tree' cut…"

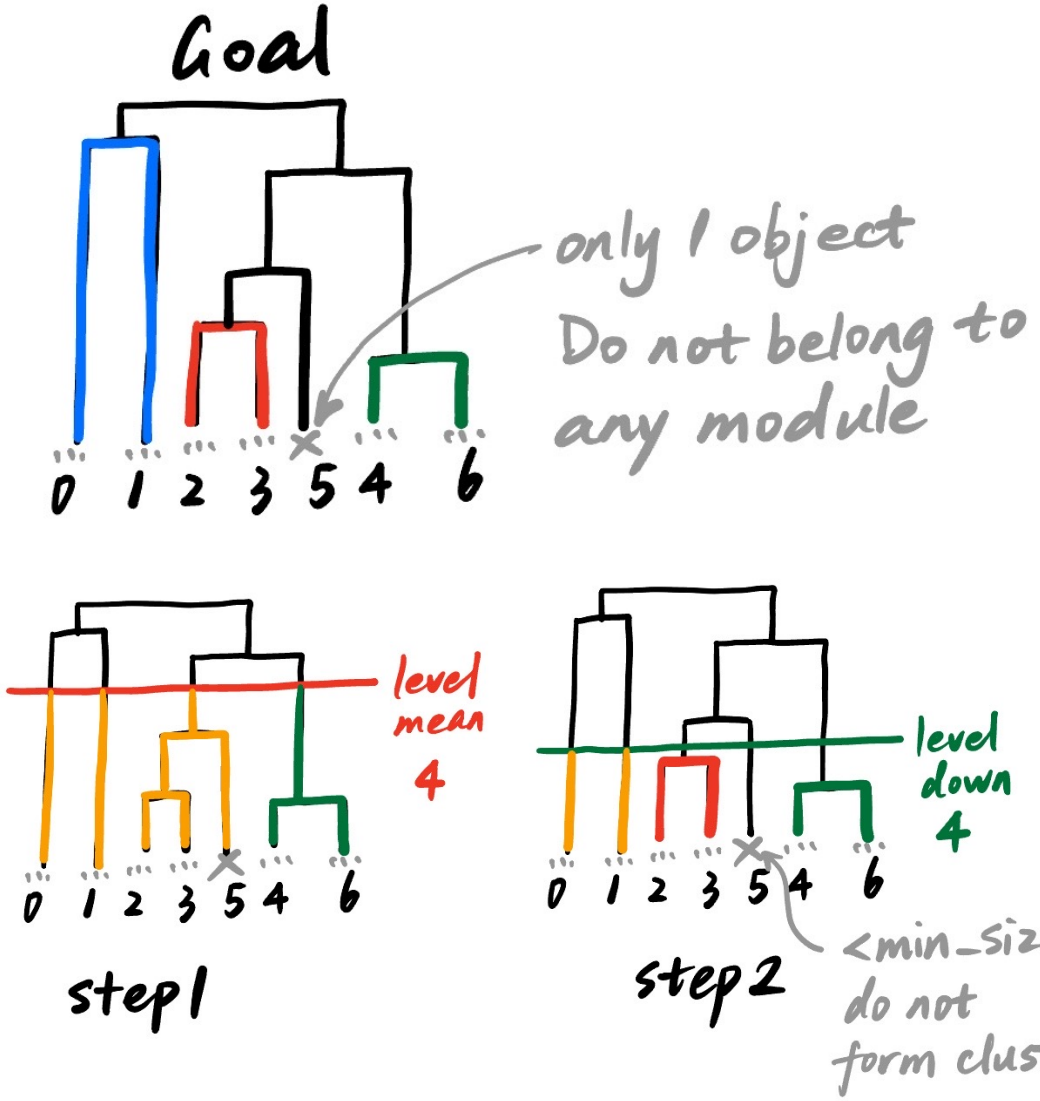

on it. This method has been utilized for identifying biologically meaningful gene modules in gene co-expression networks from several species such as yeast [2], [3], mouse [5] and human cell lines [4], but has not previously been systematically described nor m… hich we call the "Dynamic Hybrid" cut, builds the clusters fron… he dendrogram, it utilizes dissimilarity information among the … ection of outlying members of each cluster. We describe each of … rds about dendrograms

[5] A. Ghazalpour, S. Doss, B. Zhang, C. Plaisier, S. Wang, E.E. Schadt, A. Thomas, T.A. Drake, A.J. Lusis, and S. Horvath. Integrating genetics and network analysis to characterize genes related to mouse weight. PloS Genetics, 2(8):e130, 2006.

(Langfelder et al., 2008)

But existing package is the **second variant** → reproduce by myself…

DynamicTreeCut Principle



Error source 2: TreeCut Method

💡 Use my method to analyze tutorial data to evaluate.

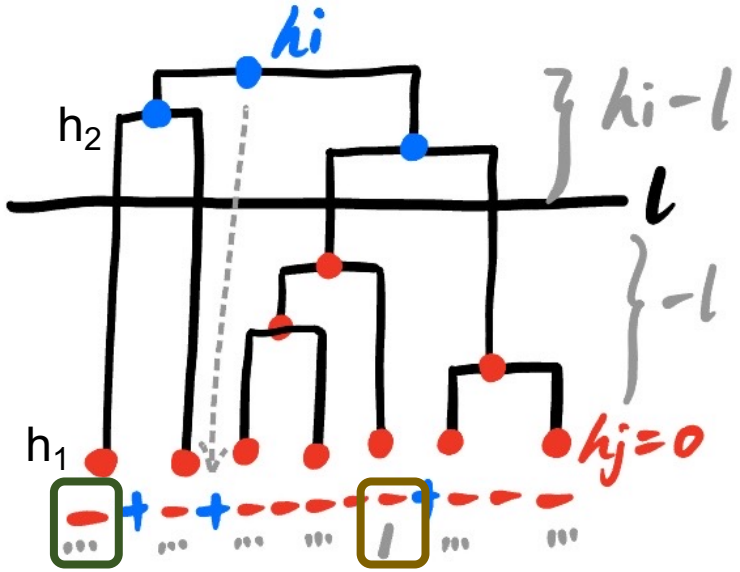
$$l_m = \frac{1}{n}\sum_{i=1}^{n} h_i\,,$$

$$l_u = \frac{1}{2}(l_m + max\{h_1, h_2, ..., h_n\})\,,$$

$$l_d = \frac{1}{2}(l_m + min\{h_1, h_2, ..., h_n\})\,.$$

Reproduce DynamicTreeCut

`def TreeCut(H,l,min_size):` My `min_size` is from the paper: 34 +33



- Nodes here including merging nodes
- # Nodes in the cluster = # **continuous** **minus-**
- If # > min_size → form a cluster

🤔**Order of H = [h_1,h_2,…]?**

- `hclust()` and `linkage()` only do return the merging process
- Order is the **same as dendrogram**!

