Inferring gene co-expression networks using a sparse factor model approach

Y. Blum[1], M. Houée[1,2], S. Lagarrigue[2] & D. Causeur[1]

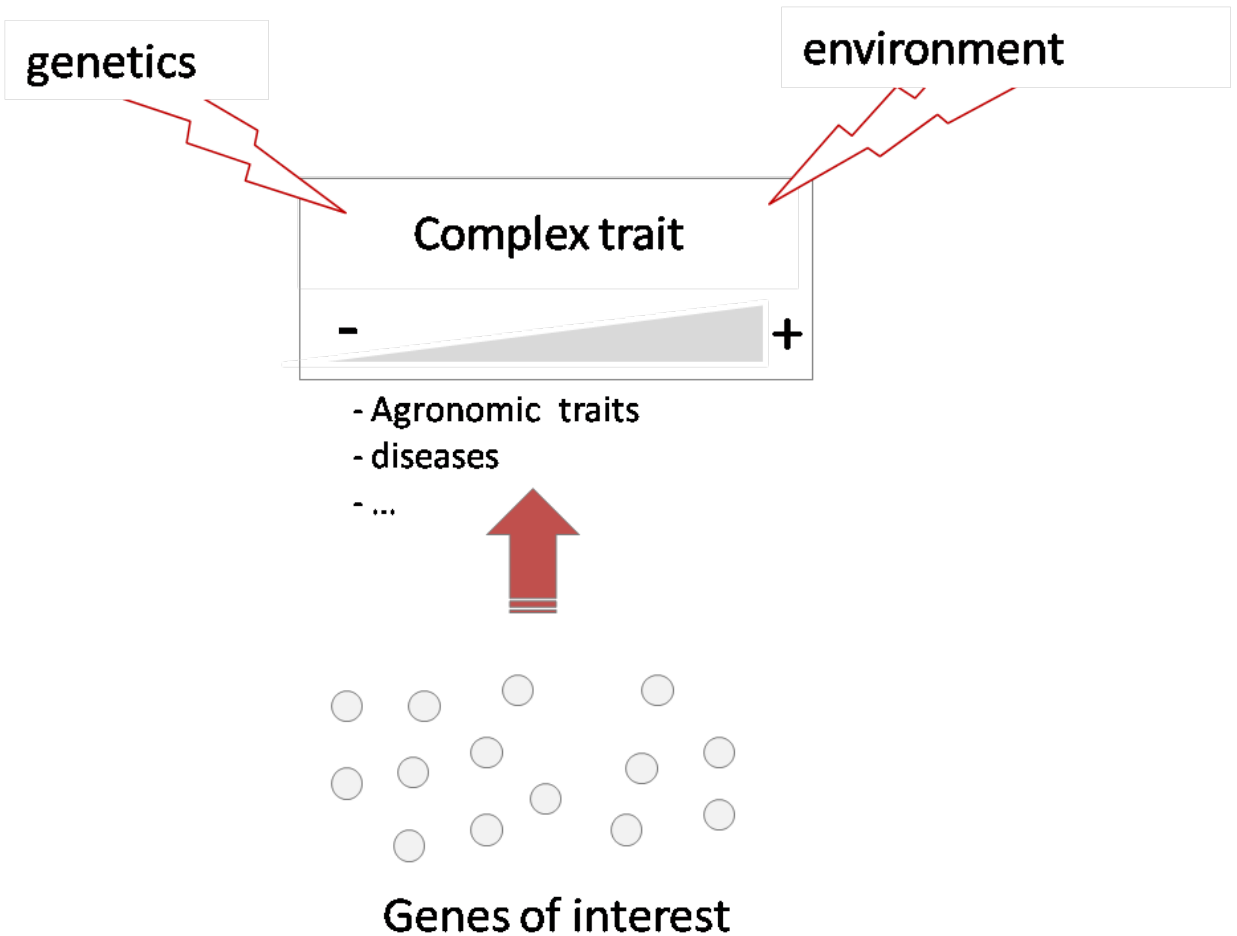
(1) Applied mathematics laboratory, (2) Genetics Genomics team PEGASE, Agrocampus Ouest, Rennes

Réunion NETBIO
September 12th 2013

Outline

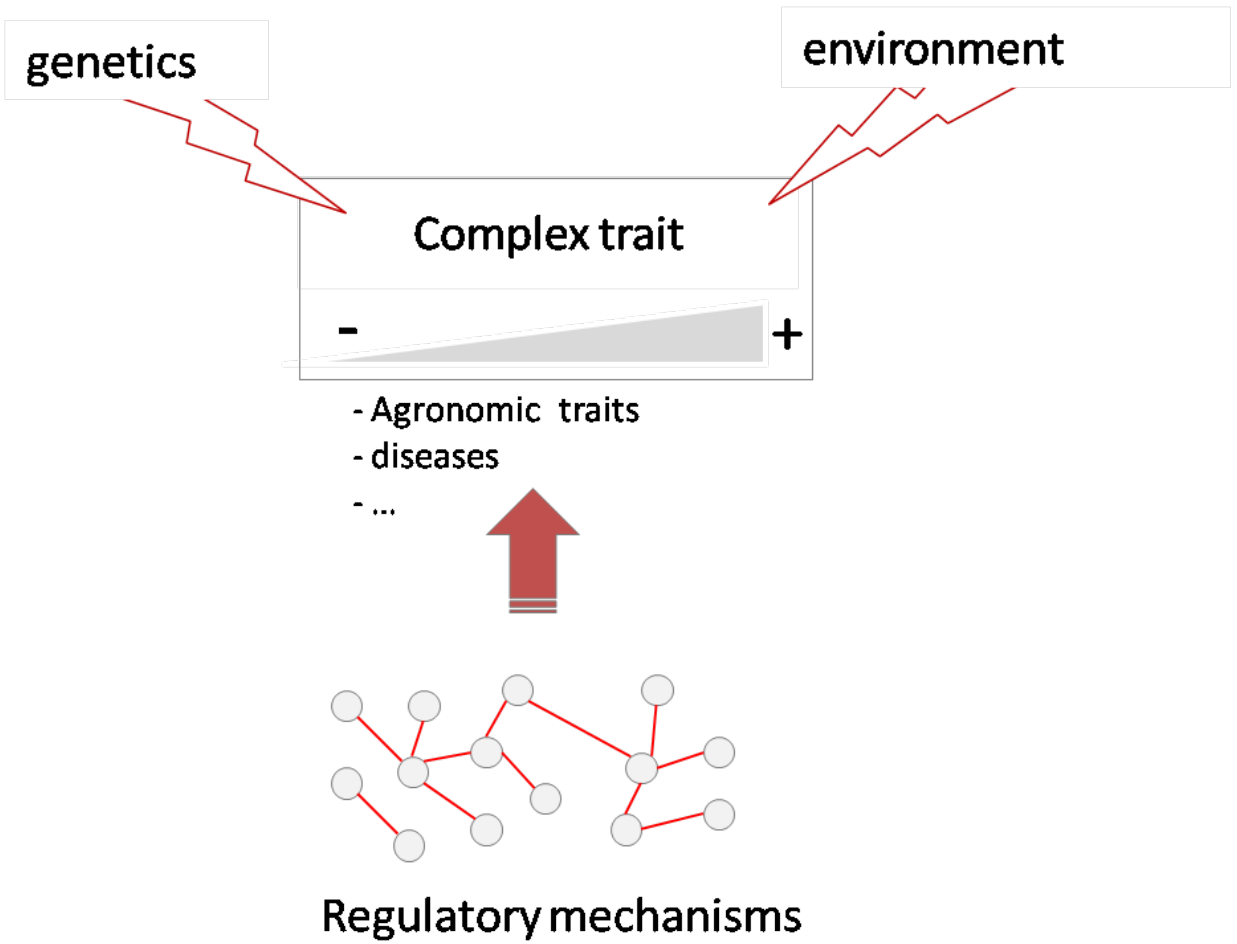
Thesis

Genetic analysis of a complex trait using transcriptomic data: contribution of gene network modeling.



Thesis

Genetic analysis of a complex trait using transcriptomic data: contribution of gene network modeling.



