WGCNA and LASSO method in the identification of key genes for phenotypes related to salinity tolerance in rice plants (Oryza Sativa)

Camila Riccio Rengifo

Doctorate student in Engineering
Pontificia Universidad Javeriana - Cali

September 13th 2019

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

WGCNA

WGCNA (Weighted gene co-expression network analysis)

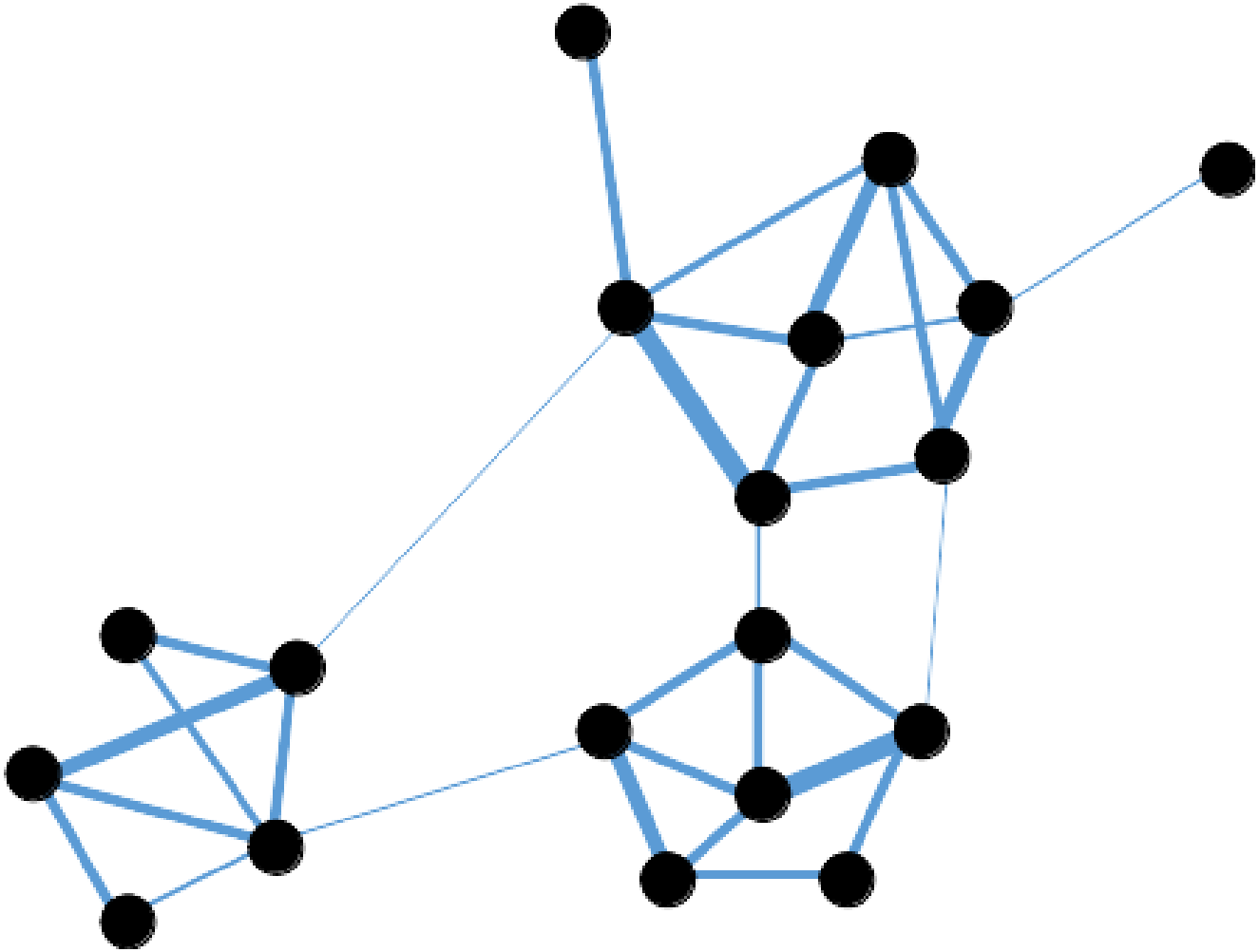
Is a widely used data mining method especially for studying biological networks based on pairwise correlations between variables.

Data for WGCNA:

- **Gene expression data** (microarray or RNA-seq)
- **Clinical/phenotypical traits** from the same individuals

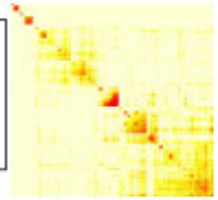
Pontificia Universidad JAVERIANA Cali

Weighted gene co-expression network

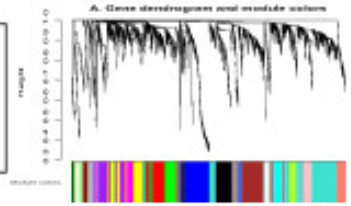


WGCNA methodology

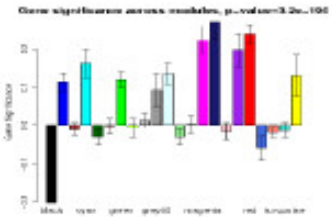
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



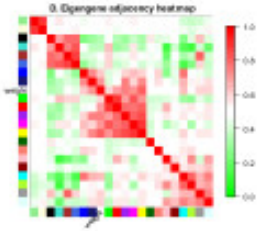
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



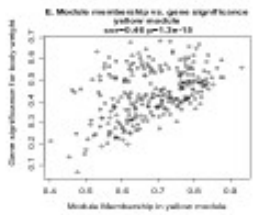
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



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



Pontificia Universidad JAVERIANA Cali

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

Gene Expression Data

Let d_{ij} be the expression level of gene i in sample j, RNA-sequencing data present the following structure:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{np} \end{pmatrix}$$

Experiment data: *Oryza sativa* gene expression levels

- Control data: $C = [c_{ij}]_{n \times p}$.
- Salt stress treatment data: $T = [t_{ij}]_{n \times p}$.