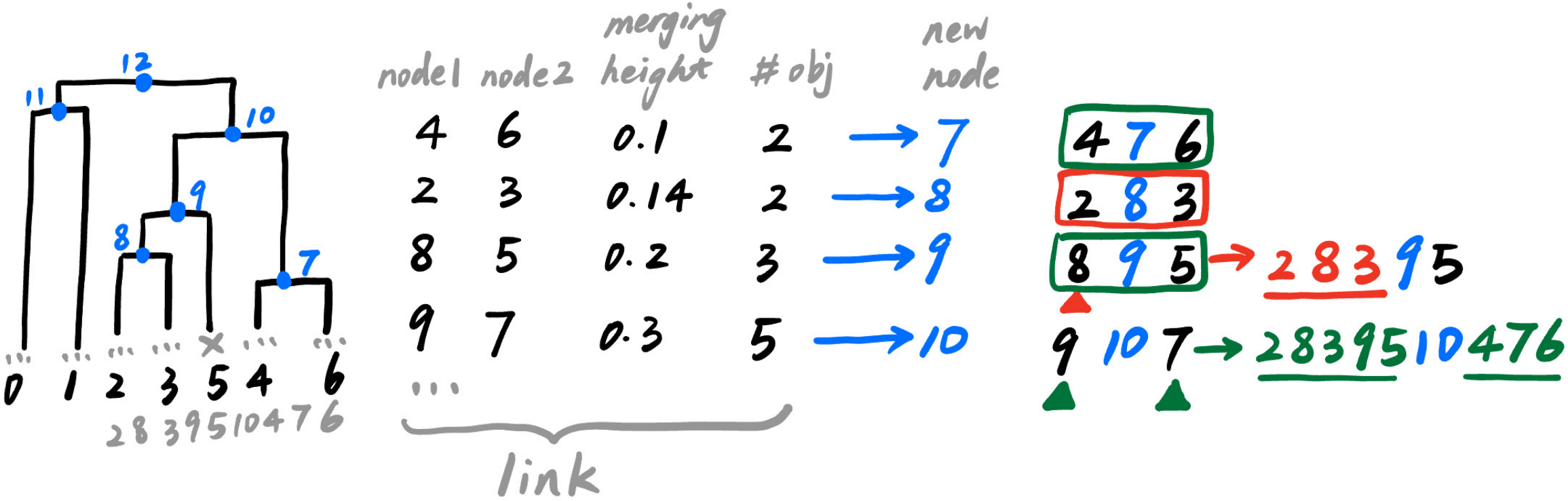
2 You're right to be confused. The only explanation that makes sense is that "dendrogram height sequence" has an unconventional meaning the authors never managed to mention. There's a possible clue in an implementation comment that says it's "the sum of the dissimilarity measure from the root to the node" http://www.bioinformatics.org/cgi-bin/viewvc.cgi/catch/branches/catch-engine/splitter/splitter.c?r1=482&r2=483&view=patch - Gene

Arduous!
Confusions in the paper!

(From stack overflow)

Reproduce DynamicTreeCut

```
def node_order(link):
```

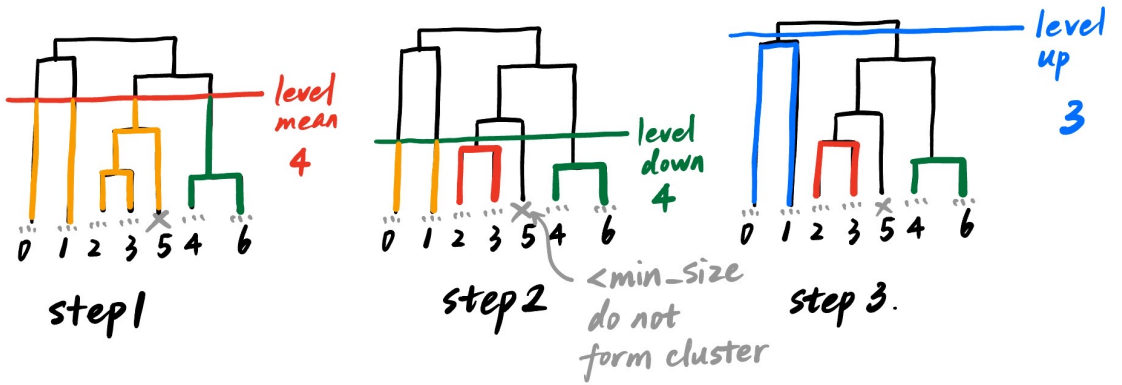


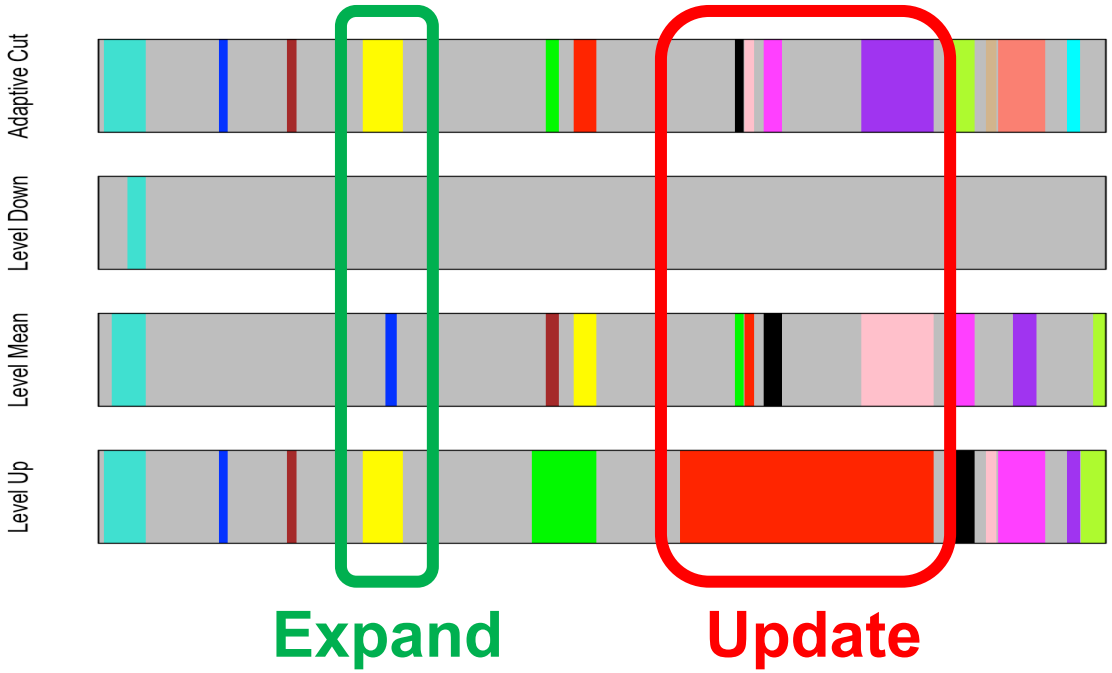
Then **map the height** to the node with link get `height_order`