Thesis

Main results

- Improvement of current genetical genomics approaches
  - **Blum Y** *et al. A Factor Model to Analyze Heterogeneity in Gene Expression.* BMC Bioinformatics, 2010, 11:368. Highly Accessed
  - **Blum Y** *et al. Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken*. BMC Genomics, 2011, 12:567
  - Mach N, **Blum Y** *et al. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (DGAT1) in the mammary gland tissue of dairy cows*. Journal of Dairy Science, 2012

- Development of new methods for gene network inference
  - **Blum Y**, Houée M, Cadoret M, Causeur, *Sparse factor models for high dimensional relevance networks*. *(submitted)*

Thesis

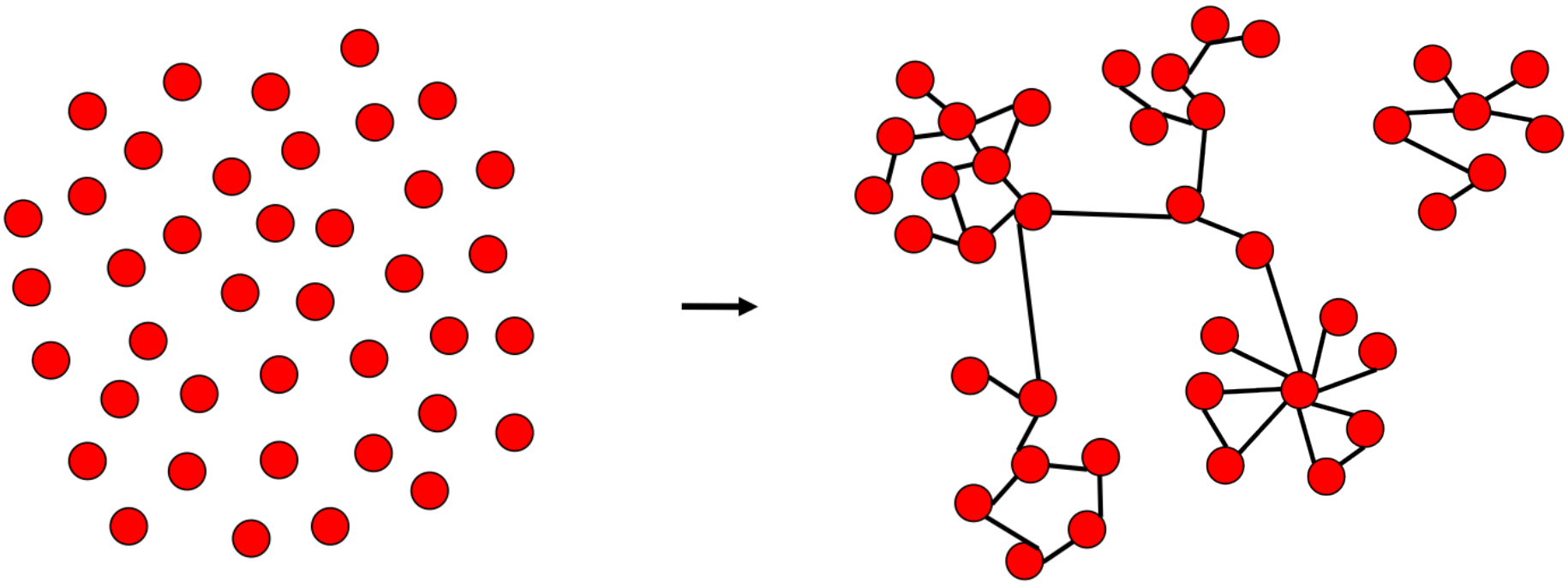
Main results

- Improvement of current genetical genomics approaches
  - **Blum Y** *et al. A Factor Model to Analyze Heterogeneity in Gene Expression.* BMC Bioinformatics, 2010, 11:368. Highly Accessed
  - **Blum Y** *et al. Complex trait subtypes identification using transcriptome profiling reveals an interaction between two QTL affecting adiposity in chicken*. BMC Genomics, 2011, 12:567
  - Mach N, **Blum Y** *et al. Pleiotropic effects of polymorphism of the gene diacylglycerol-o-transferase 1 (DGAT1) in the mammary gland tissue of dairy cows*. Journal of Dairy Science, 2012

- Development of new methods for gene network inference
  - **Blum Y**, Houée M, Cadoret M, Causeur, *Sparse factor models for high dimensional relevance networks*. *(submitted)*