The expression changes are measure through the differential expression matrix

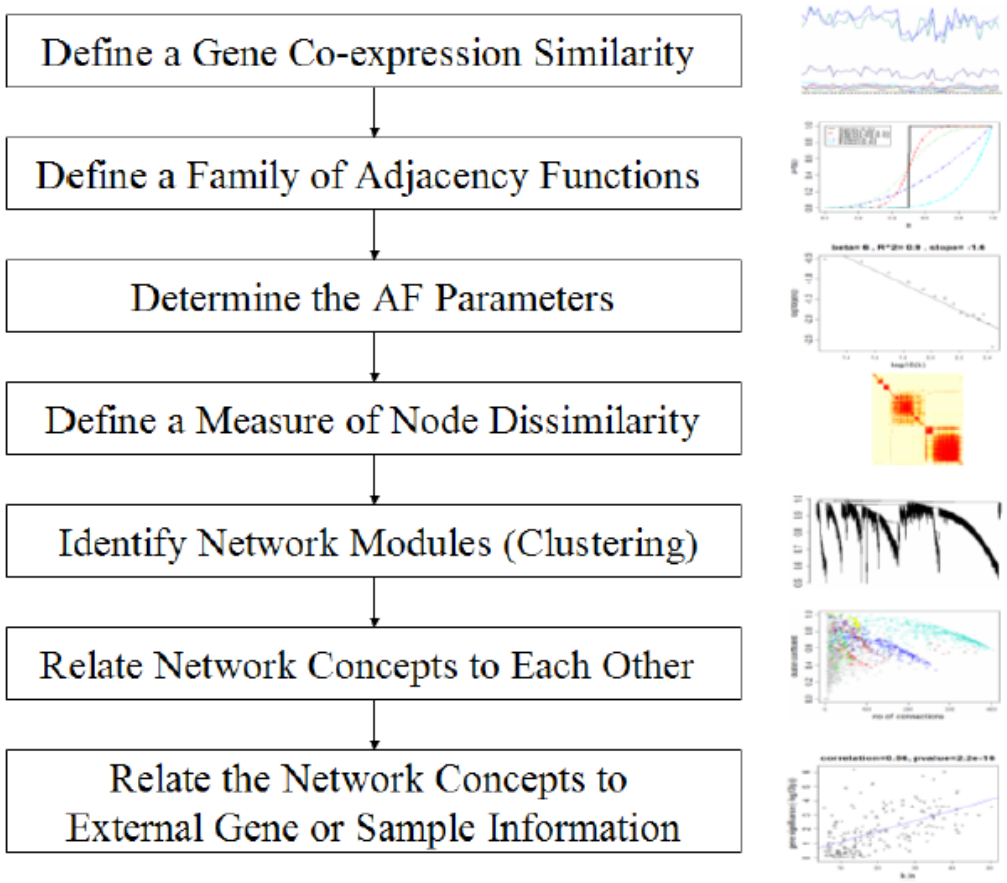
$$E = [e_{ij}]_{n \times p}, \qquad e_{ij} = \log_2(t_{ij}/c_{ij}).$$

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

Gene co-expression network construction



Pontificia Universidad JAVERIANA Cali

Gene Co-expression Similarity

Similarity

The similarity matrix $S = [s_{ij}]_{n \times n}$ measures the level of concordance between gene expression profiles across the experiments.

$$s_{ij} = |cor(g_i, g_j)|, \qquad s_{ij} \in [0, 1]$$

Differential Expression matrix

	GSM2596381	GSM2596385	GSM2596389	GSM2596393	...
13102.t05279	1.0979195	0.6549420	0.7451395	1.3129533	
13103.t00726	0.7776076	0.9593580	1.0000000	0.8845228	
13106.t04257	0.9696264	0.5077946	0.6903155	1.3107875	
13102.t03177	0.9781642	0.7330076	0.7728225	1.1570437	
13105.t03352	1.1710778	0.6765930	0.7004397	1.3006595	
13109.t02369	1.1069152	1.7369656	1.0000000	1.3692338	
⋮					

Similarity matrix

	13102.t05279	13103.t00726	13106.t04257	13102.t03177	...
13102.t05279	1.0000000	0.27969713	0.6693985	0.8126950	
13103.t00726	0.2796971	1.00000000	0.2424204	0.1857460	
13106.t04257	0.6693985	0.24242041	1.0000000	0.5611319	
13102.t03177	0.8126950	0.18574604	0.5611319	1.0000000	
13105.t03352	0.8489128	0.17629182	0.7025265	0.7878663	
13109.t02369	0.1698732	0.01098545	0.2347910	0.2485275	
⋮					