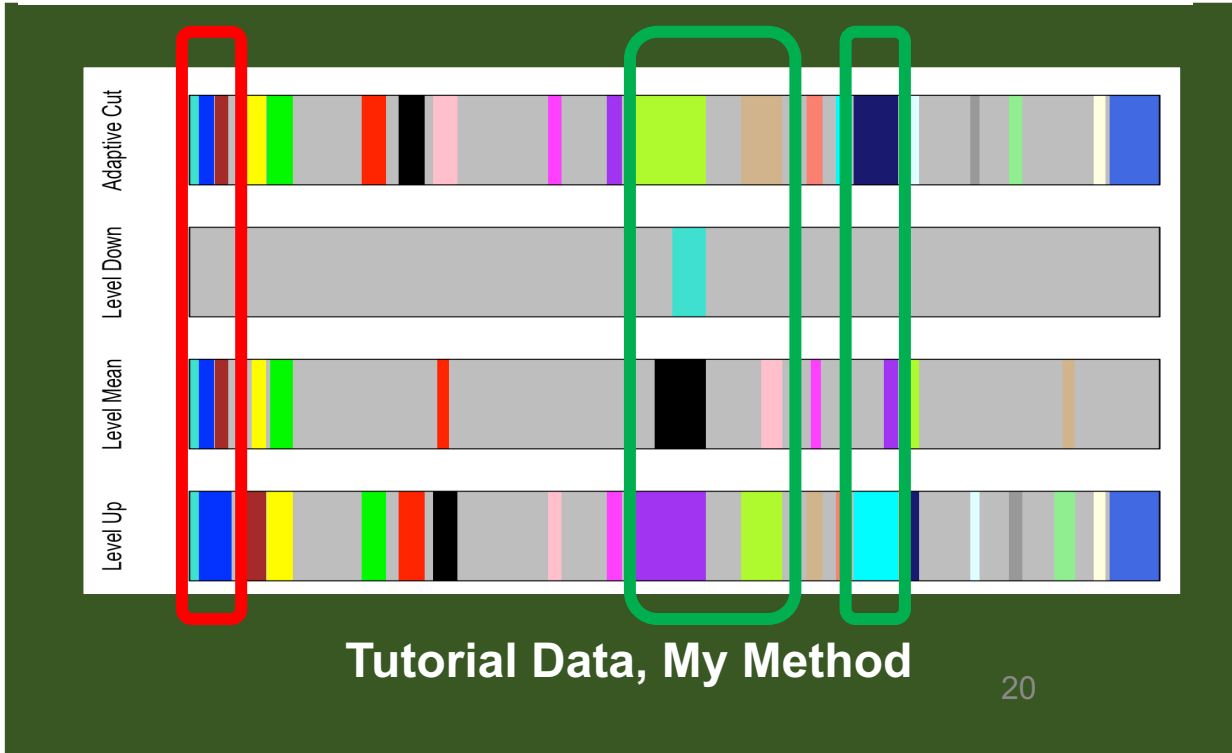
Reproduce DynamicTreeCut

```
Def AdaptiveTreeCut(H):
```

- If **≥ 2** sub-clusters in the lower cut level result, **update**
- Or else, **expand** the cluster range