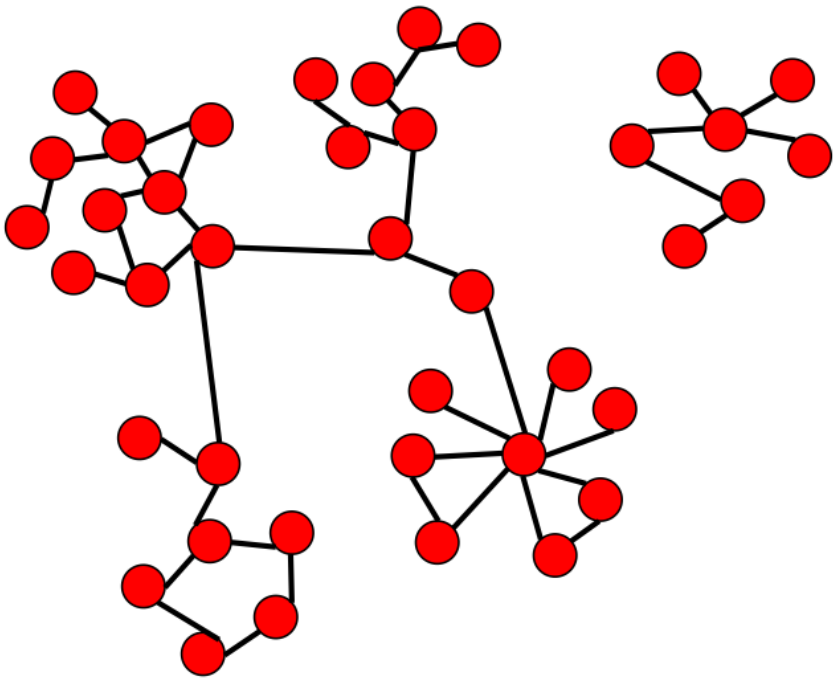
Gene network modeling

Genes of interest



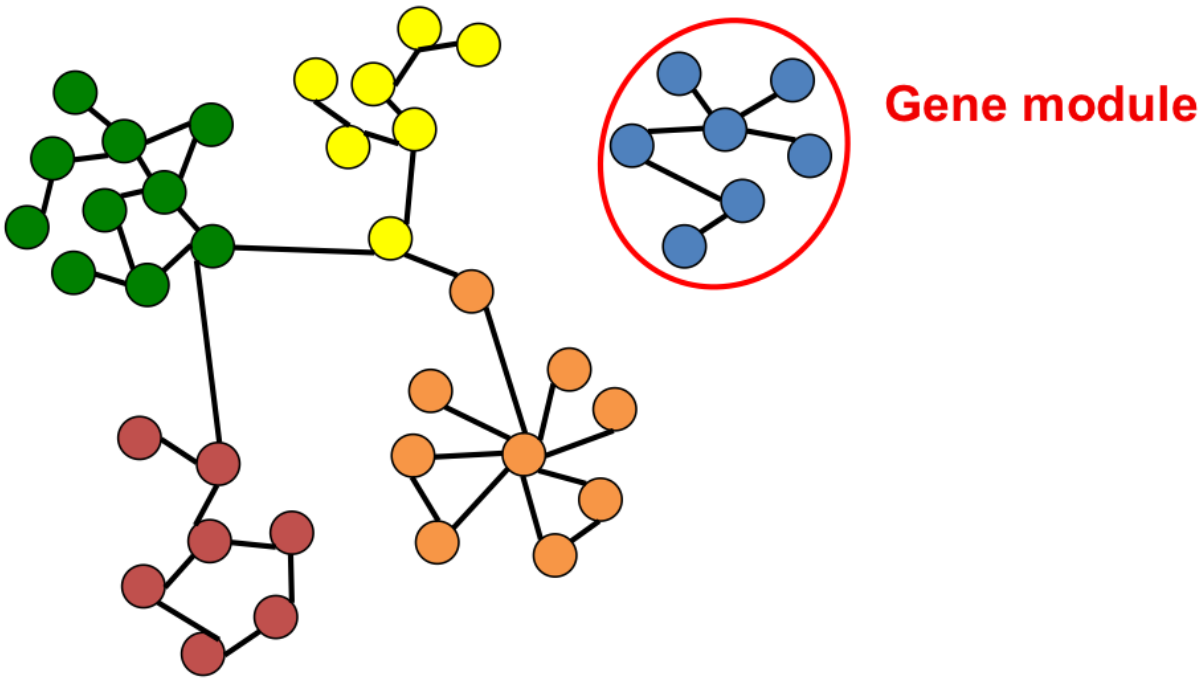
Gene network modeling

node=gene
edge=link between 2 genes



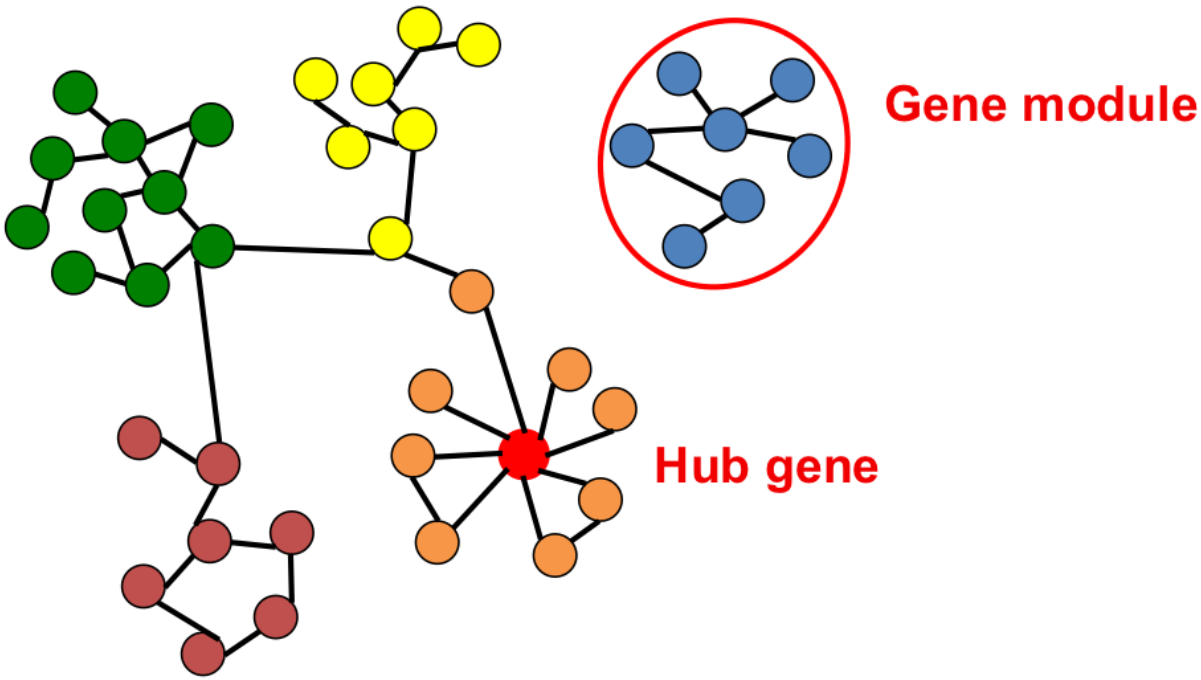
Gene network modeling

node=gene
edge=link between 2 genes



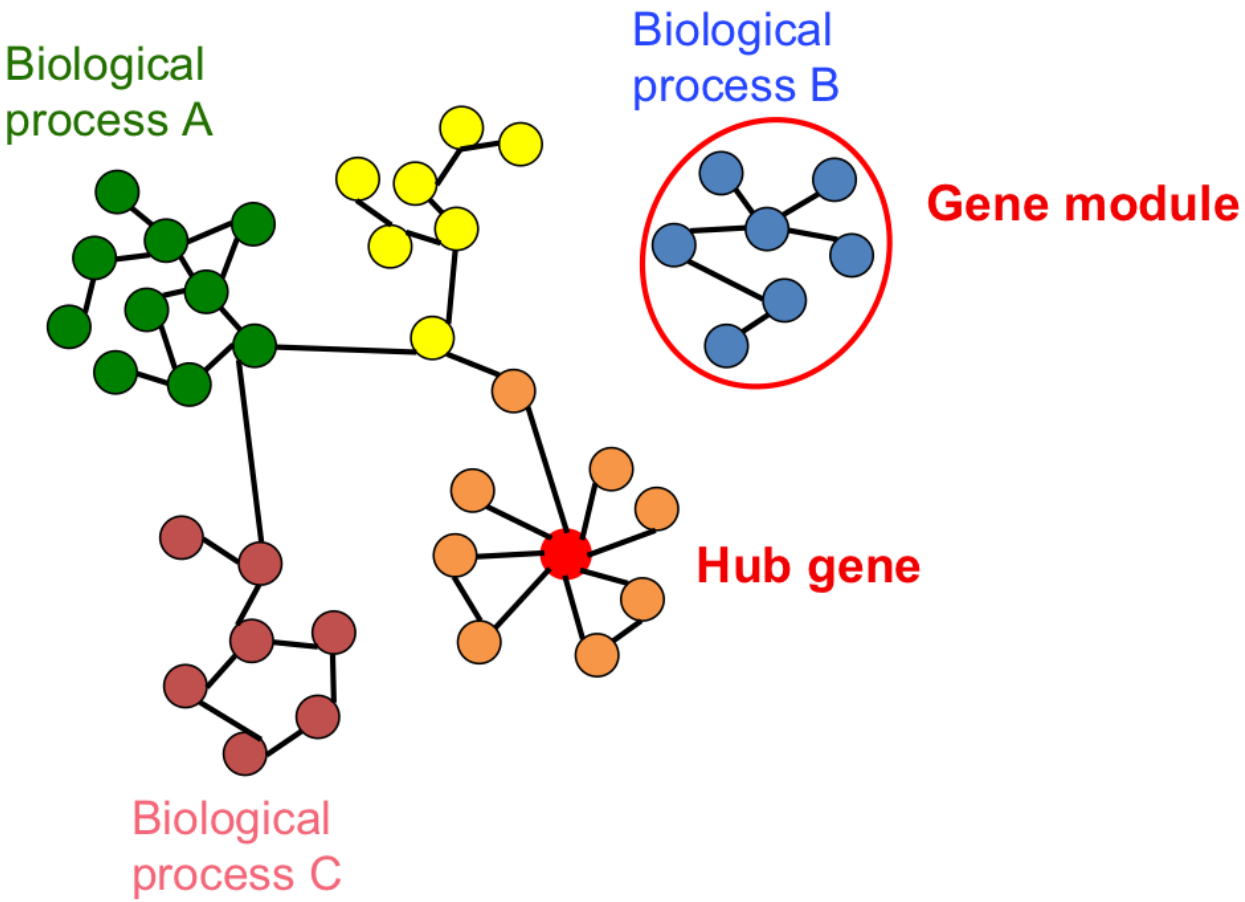
Gene network modeling

node=gene
edge=link between 2 genes



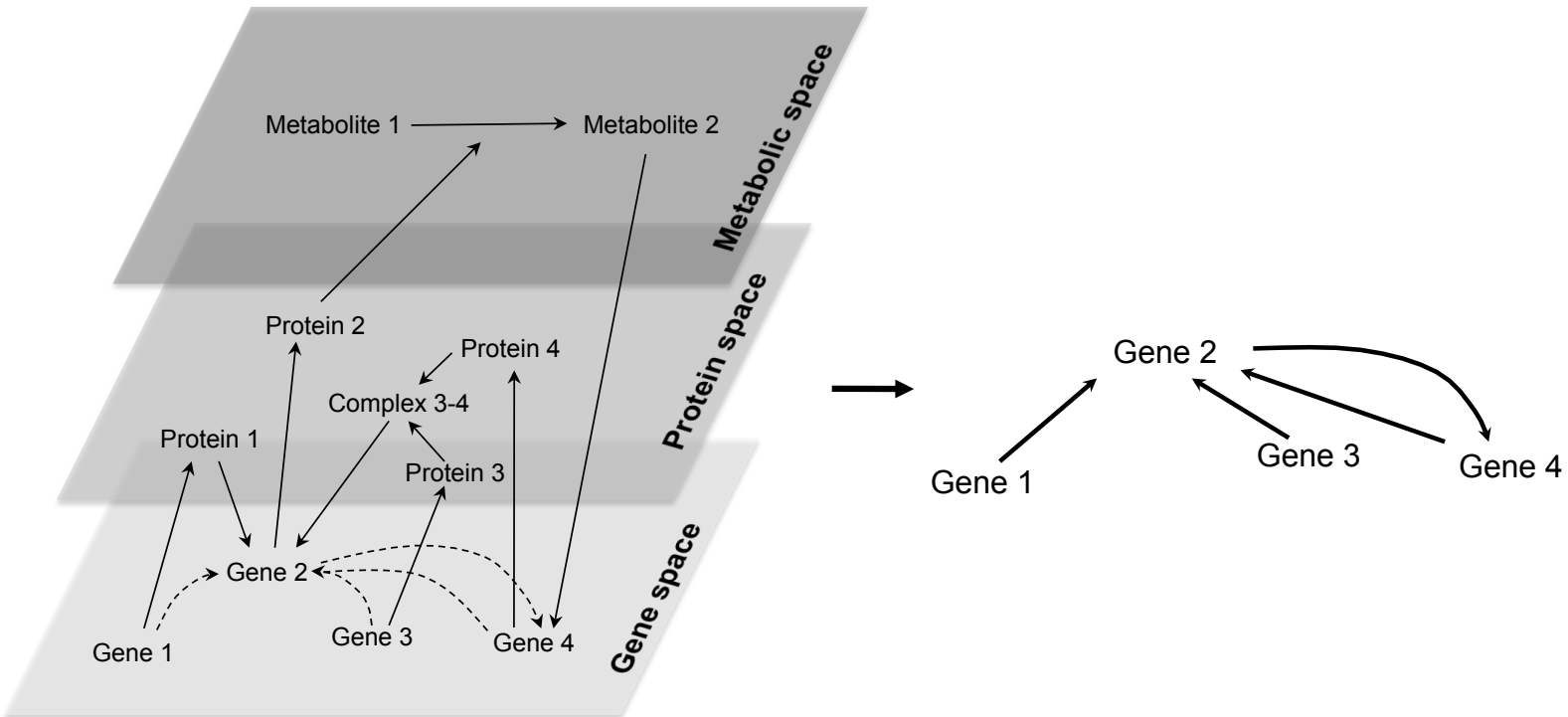
Gene network modeling

node=gene
edge=link between 2 genes



Gene network modeling

Projection of all interactions to the gene space

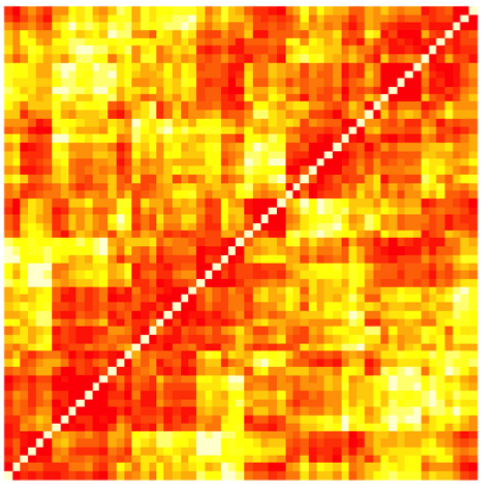


(Brazhnik et al., 2002)