Number of genes: $n = 5142$

Number of samples: $p = 91$

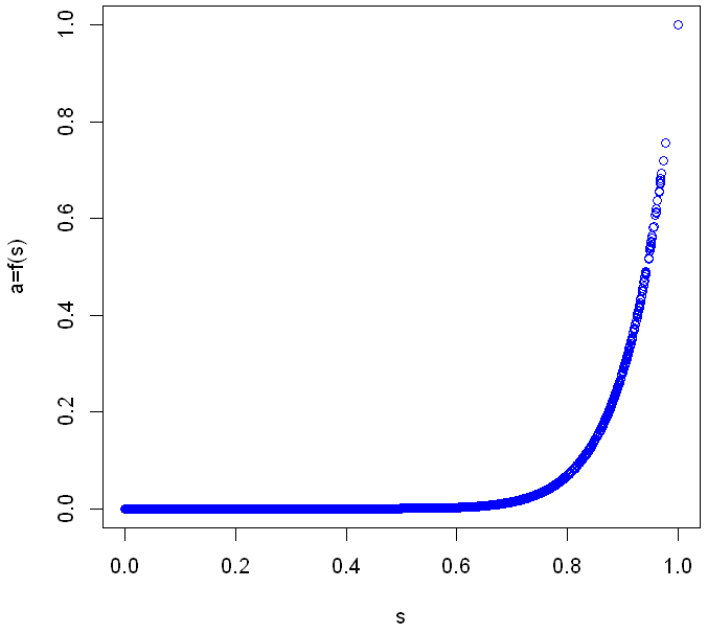
Pontificia Universidad JAVERIANA Cali

Adjacency Function

Power adjacency function

The adjacency matrix $A = [a_{ij}]_{n \times n}$ encodes the connection strength between each pair of nodes (genes).

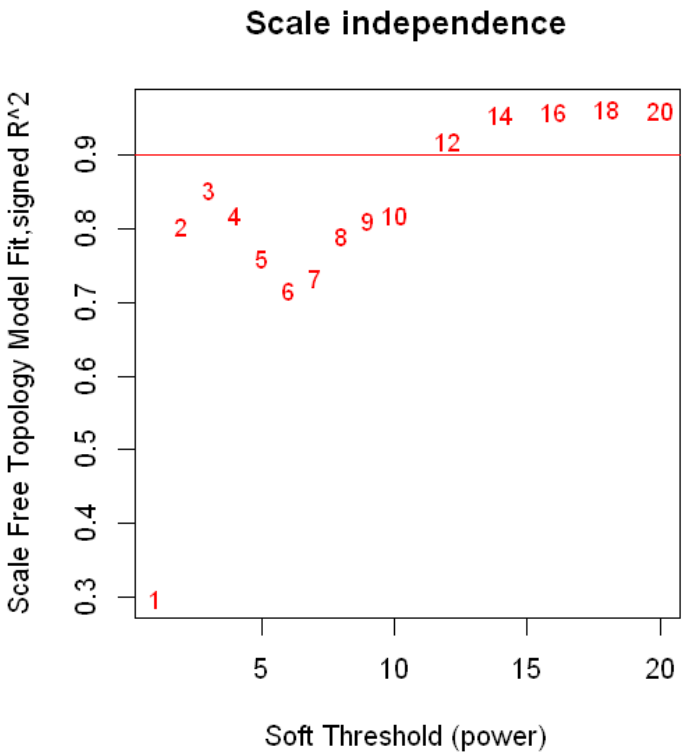
$$a_{ij} = power(s_{ij}, \beta) = s_{ij}^{\beta}, \qquad \beta \geq 1., \qquad a_{ij} \in [0, 1]$$

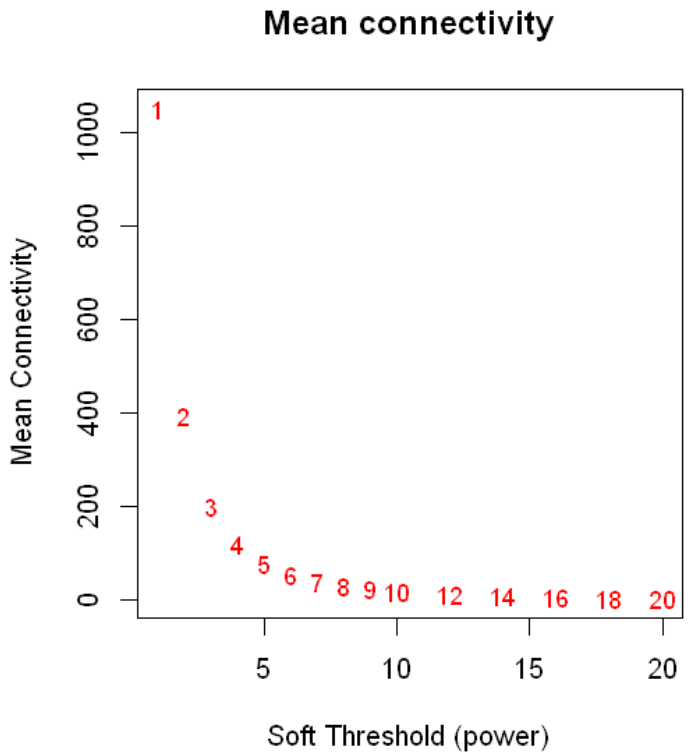


Determining the Parameter of the Adjacency Function

Scale-free Topology Criterion

Use the first parameter value that lead to a network satisfying scale-free topology at least approximately, e.g. R^2 between $\log(p(k))$ and $\log(k)$, greater than 0.9.





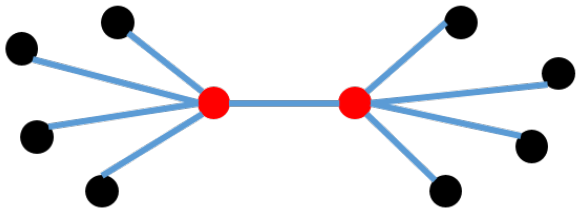
Measure of Node Dissimilarity

TOM: Topological Overlap Matrix

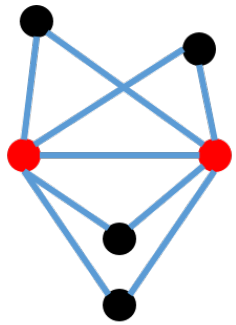
The topological overlap matrix $\Omega = [\omega_{ij}]$ measures direct connection + shared neighbours:

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

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



No shared neighbours: low TOM



Many shared neighbours: high TOM

The topological overlap based dissimilarity measure is defined by $d_{ij}^{\omega} = 1 - \omega_{ij}$.