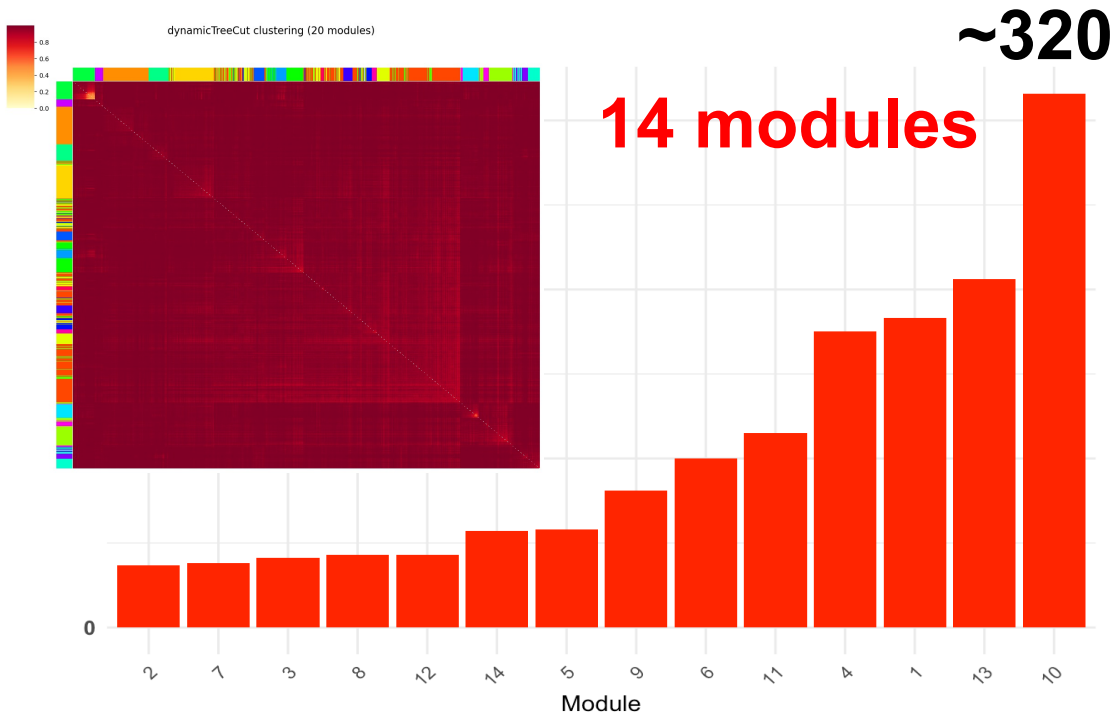




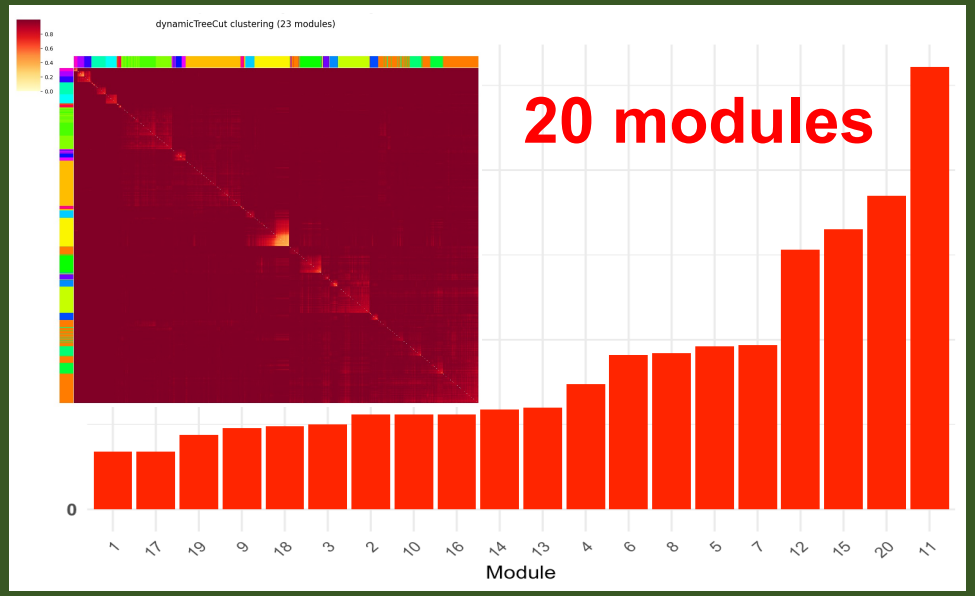


Number of Clusters

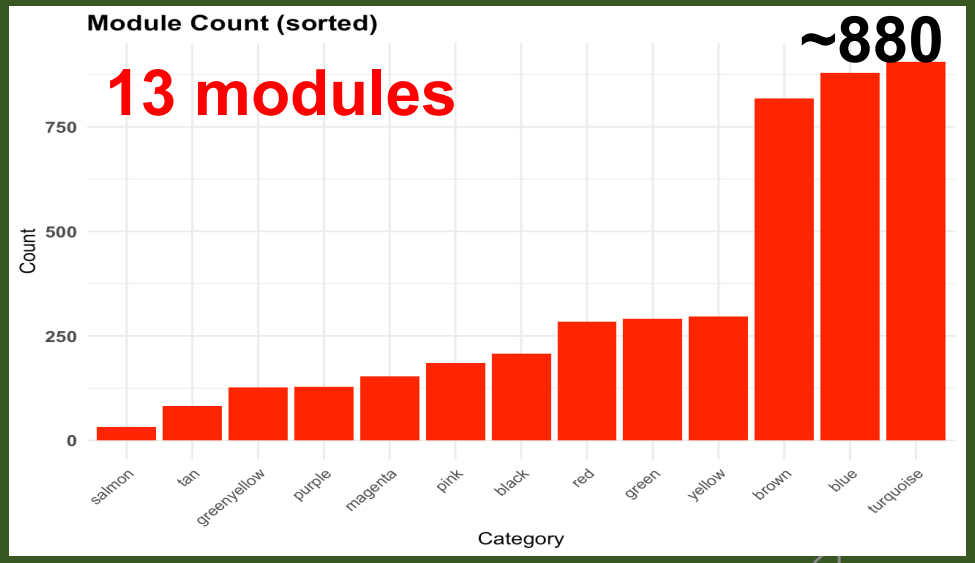
dTOM plot is better in determining the number of



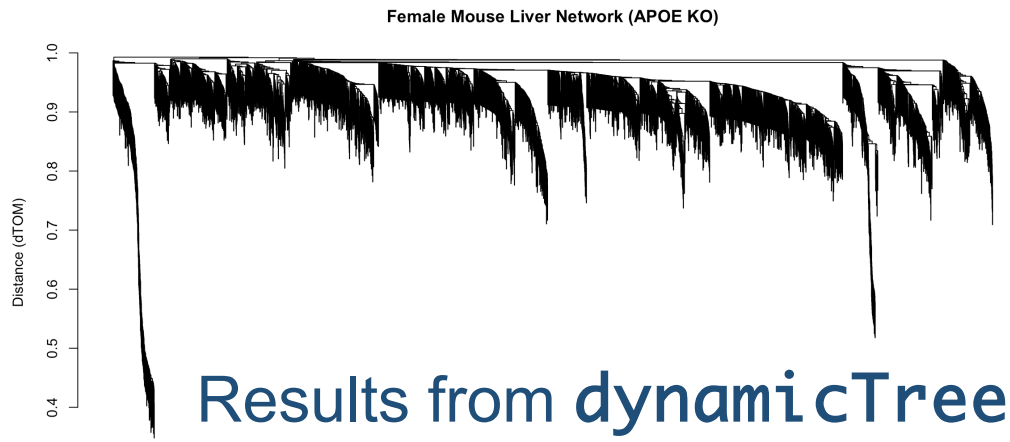
Tutorial Data, My Method



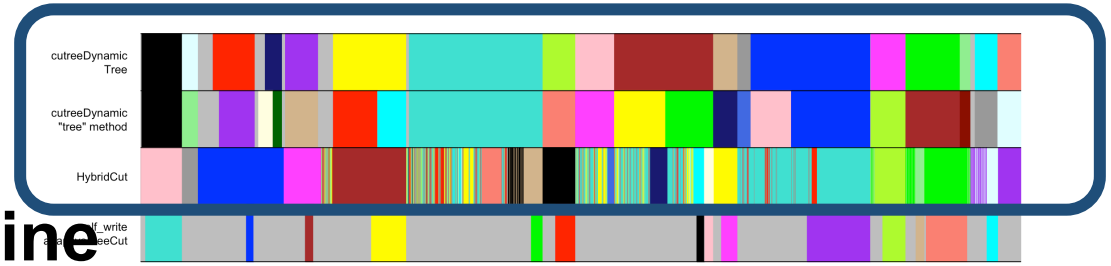
WGCNA Package, My Data



DynamicTreeCut



Results from `dynamicTreeCut()`

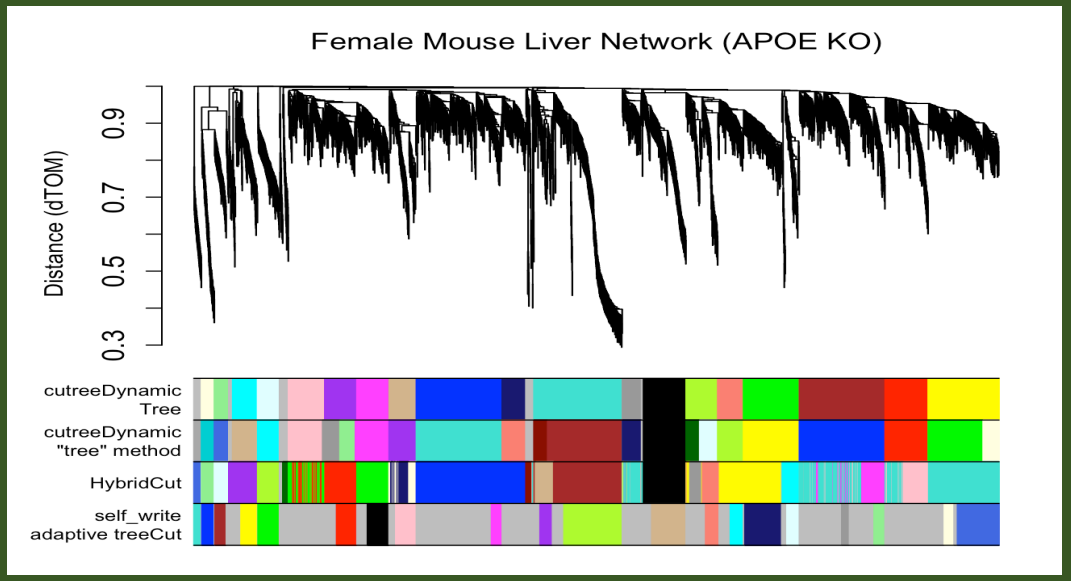


mine

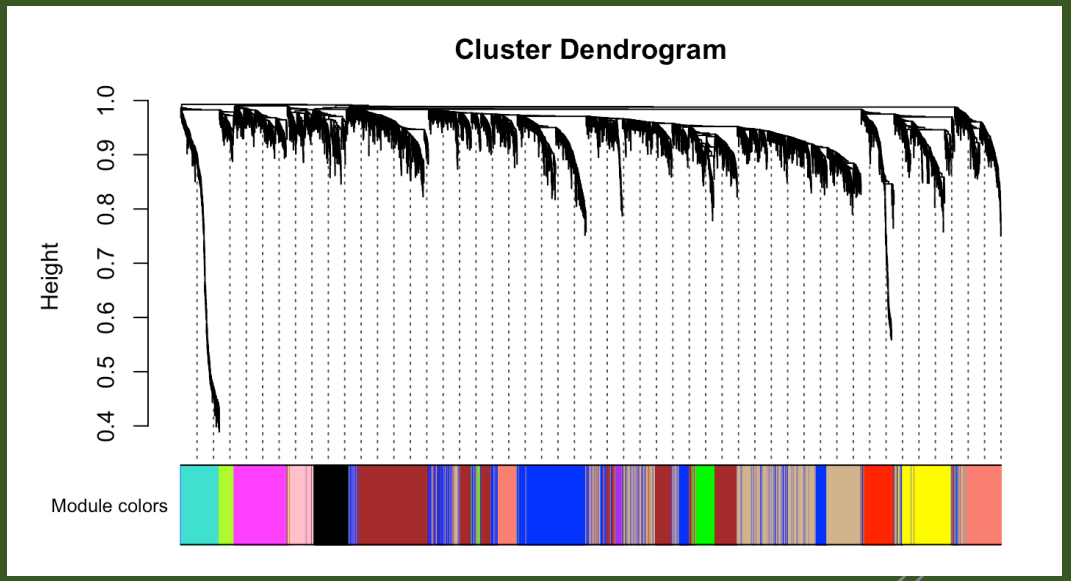
grey means unassigned

🤔 my results are much sparser because do not do assignment (variant 2) —hard to evaluate…