Outline

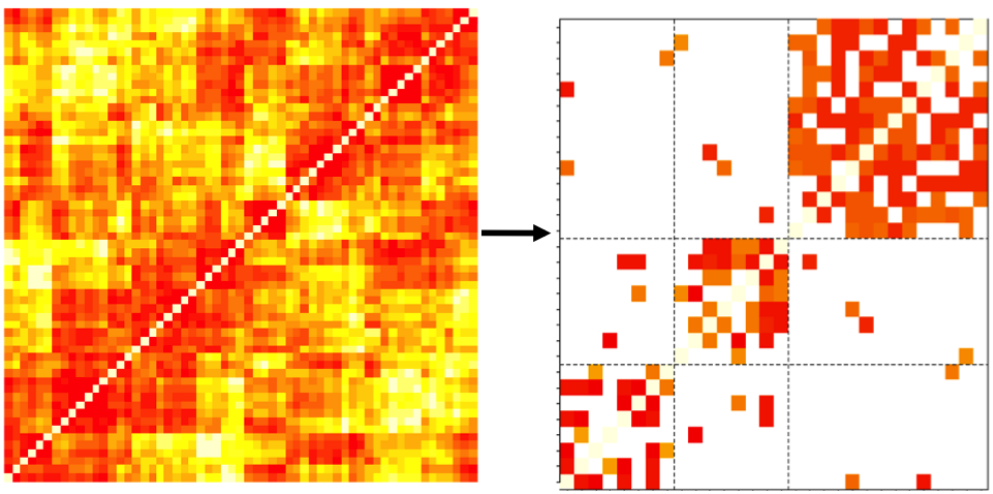
Network construction

1. Choose a measure L of the link between 2 genes



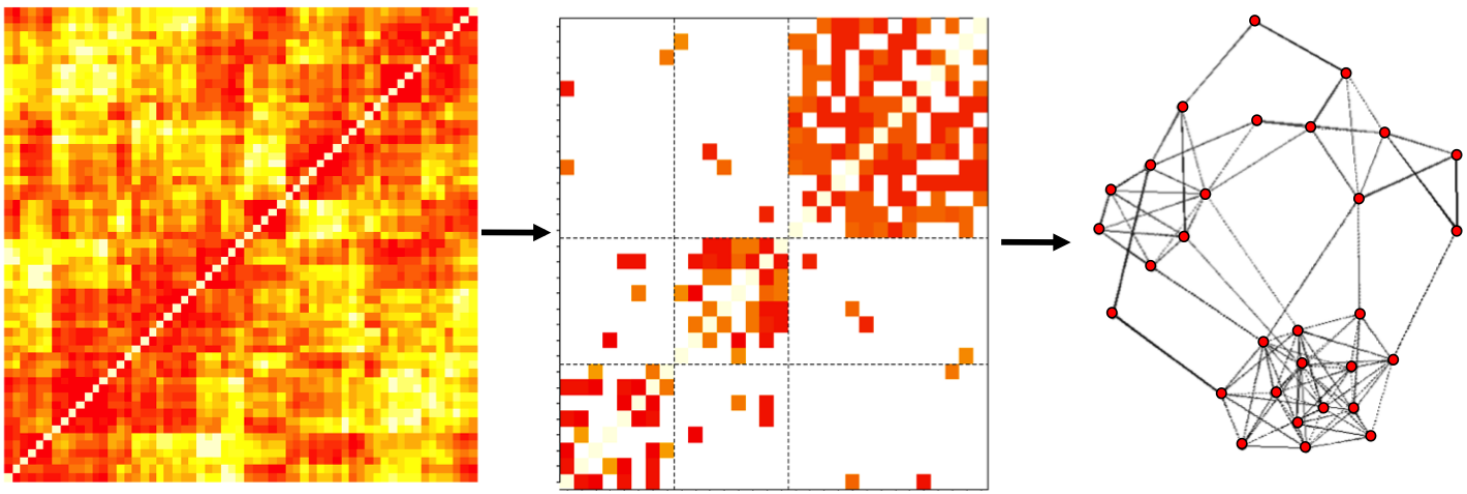
Network construction

1. Choose a measure L of the link between 2 genes
2. Decision rule: is L(y_i,y_j) different from 0 ?



Network construction

1. Choose a measure L of the link between 2 genes
2. Decision rule: is $L(y_i, y_j)$ different from 0 ?
3. Vizualisation through a graph



Measure of the link

Linear measures:

- Pearson correlation: $corr(y_i, y_j) = \frac{Cov(y_i, y_j)}{\sqrt{Var(y_i)Var(y_j)}}$

  $\Rightarrow$ Relevance network

  (Butte & Kohane 2000, Langfelder & Horvath 2008 `WGCNA`)

- Partial correlation: $corr(y_i, y_j | y_{\setminus i,j})$

  $\Rightarrow$ Gaussian Graphical Model (GGM)

  (Schäfer & Strimmer 2005 `GeneNet`, Peng *et al.* 2009 `SPACE`)

Non-linear measure:

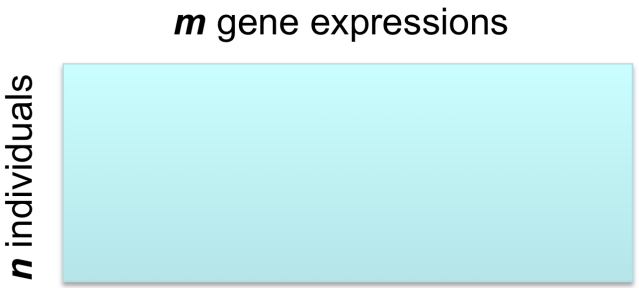
- Mutual information MI:
  $MI(x, y) = \sum_{i=1}^{r} \sum_{j=1}^{r} P(x = i, y = i) log \frac{P(x=i, y=j)}{P(x=i)P(y=j)}$

  $\Rightarrow$ information-theory-based methods

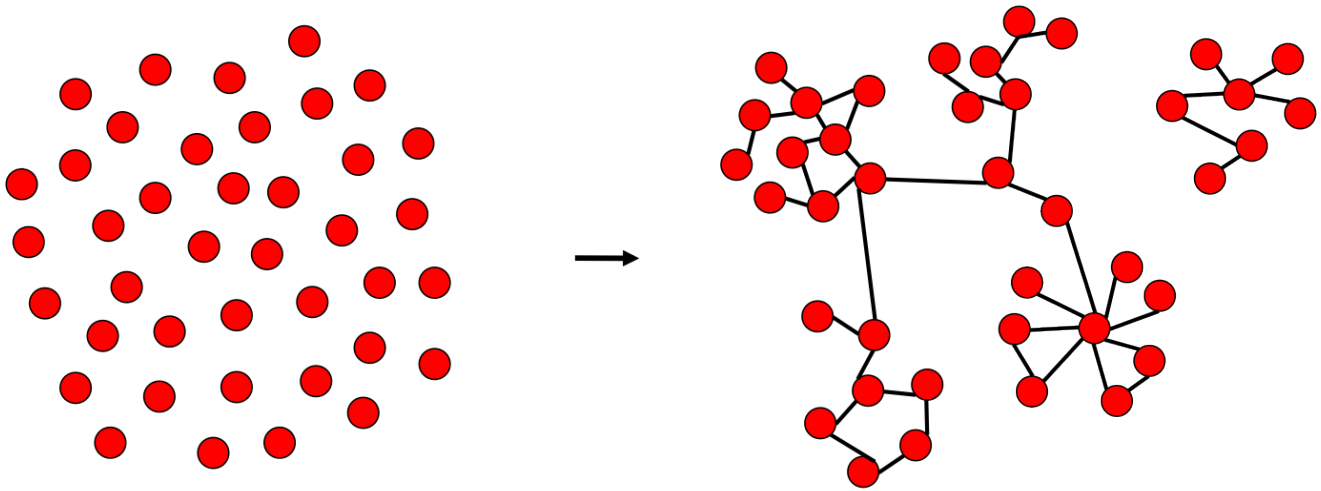
  (Margolin *et al.* 2006 `ARACNE`)

A challenging task

- High dimension: $n << m$



- Sparsity assumption: within a set of genes, only a few are interacting (Tegner et al 2003).



Relevance networks: the WGCNA approach

WGCNA: Weighted Gene Co-expression network Zuang et *et al.* 2005, Langfelder & Horvath 2008

Network estimation:

- Σ is estimated using the empirical estimator S.
- Power function on the empirical correlations:

$$a_{ij} = |s_{ij}|^{\beta}$$

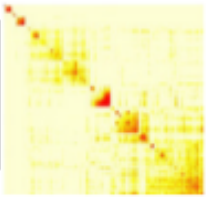
where β is chosen to satisfy a parsimonious criteria

$\Rightarrow$ noise is decreased, better extraction of the modular structure(Horvath *et al.*, 2005)

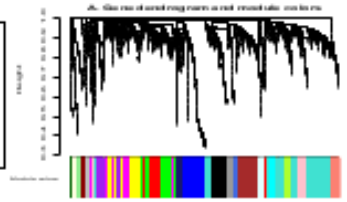
Module detection: Topological Overlap Measure (normalized number of shared neighbors).