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

Identifying Gene Modules

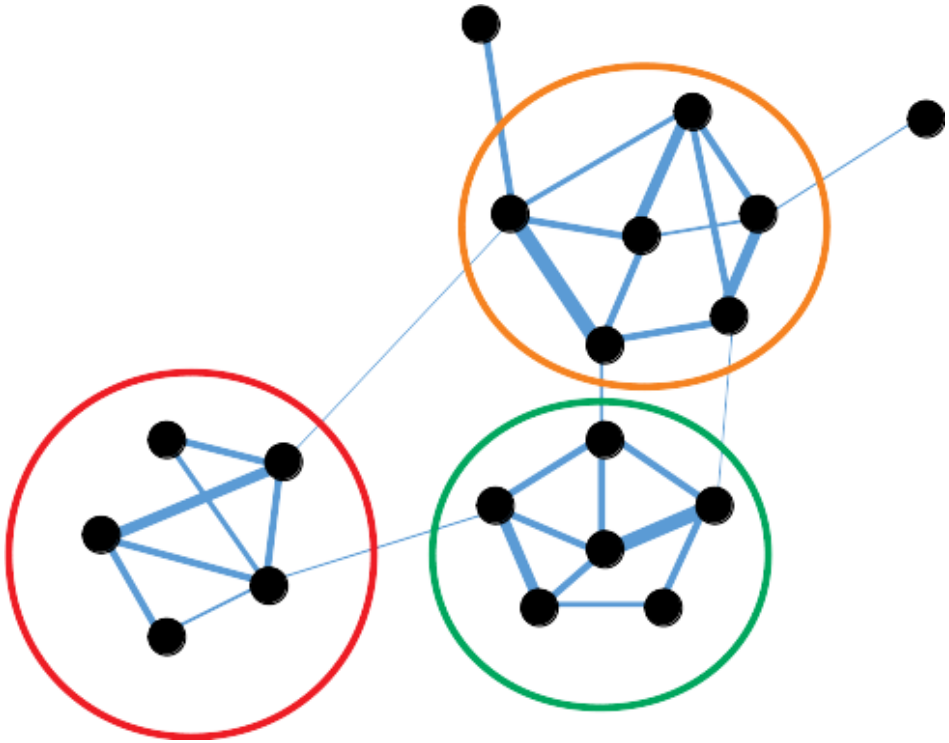
Modules

Modules are groups of nodes with high topological overlap. Intuitively speaking, modules are groups of genes whose expression profiles are highly correlated across the samples.

Identifying Gene Modules

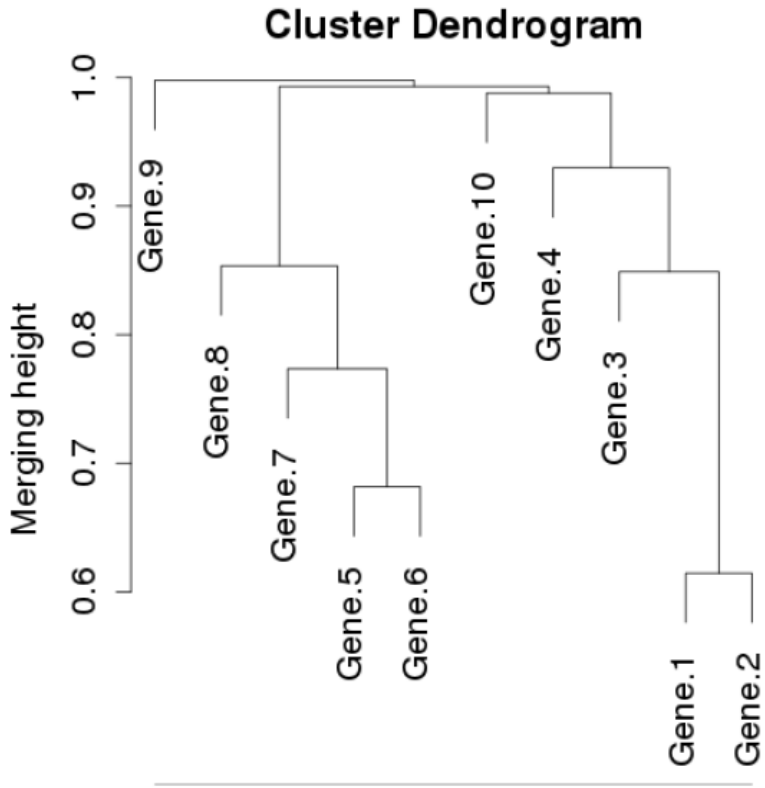
Modules

Modules are groups of nodes with high topological overlap. Intuitively speaking, modules are groups of genes whose expression profiles are highly correlated across the samples.



Hierarchical clustering steps

Step one: Construct a hierarchical clustering tree (dendogram) that provides information on how objects are iteratively merged together.