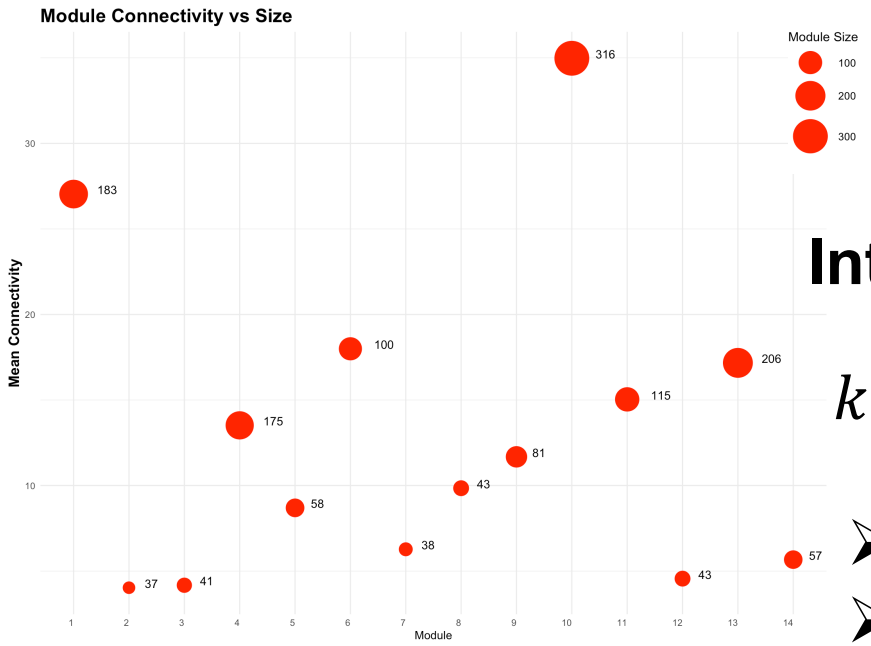
Tutorial Data, My Method

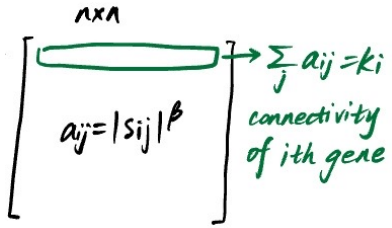


WGCNA Package, My Data



Cluster Evalutation





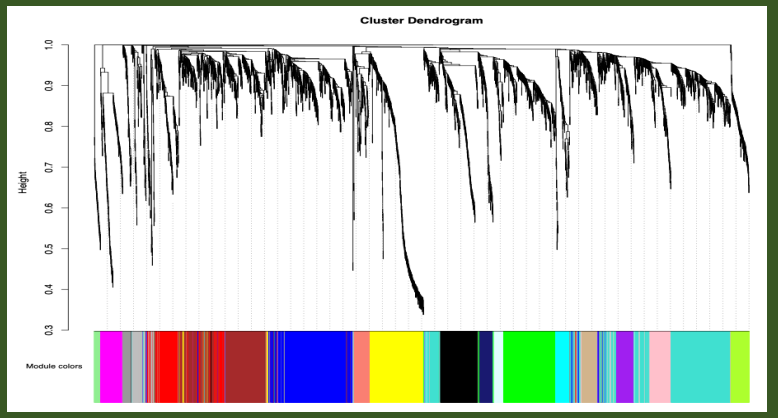
Intramodular Connectivity

$$k_{in} = \frac{\sum_i k_i}{\#\ genes\ in\ module}$$

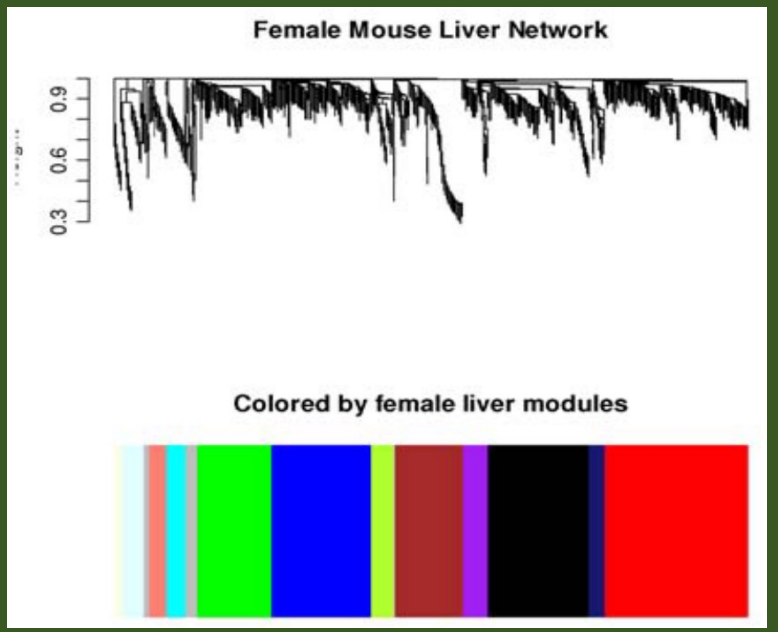
- **Larger** module size
- Generally **higher** k_{in}

🤔 Even tutorial data processed with WGCNA package **do not fully align** to the paper result! Maybe should focus on the **biological interpretation** of the result to **evaluate**…