WGCNA: Weighted Gene Co-expression network

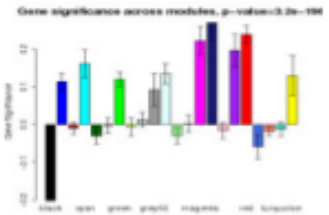
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



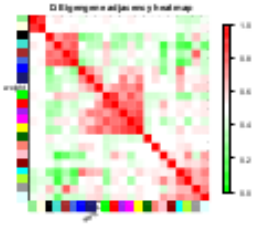
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



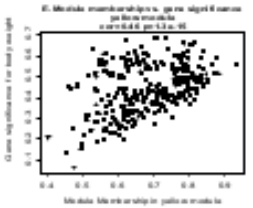
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



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



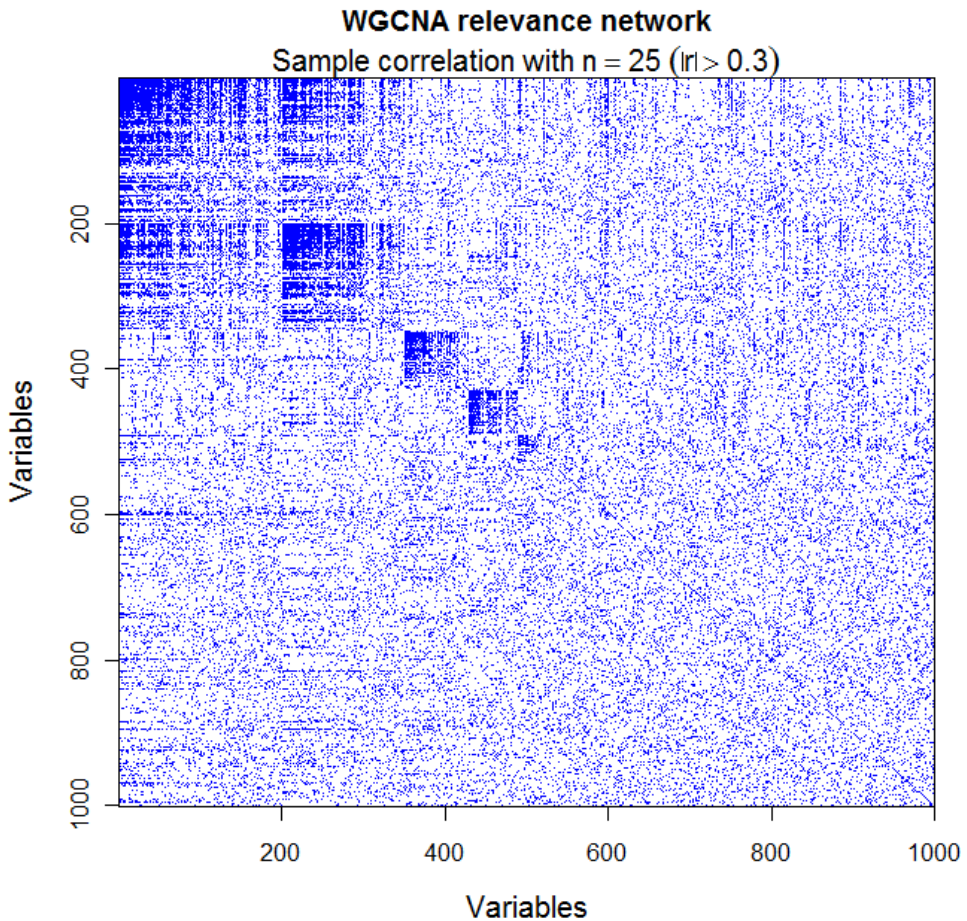
Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



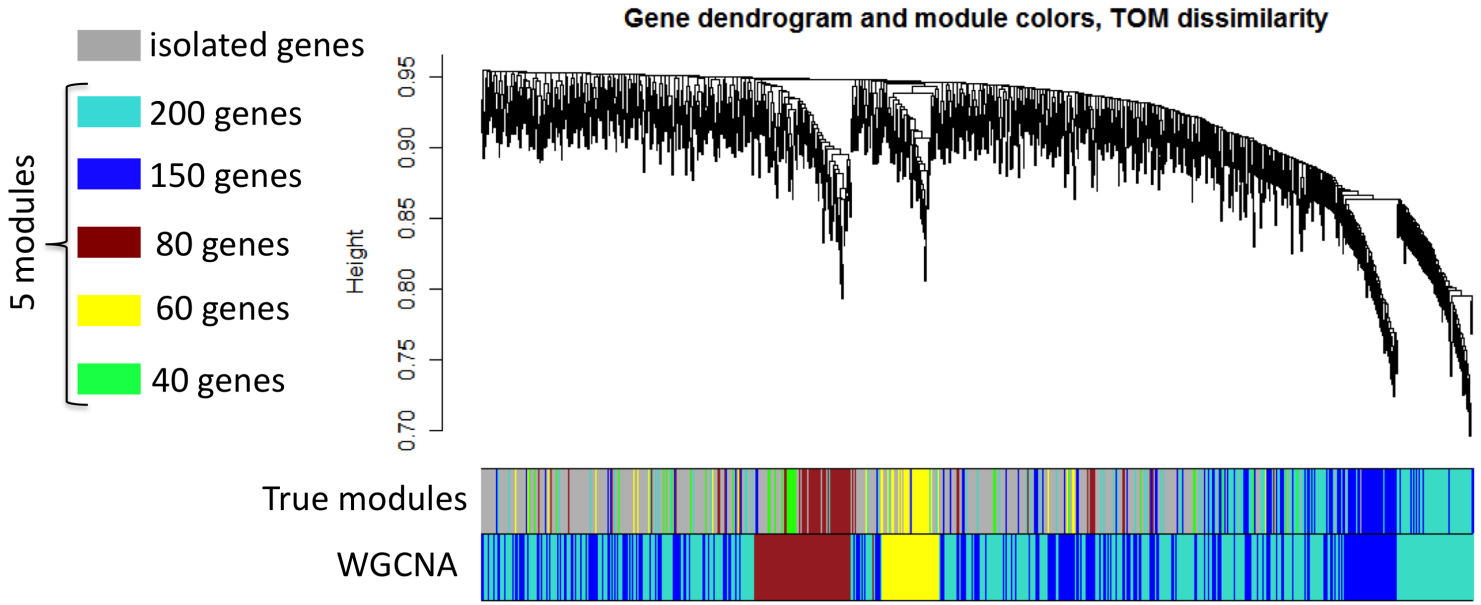
Simulated example

Dataset simulated using the R package WGCNA with $m = 1000$ and $n = 25$

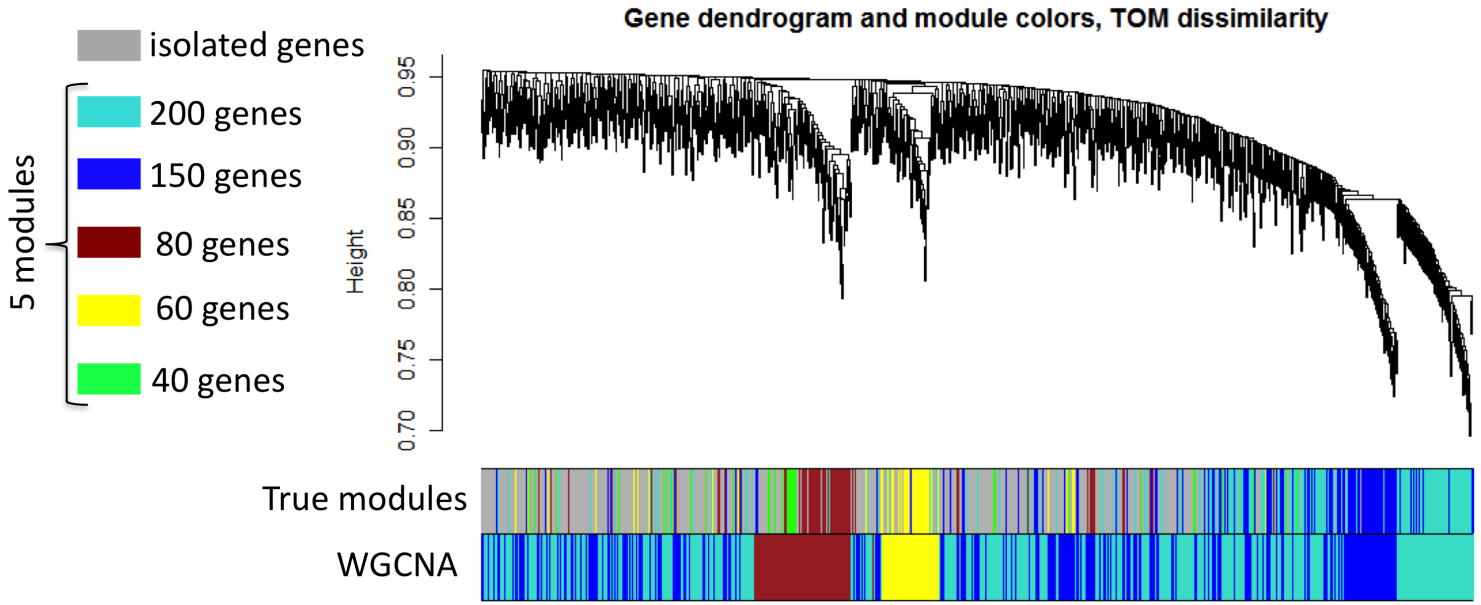
5 modules: $m_1 = 200$ genes, $m_2 = 150$, $m_3 = 80$, $m_4 = 60$, $m_5 = 40$



WGCNA results based on the empirical correlations



WGCNA results based on the empirical correlations



TOM classification								
Weighted adjacency with $\beta = 2$ - Rand index = 0.603								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**0.00**	56.60	33.20	7.20	3.00	0.00	100.00	470.00
turquoise	0.00	**86.00**	12.00	0.50	1.50	0.00	100.00	200.00
blue	0.00	26.70	**70.70**	2.00	0.70	0.00	100.10	150.00
brown	0.00	16.20	22.50	**60.00**	1.20	0.00	99.90	80.00
yellow	0.00	23.30	13.30	0.00	**63.30**	0.00	99.90	60.00
green	0.00	55.00	12.50	32.50	0.00	**0.00**	100.00	40.00