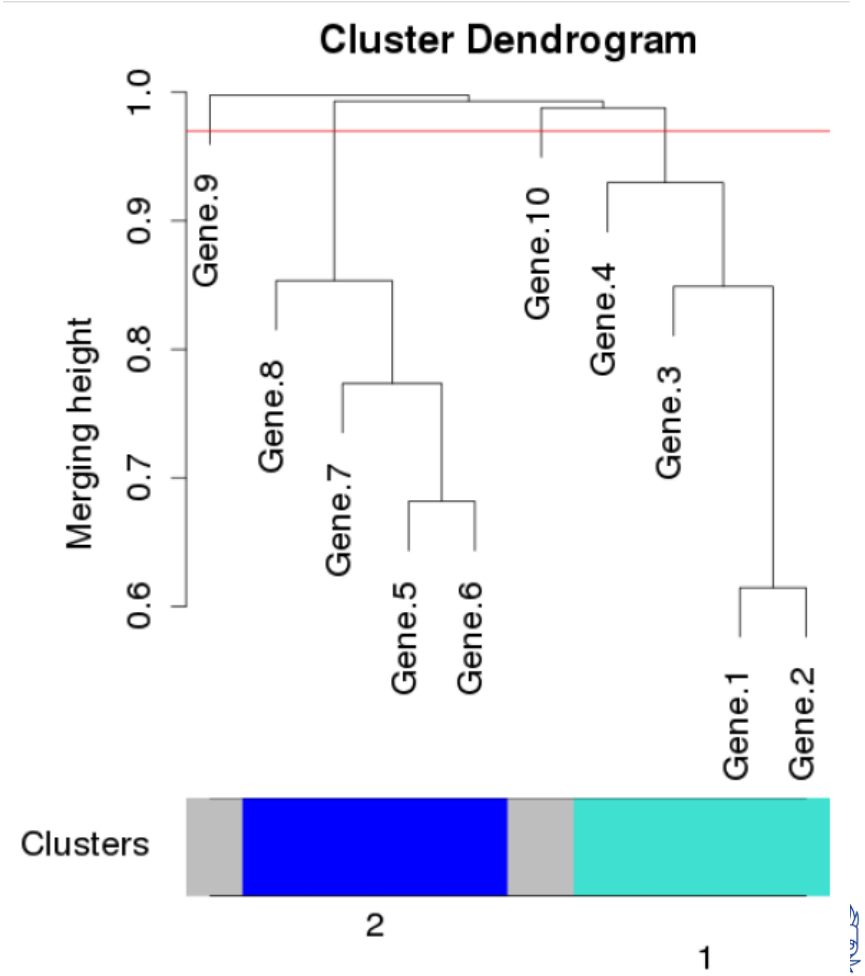


Pontificia Universidad JAVERIANA Cali

Hierarchical clustering steps

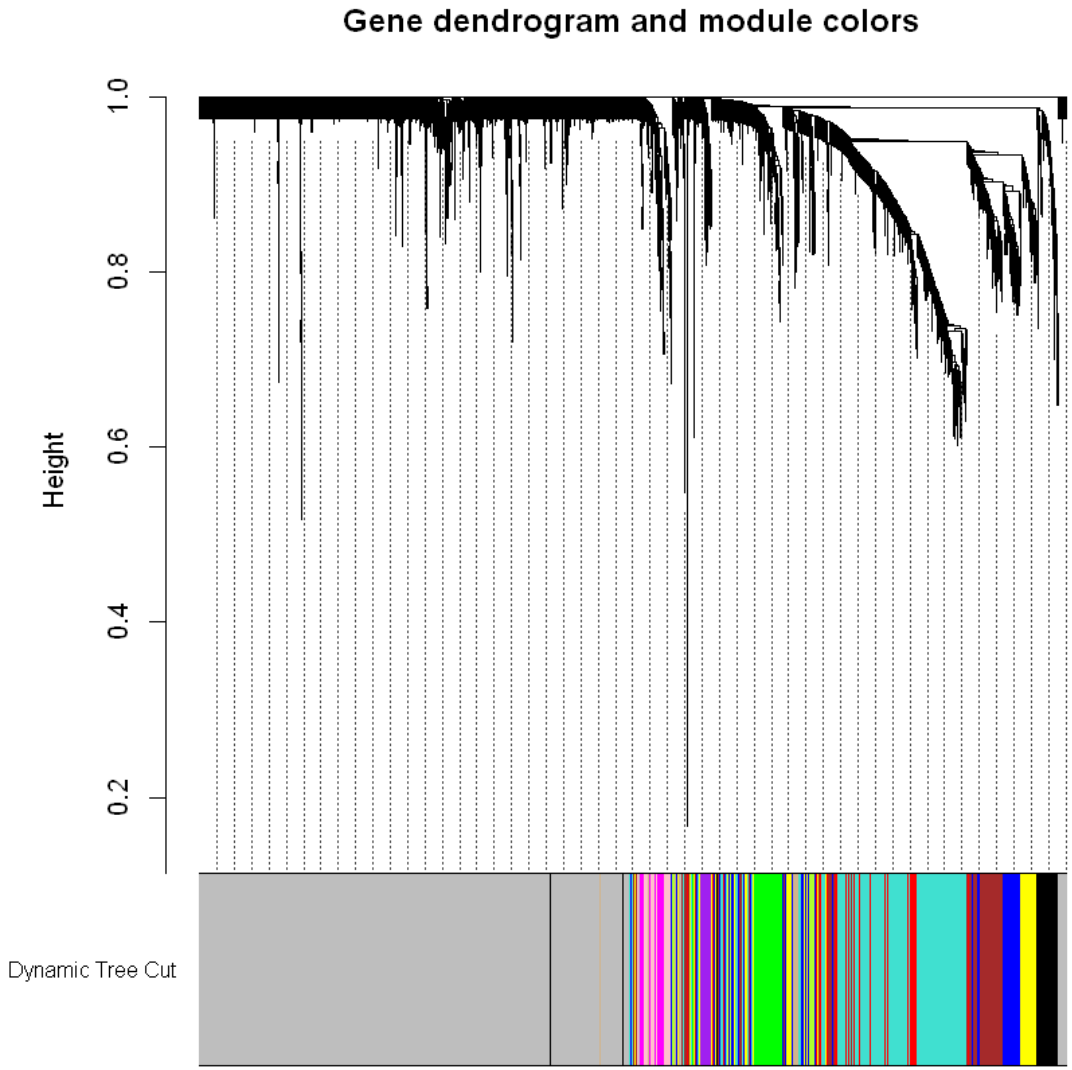
Step one: Construct a hierarchical clustering tree (dendogram) that provides information on how objects are iteratively merged together.