Tutorial data and WGCNA package

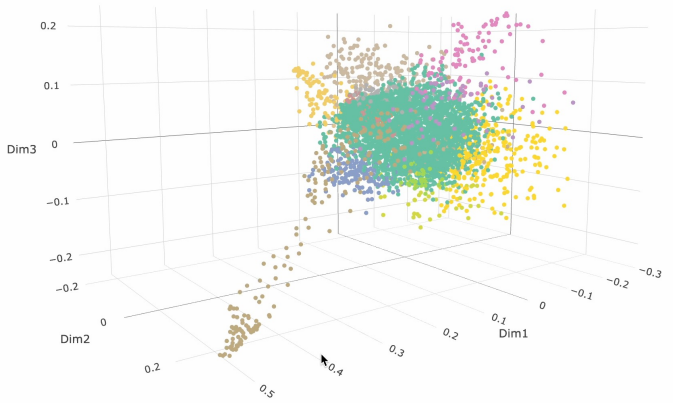


Paper Figure

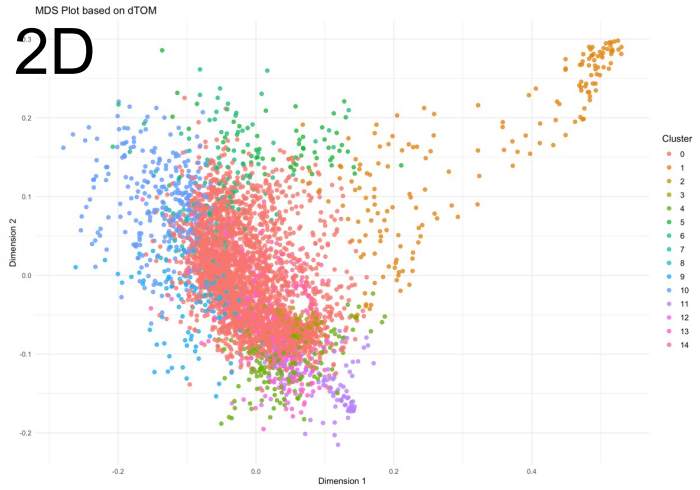


Evaluation--MDS Plot

3D

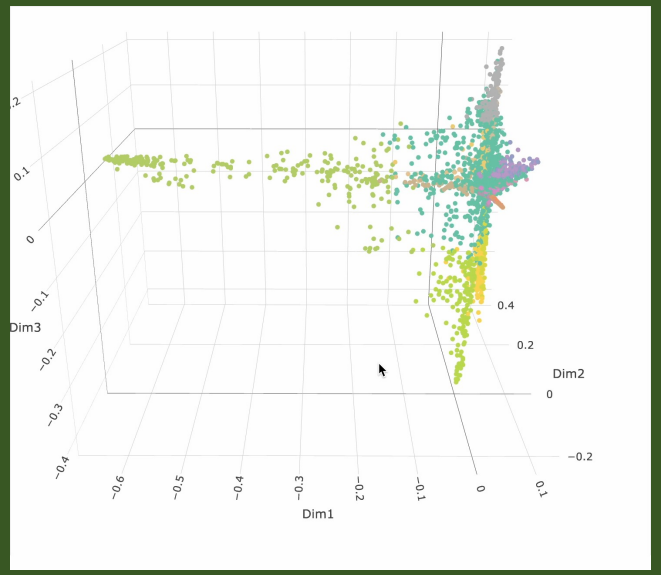


2D

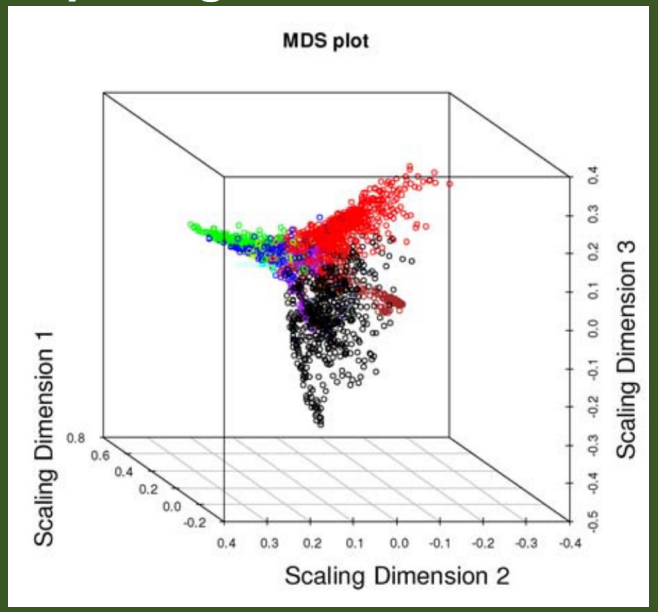


🤔 Similar **shape** → my **data** may work **unassigned** genes in the **center** → clustering performs well

Tutorial Data, My Method



Paper Figure



Module Significance (MS)