Confusion matrix for clustering with WGCNA based on empirical correlations

Outline

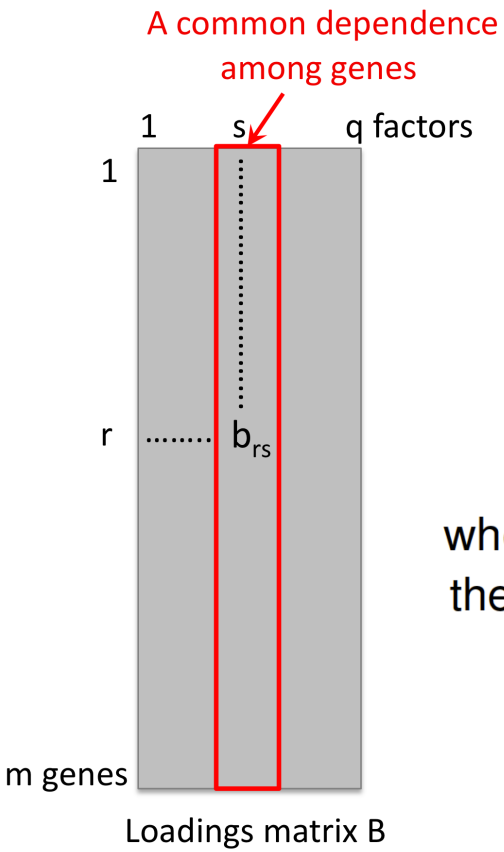
Our proposal

Factor Model:

Y dataset with n rows (individuals) and p columns (genes).

$Y \sim \mathcal{N}_p(\mu, \Sigma)$

Correlations between genes are described by a small number q of factors containing a common dependence:



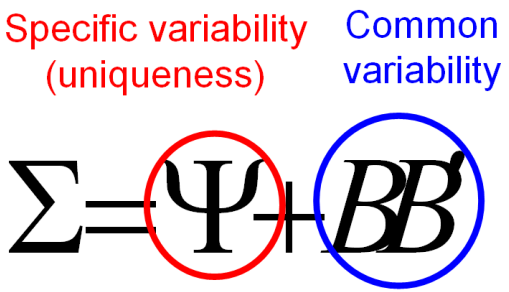
Specific variability (uniqueness) Common variability

$$\Sigma = \Psi + BB'$$

where Ψ is a diagonal matrix, B represents the $m \times q$ matrix of loadings b_k.

Our proposal

In WGCNA procedure:
use correlations estimated by a **factor model**.



Deviance: $\mathcal{D}(\Psi, B) \propto \log\det(\Psi + BB') + \text{trace}[S(\Psi + BB')^{-1}]$

- Estimation of the Ψ and B: EM algorithm (Rubin & Thayer, 1982)

Our proposal

Maximum Likelihood estimation: EM algorithm

- E-step: calculation of the expectation of $\mathcal{D}(\Psi, B; Z)$

$$n^{-1}\mathbb{E}_y\mathcal{D}(\Psi, B; Z) = \sum_{j=1}^{m} \log \psi_j^2 + [(\Psi^{-1}S) - 2(\Psi^{-1}BC'_{yz}) + (B'\Psi^{-1}BC_{zz})] + (C_{zz})$$

where $C_{yz} = \mathbb{E}_y(S_{yz})$ and $C_{zz} = \mathbb{E}_y(S_{zz})$.

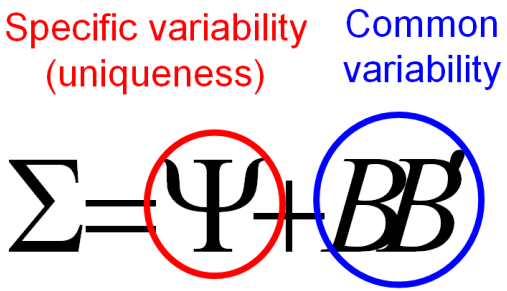
- M-step: minimization of the expected deviance

$$\hat{B} = C_{yz}C_{zz}^{-1}$$

$$\hat{\psi}_j^2 = S_{jj} - 2(\hat{B}C'_{yz})_{jj} + (\hat{B}C_{zz}\hat{B}')_{jj}$$

Our proposal

In WGCNA procedure:
use correlations estimated by a **factor model**.

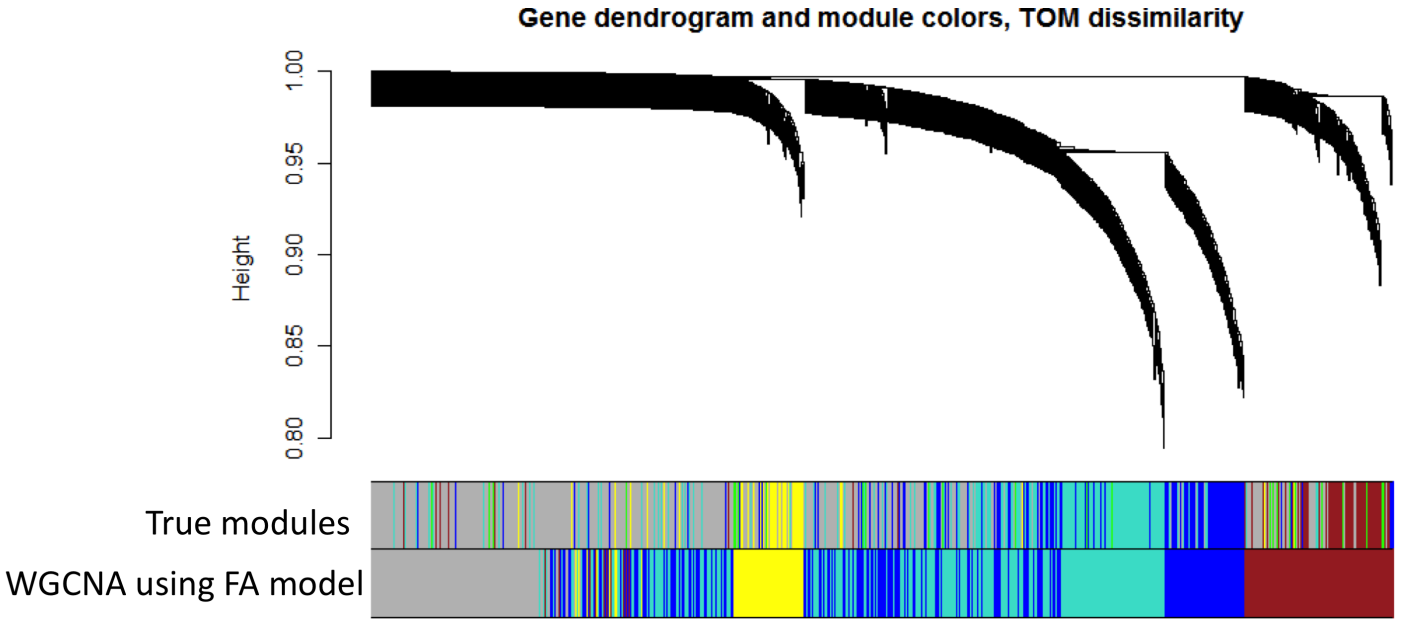


Deviance: $\mathcal{D}(\Psi, B) \propto \log\det(\Psi + BB') + \text{trace}[S(\Psi + BB')^{-1}]$

- Estimation of the Ψ and B: EM algorithm (Rubin & Thayer, 1982) $\rightarrow$ Calculate and minimize the expectation of $\mathcal{D}(\Psi, B; Z)$

- Number of factors: parallel analysis (Buja & Eyuboglu, 1994)