Step two: Identify branches that correspond to clusters. Label branches by numbers and/or colors.

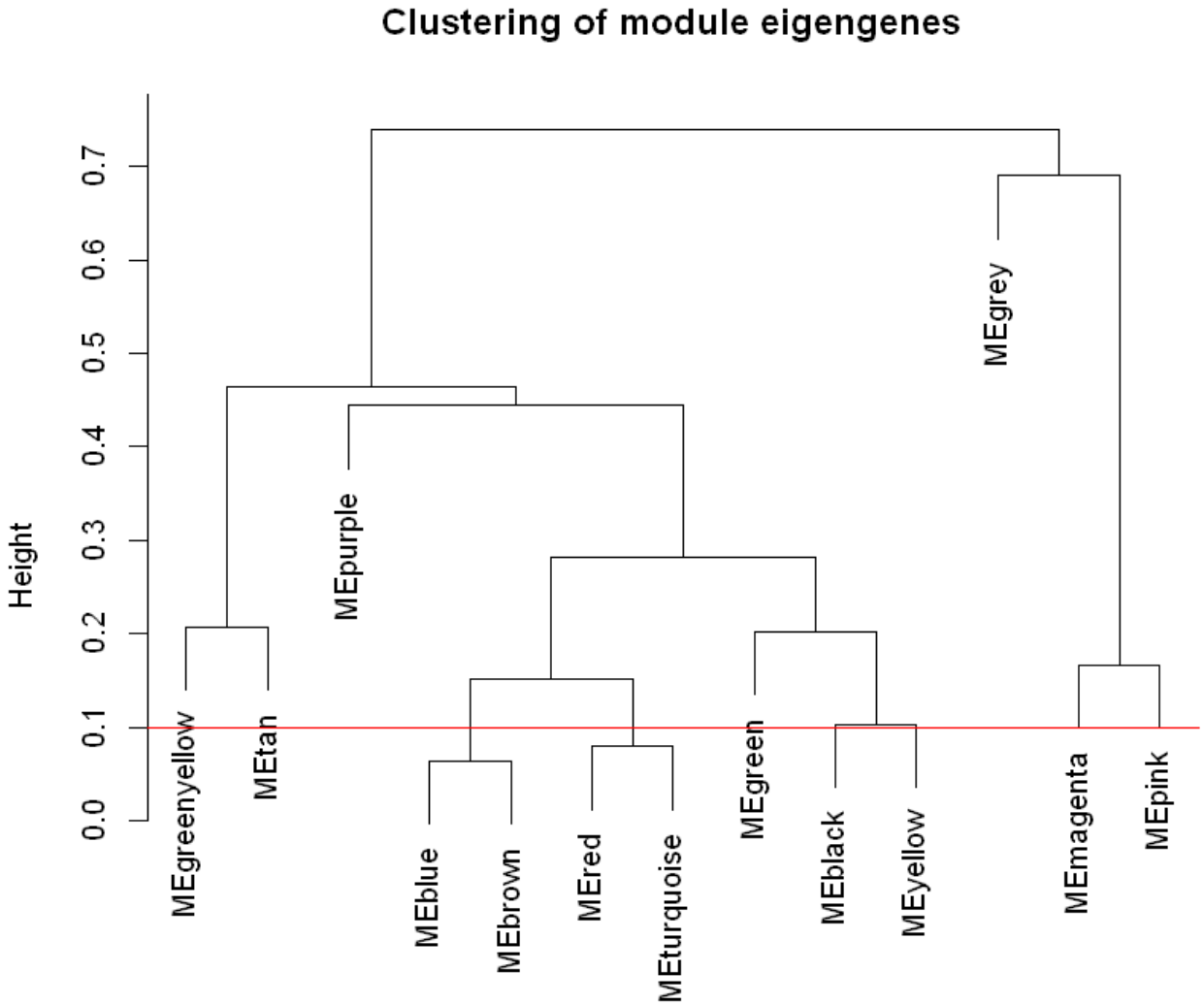


Pontificia Universidad JAVERIANA Cali

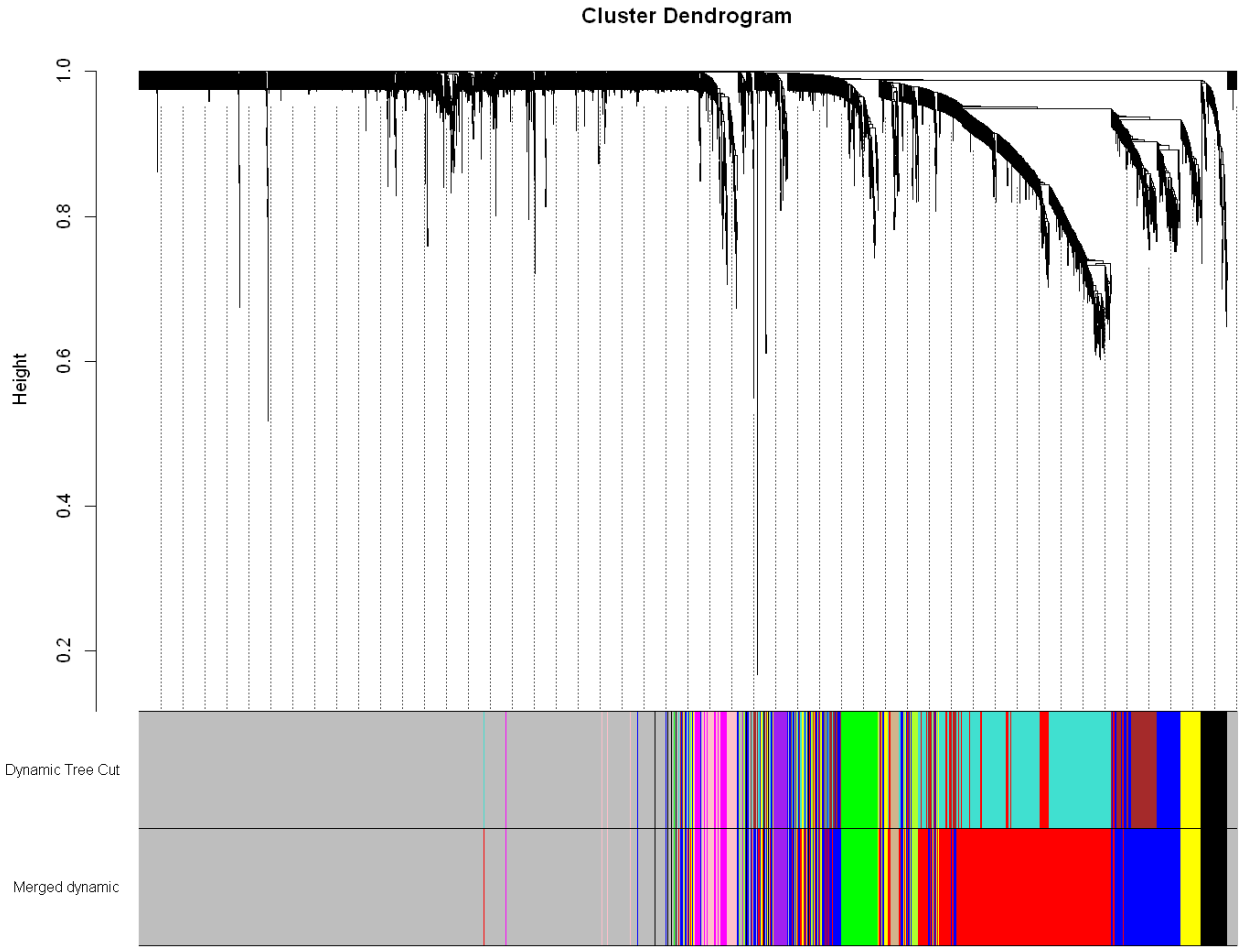
Identifying Gene Modules



Identifying Gene Modules



Merging of modules whose expression profiles are very similar



Merging of modules whose expression profiles are very similar

Module	Color	Dynamic Tree Cut	Merged Dynamic
0	grey	2595	2595
1	turquoise	841	
2	blue	321	618
3	brown	297	
4	yellow	224	224
5	green	182	182
6	red	172	1013
7	black	144	144
8	pink	127	127
9	magenta	71	71
10	purple	67	67
11	greenyellow	57	57
12	tan	44	44

Table: Number of genes and colors assigned to each module

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

LASSO method

LASSO (Least Absolute Shrinkage and Selection Operator)

Is a regularized linear regression technique, a method that combines a regression model with a procedure of contraction of some parameters towards zero and selection of variables, imposing a restriction or a penalty on the regression coefficients.

LASSO method

LASSO (Least Absolute Shrinkage and Selection Operator)

Is a regularized linear regression technique, a method that combines a regression model with a procedure of contraction of some parameters towards zero and selection of variables, imposing a restriction or a penalty on the regression coefficients.

Very usefull in problems where the number of variables (genes) n is much greater than the number of samples p ($n \gg p$).

Lasso solves the least squares problem with restriction on the L_1-norm of the coefficient vector:

$$\min \left\{ \sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \right\}, \text{sujeto a} \sum_{j=1}^{n} |\beta_j| \leq s$$

Or equivalently minimizing:

$$\sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} |\beta_j|$$

being s, $\lambda \geq 0$ the respective penalty parameters for complexity.

Lasso solves the least squares problem with restriction on the L_1-norm of the coefficient vector:

$$\min \left\{ \sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \right\}, \text{sujeto a} \sum_{j=1}^{n} |\beta_j| \leq s$$

Or equivalently minimizing:

$$\sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} |\beta_j|$$

being s, $\lambda \geq 0$ the respective penalty parameters for complexity.

LASSO produces parameter estimation and simultaneous variable selection for increasing values of λ.