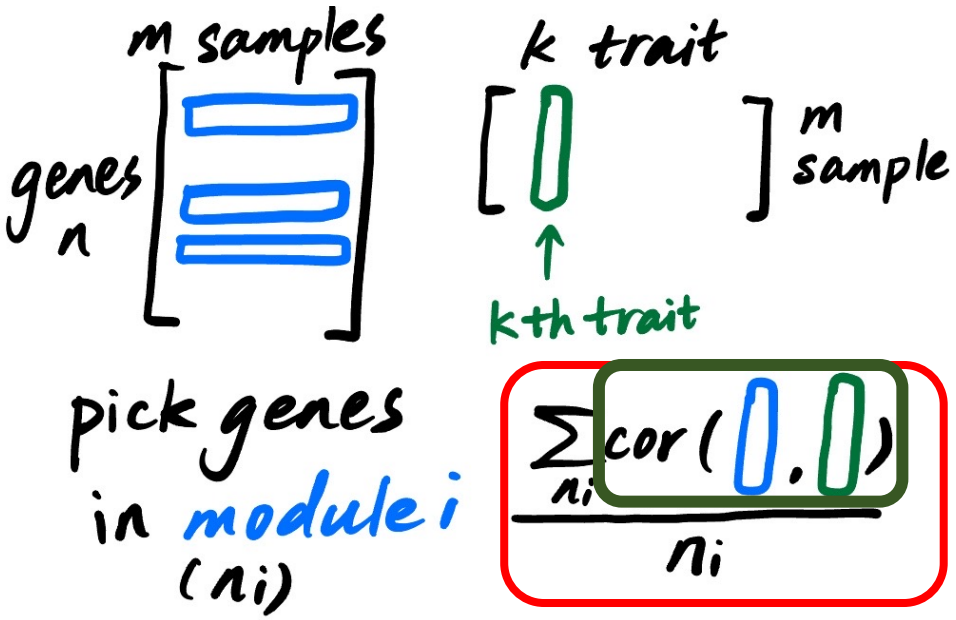
❖ **Gene significance (GSi)**

the **relationship strength** between ith gene and kth trait

$$\mathbf{GS_i = cor(x_i, T_k)}$$

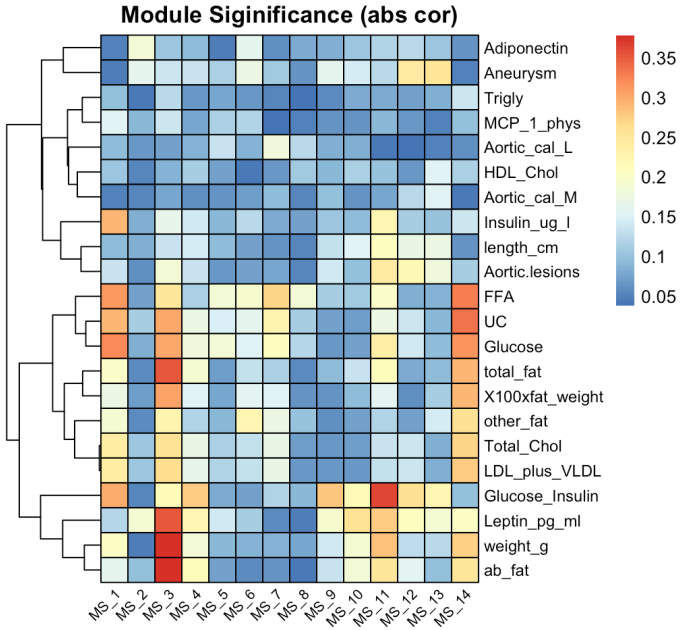
❖ **Module significance (MS)**

$$MS = \sum GS_i \text{ , } GS_i \text{ of module}$$

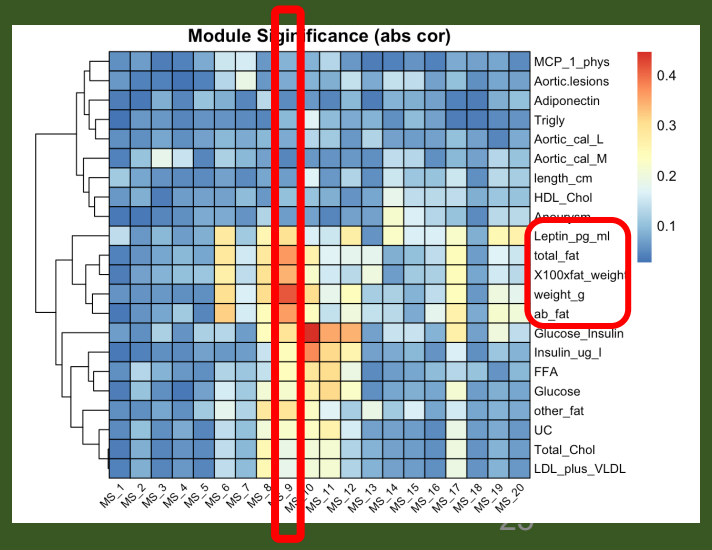


GS

MS: Value of one block

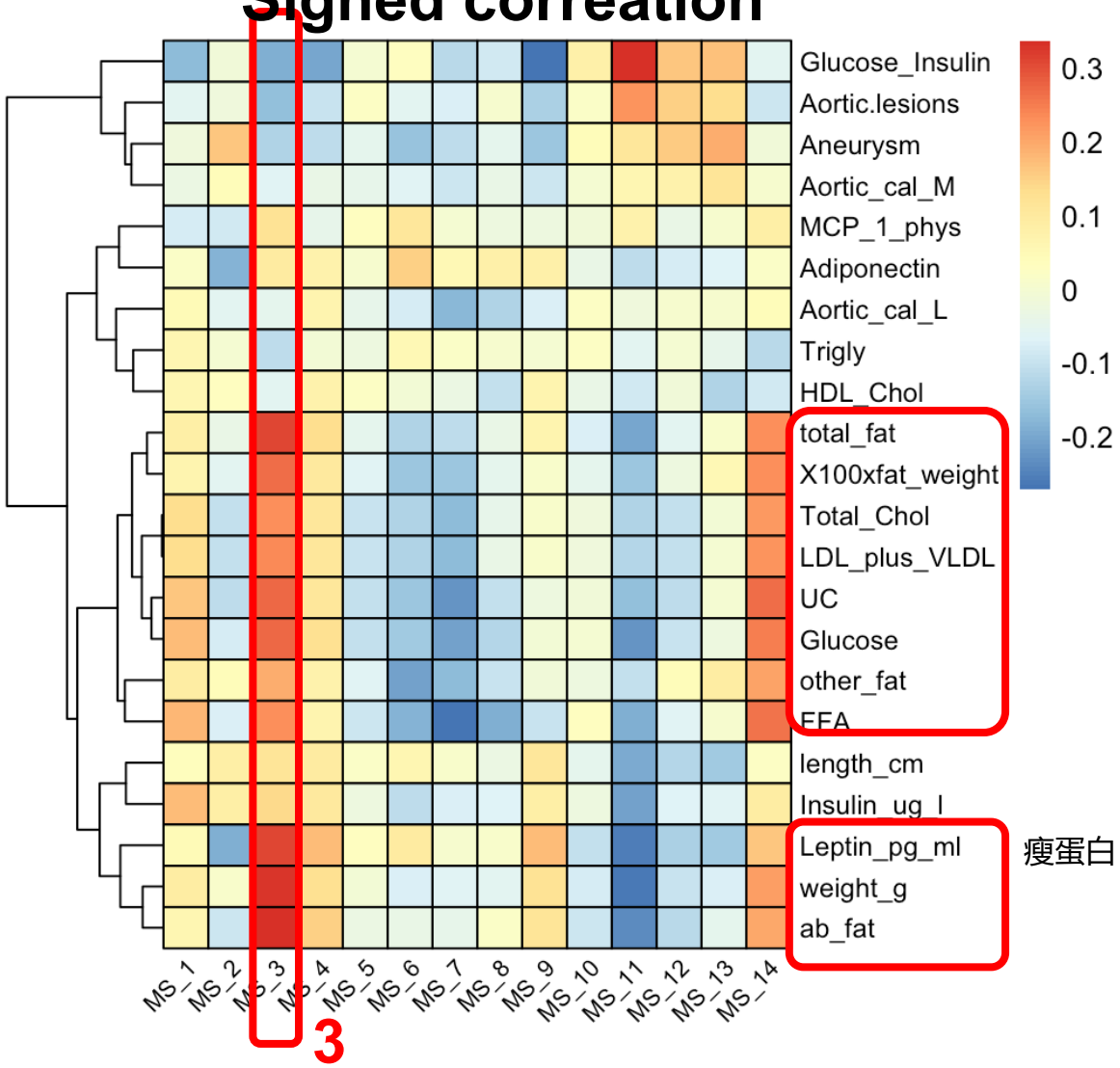


Tutorial Data, My Method
Absolute correlation result

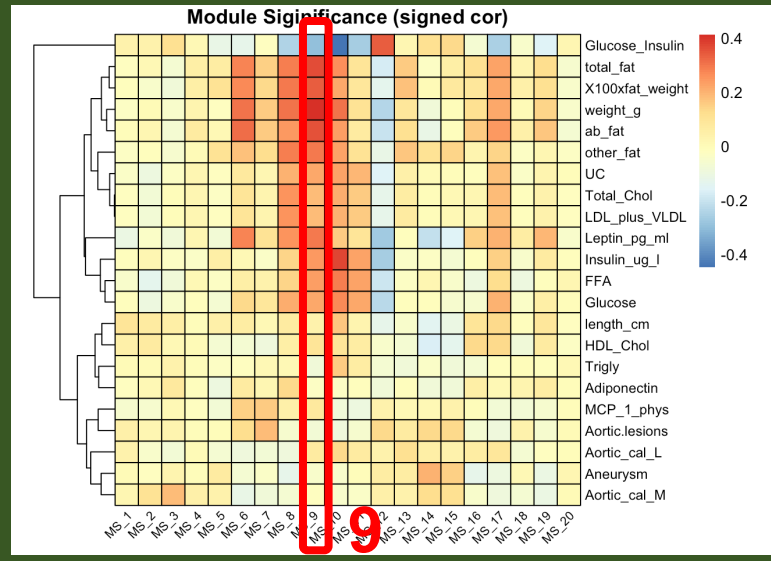


Module Significance (MS)