FA model in WGCNA

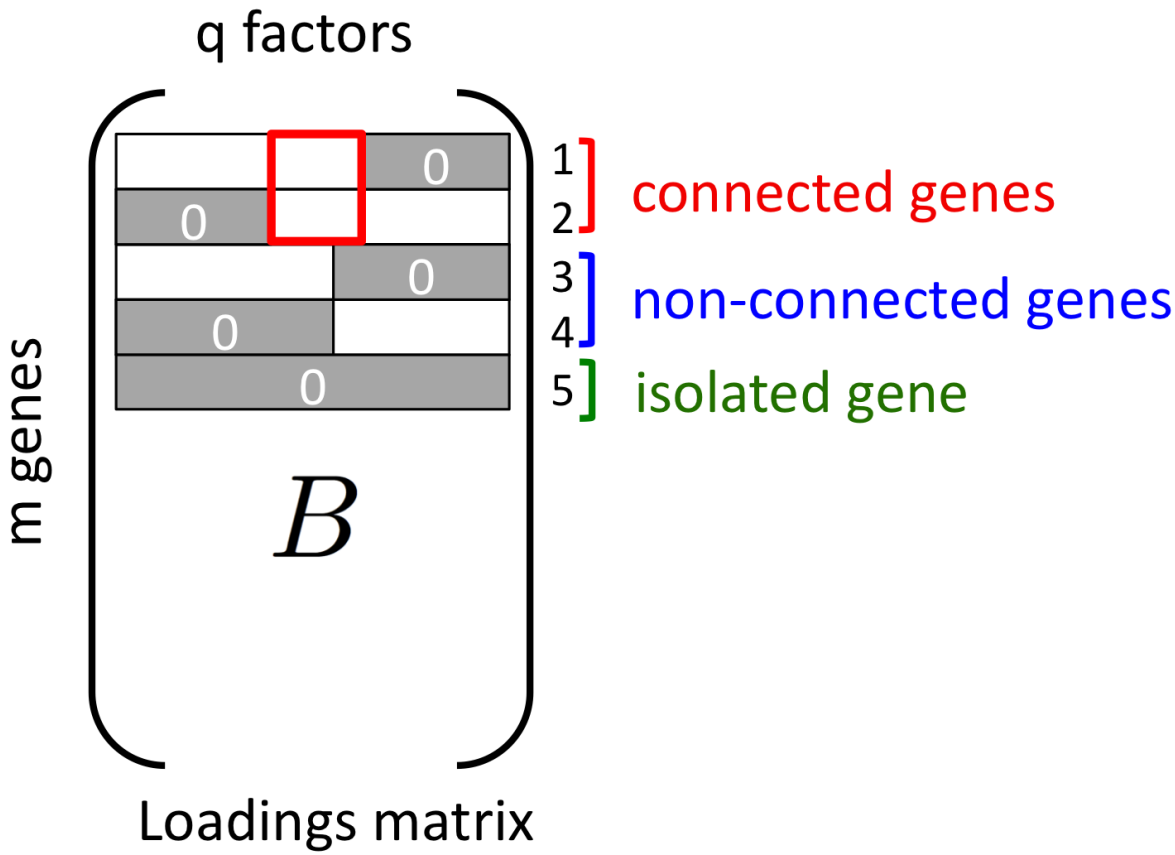


TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.654								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**30.20**	27.20	23.00	14.50	5.10	0.00	100.00	470.00
turquoise	7.00	**71.00**	17.00	2.00	3.00	0.00	100.00	200.00
blue	4.00	22.00	**67.30**	3.30	3.30	0.00	99.90	150.00
brown	10.00	2.50	5.00	**81.20**	1.20	0.00	99.90	80.00
yellow	5.00	13.30	11.70	5.00	**65.00**	0.00	100.00	60.00
green	10.00	35.00	7.50	35.00	12.50	**0.00**	100.00	40.00

Confusion matrix for clustering with WGCNA based on FA model for correlations

Sparse factor model

The topology of the network can be deduced from the loadings matrix B:



Sparse factor model

Inference on sparse matrix:

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

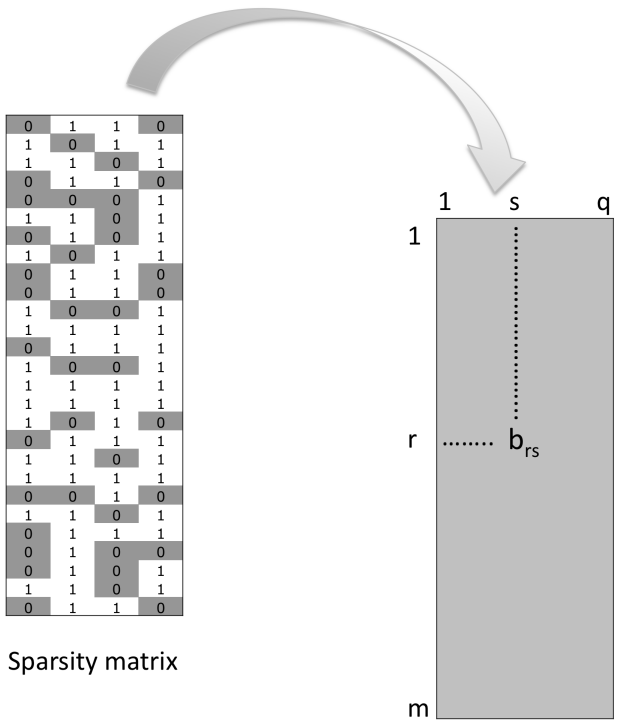
Sparsity matrix

Sparse factor model

EM algorithm for sparse factor structure:

M-step: minimizing the expected deviance

$$\mathcal{D}(\Psi, B; Z)$$

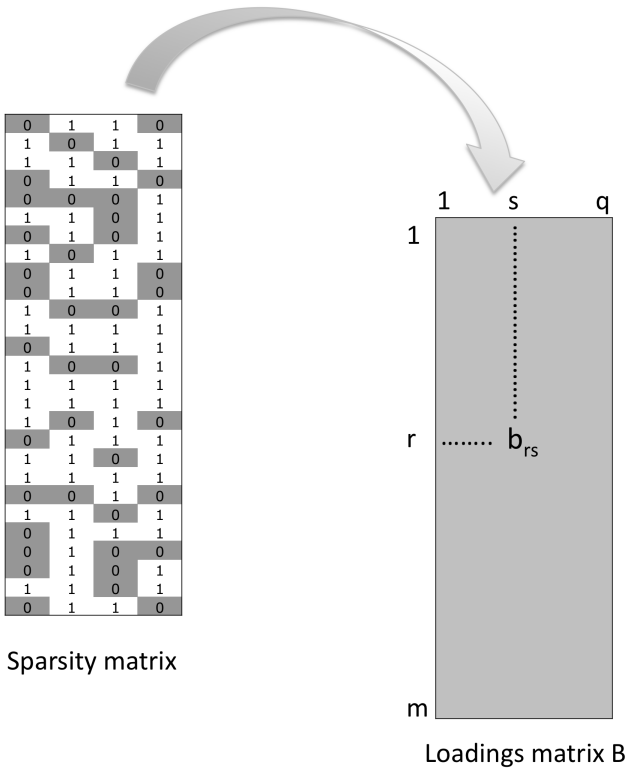


Sparse factor model

EM algorithm for sparse factor structure:
M-step: minimizing the expected deviance under sparsity constraints

$$\mathcal{D}(\Psi, B; Z) + \lambda' R' vec(B) \qquad \textit{Lagrange multiplier approach}$$

where λ is the vector of Lagrange multipliers and R the constraints matrix deduced from the sparsity matrix.



Estimation of the model parameters

Maximum Likelihood estimation: EM algorithm

- E-step: calculation of the expectation of $\mathcal{D}(\Psi, B; Z)$

$$n^{-1}\mathbb{E}_y\mathcal{D}(\Psi, B; Z) = \sum_{j=1}^{m} \log \psi_j^2 + [(\Psi^{-1}S) - 2(\Psi^{-1}BC'_{yz}) + (B'\Psi^{-1}BC_{zz})] + (C_{zz})$$

where $C_{yz} = \mathbb{E}_y(S_{yz})$ and $C_{zz} = \mathbb{E}_y(S_{zz})$.

- M-step: minimization of $\mathcal{D}(\Psi, B; Z)$

$$\hat{b}_r = [C_{zz}^{-1} - C_{zz}^{-1} C_r^{\star} C_{zz}^{-1}] C_{yz,(r)}$$

where $C_{yz,(r)}$ stands for the rth row of C_{yz} and $C_r^{\star}$ is a $q \times q$ symmetric matrix which entry (i, j) is zero if the corresponding loadings b_{ri} and b_{rj} are nonzero

Sparse factor model

Topology of the network depends on B

Inference on the sparsity of B

- Tests $b_{ki} = 0$
- ℓ_1-regularization

$$\min_{\Psi, B} \mathcal{D}(\Psi, B) + \lambda \sum_{k=1}^{m} \sum_{i=1}^{q} |b_{ki}|$$

λ chosen by minimization of BIC

$$\text{BIC}(\lambda) = \mathcal{D}(\hat{\Psi}_\lambda, \hat{B}_\lambda) + 2\#\{(k,i),\ \hat{b}_{\lambda,ki} \neq 0\}$$

- External information (Gene Ontology, KEGG, ...)

Results using a sparse factor model

Test for significance of loadings

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.710								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**62.60**	12.30	3.20	13.60	8.30	0.00	100.00	470.00
turquoise	15.00	**73.50**	4.50	4.50	2.50	0.00	100.00	200.00
blue	12.70	34.70	**42.00**	6.00	4.70	0.00	100.10	150.00
brown	13.80	3.80	1.20	**77.50**	3.80	0.00	100.10	80.00
yellow	15.00	3.30	1.70	3.30	**76.70**	0.00	100.00	60.00
green	32.50	22.50	2.50	32.50	10.00	**0.00**	100.00	40.00