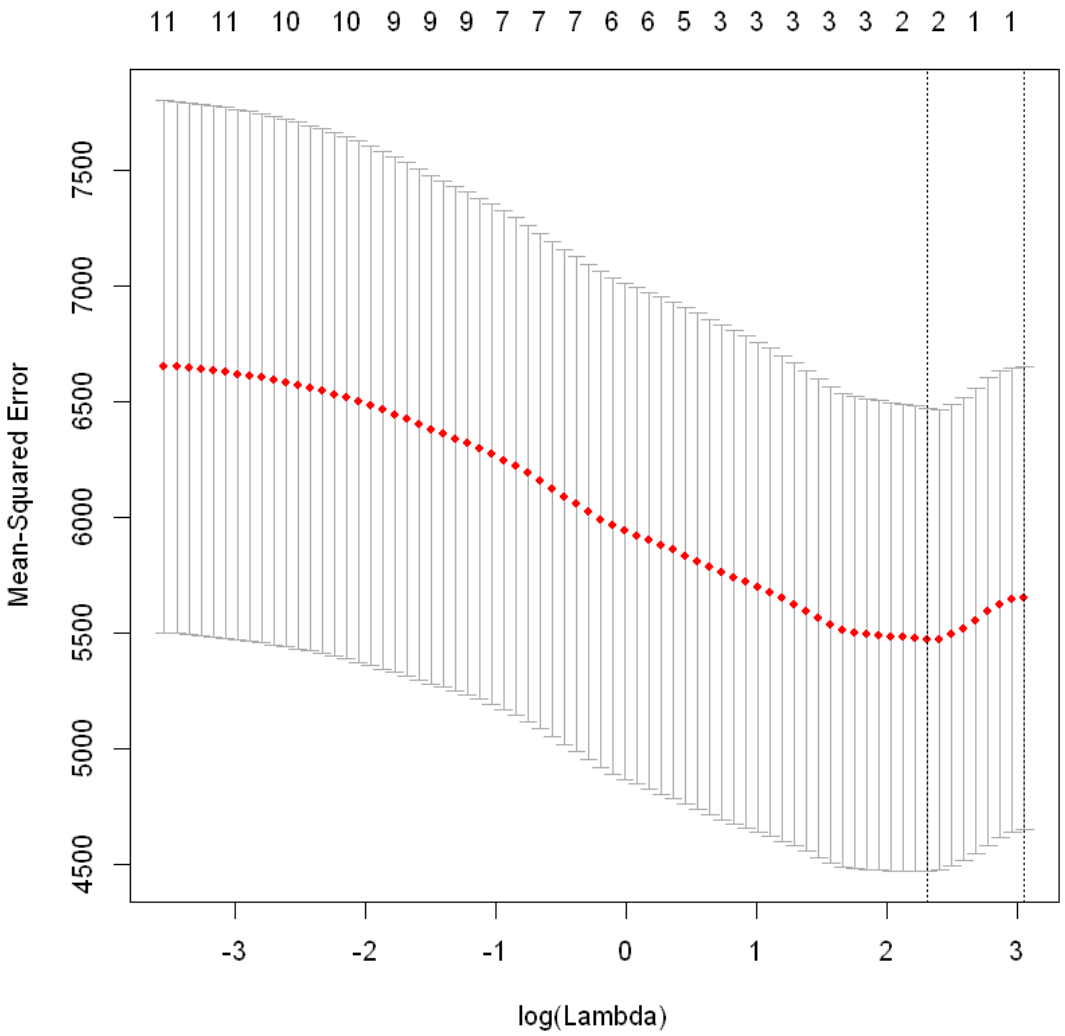
Phenotypic data

Phenotypic trait: $[Na^+]/[K^+]$ ratio in rice roots as an indicator of salinity tolerance, measured in the 91 accessions.

Salt's toxic effects:

- Osmotic stress: $[Na]_{pl} < [Na]_s$ impedes water uptake affecting cell expansion and growth. reduce stomatal conductance, transpiration, and carbon assimilation.
- Ionic stress: $[Na^+]/[K^+] > T$ induces apoptosis, the growth of young leaves is delayed and the senescence of old leaves is accelerated.

Find the best λ using cross-validation



Results

For $\lambda = 10.14$ two modules are seleted:

Beta	Module	Genes
505.92421	(Intercept)	
-12.17368	MEmagenta	71
-102.28731	MEgrey	2595

Contents

1. WGCNA
2. Data Structure
3. Network Construction
   - Gene co-expression similarity
   - Adjacency Function
   - Parameter of the Adjacency Function
   - Dissimilarity measure
4. Identifying Gene Modules
5. Relating modules to phenotypic data
   - LASSO method
   - Phenotypic data
   - Results
6. Bibliografía

Pontificia Universidad JAVERIANA Cali

Bibliografía

Malachy T Campbell, Nonoy Bandillo, Fouad Razzaq A Al Shiblawi, Sandeep Sharma, Kan Liu, Qian Du, Aaron J Schmitz, Chi Zhang, Anne-Aliénor Véry, Aaron J Lorenz, et al.
Allelic variants of oshkt1; 1 underlie the divergence between indica and japonica subspecies of rice (oryza sativa) for root sodium content.
PLoS genetics, 13(6):e1006823, 2017.

Kevin L Childs, Rebecca M Davidson, and C Robin Buell.
Gene coexpression network analysis as a source of functional annotation for rice genes.
PloS one, 6(7):e22196, 2011.

Qian Du, Malachy Campbell, Huihui Yu, Kan Liu, Harkamal Walia, Qi Zhang, and Chi Zhang.
Network-based feature selection reveals substructures of gene modules responding to salt stress in rice.
Plant direct, 3(8):e00154, 2019.

Peter Langfelder and Steve Horvath.
Wgcna: an r package for weighted correlation network analysis.
BMC bioinformatics, 9(1):559, 2008.

Yang Xu, Xin Wang, Xiaowen Ding, Xingfei Zheng, Zefeng Yang, Chenwu Xu, and Zhongli Hu.
Genomic selection of agronomic traits in hybrid rice using an ncii population.
Rice, 11(1):32, 2018.

¡¡GRACIAS!!