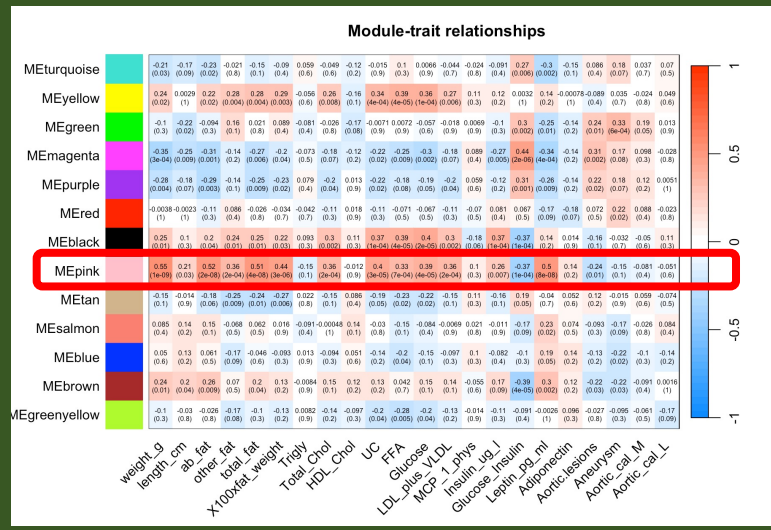
Signed correation



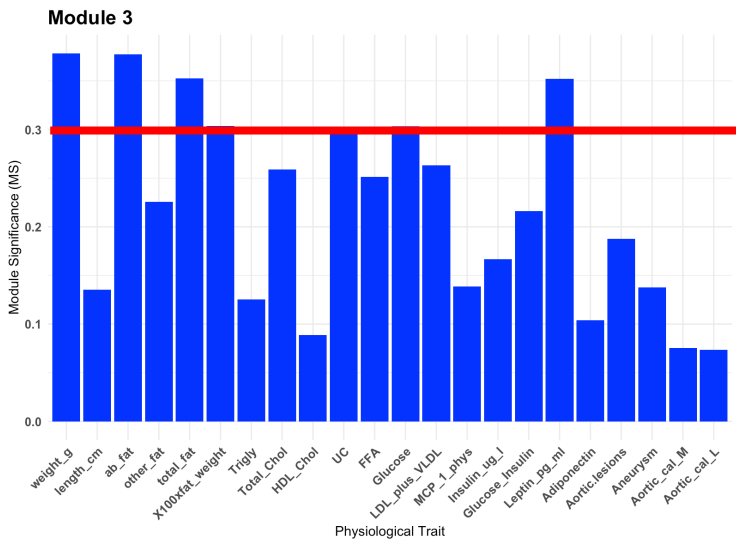
Tutorial Data, My Method



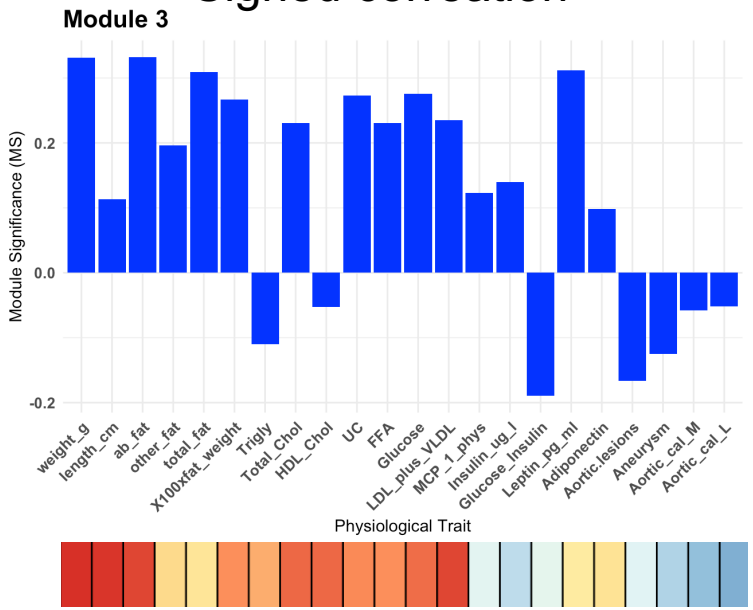
WGCNA Package, My Data



Weight-related Module



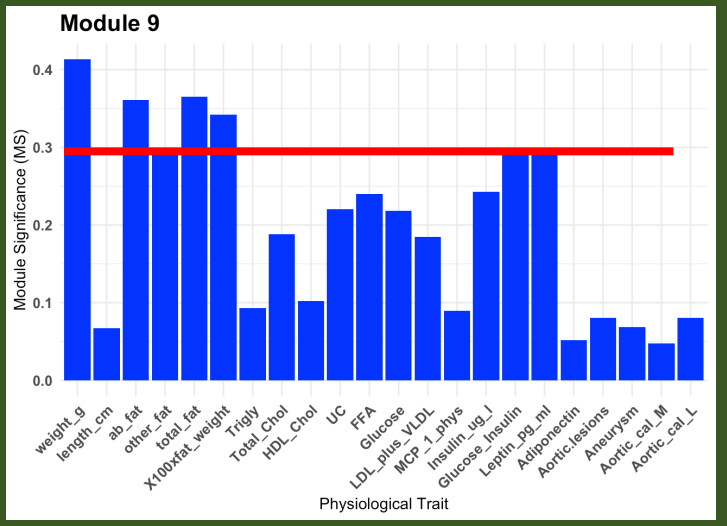
Signed correation



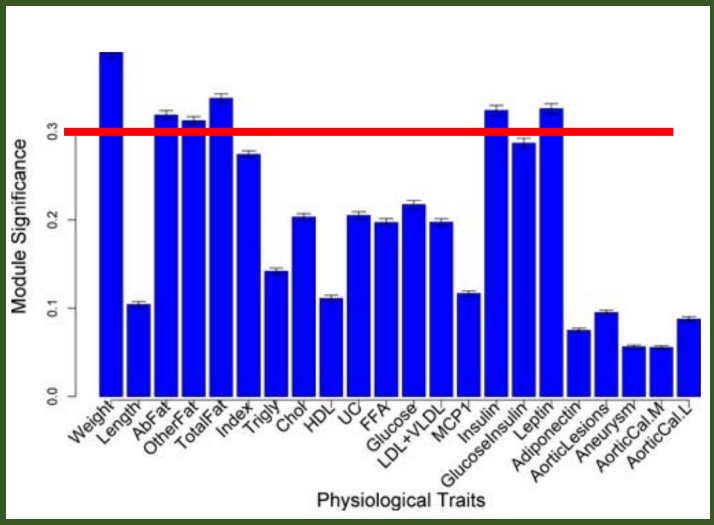
🤔 my **compact module** show **higher correlation** with the traits

Mostly the **same trend** as the paper result → my **method** seems to work!

Tutorial Data, My Method



Paper Result



Wrap Up

- **Done**
  - Reproduce: **data processing**, **network construction**, **module detection** (with “my” clustering method), and **module significance evaluation**.
  - From visualization, both **my data and method seem to work**. They can provide **biological insights** we want. **Do downstream analyses** are required to further validate.
- **To be improved…**
  - Should **simplify codes**
  - Should deepen **understanding** network property and similarity measurement
- **Overall, thanks this project!**
  - I've become **stronger**—both mentally and technically
  - I just learned that analysis can be done **without explicit control group**!

Reference List

1. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, et al. (2006) Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. PLOS Genetics 2(8): e130. https://doi.org/10.1371/journal.pgen.0020130

2. Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

3. Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics*, *24*(5), 719–720. https://doi.org/10.1093/bioinformatics/btm563

4. Zhang, B., & Horvath, S. (2005). A general framework for weighted gene Co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128