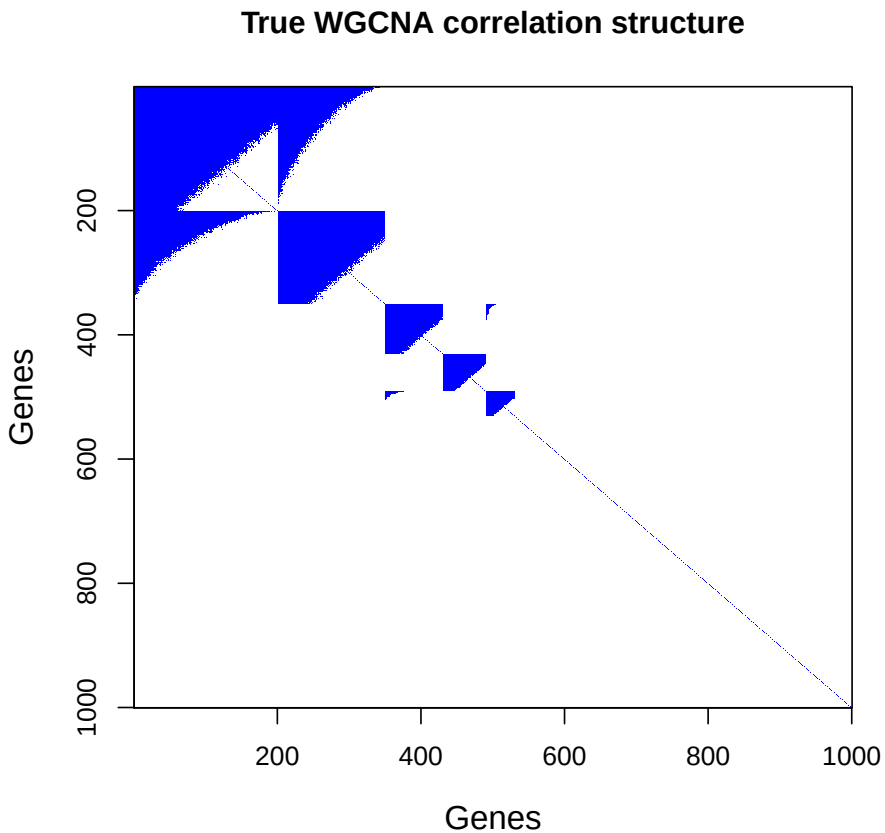
LASSO estimation

TOM classification								
Weighted adjacency with $\beta = 4$ - Rand index = 0.670								
True modules	grey	turquoise	blue	brown	yellow	green	Total	Count
grey	**43.40**	24.30	9.80	14.30	8.30	0.00	100.10	470.00
turquoise	10.00	**71.50**	12.50	3.00	3.00	0.00	100.00	200.00
blue	8.70	38.00	**49.30**	2.00	2.00	0.00	100.00	150.00
brown	11.20	2.50	8.80	**75.00**	2.50	0.00	100.00	80.00
yellow	10.00	6.70	1.70	1.70	**80.00**	0.00	100.10	60.00
green	20.00	22.50	10.00	32.50	15.00	**0.00**	100.00	40.00

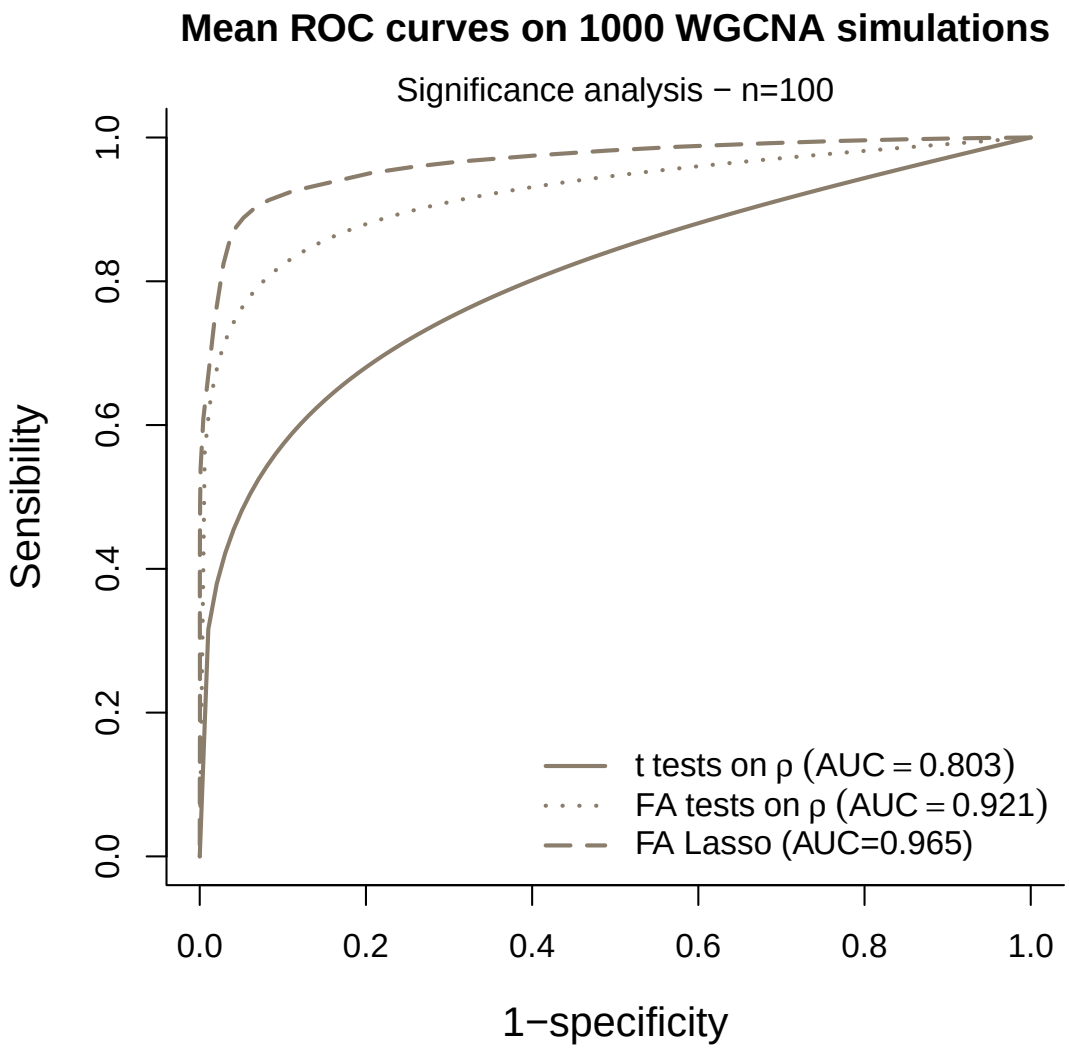
Simulation study

1000 datasets simulated from this true correlation structure (using n=10000).

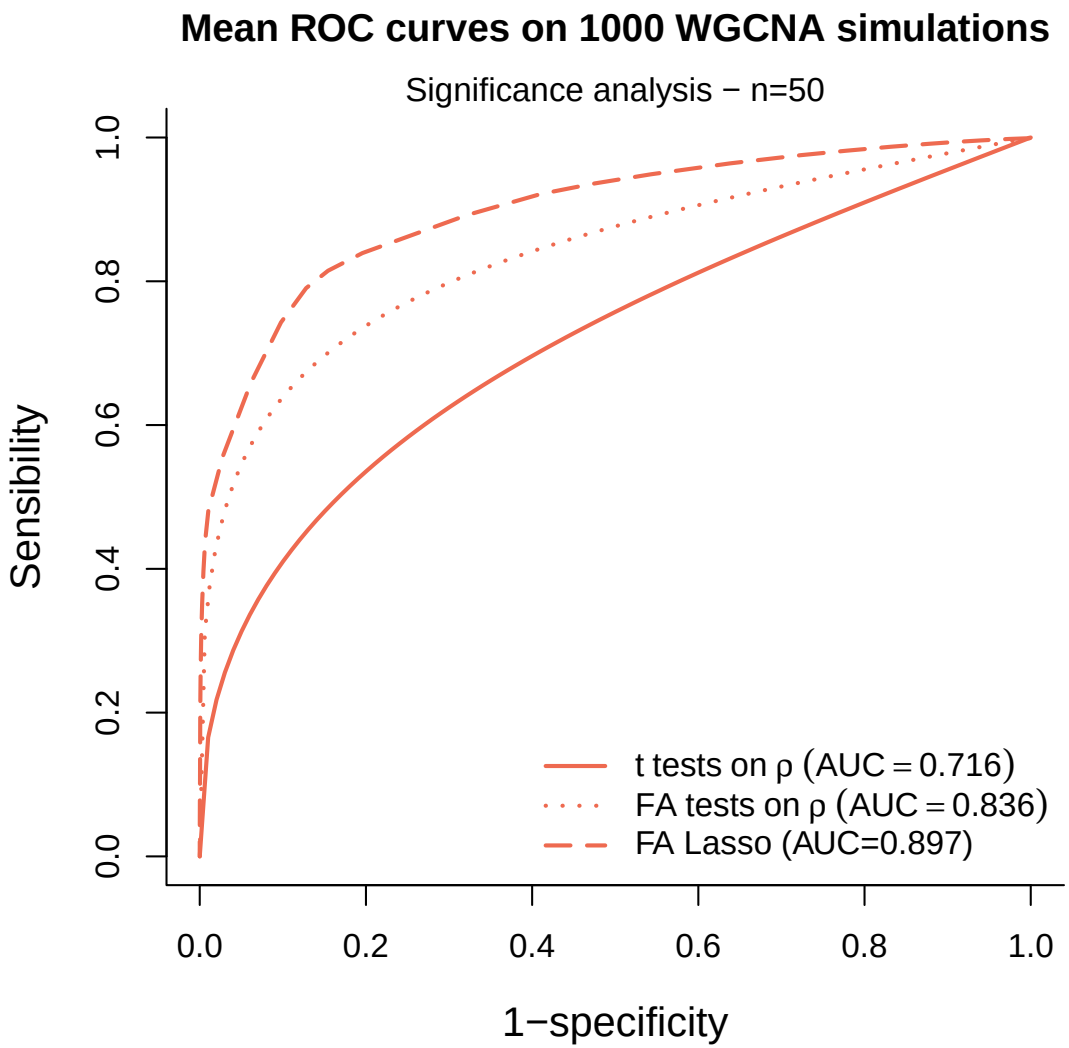
5 modules: $m_1 = 200$ genes, $m_2 = 150$, $m_3 = 80$, $m_4 = 60$, $m_5 = 40$



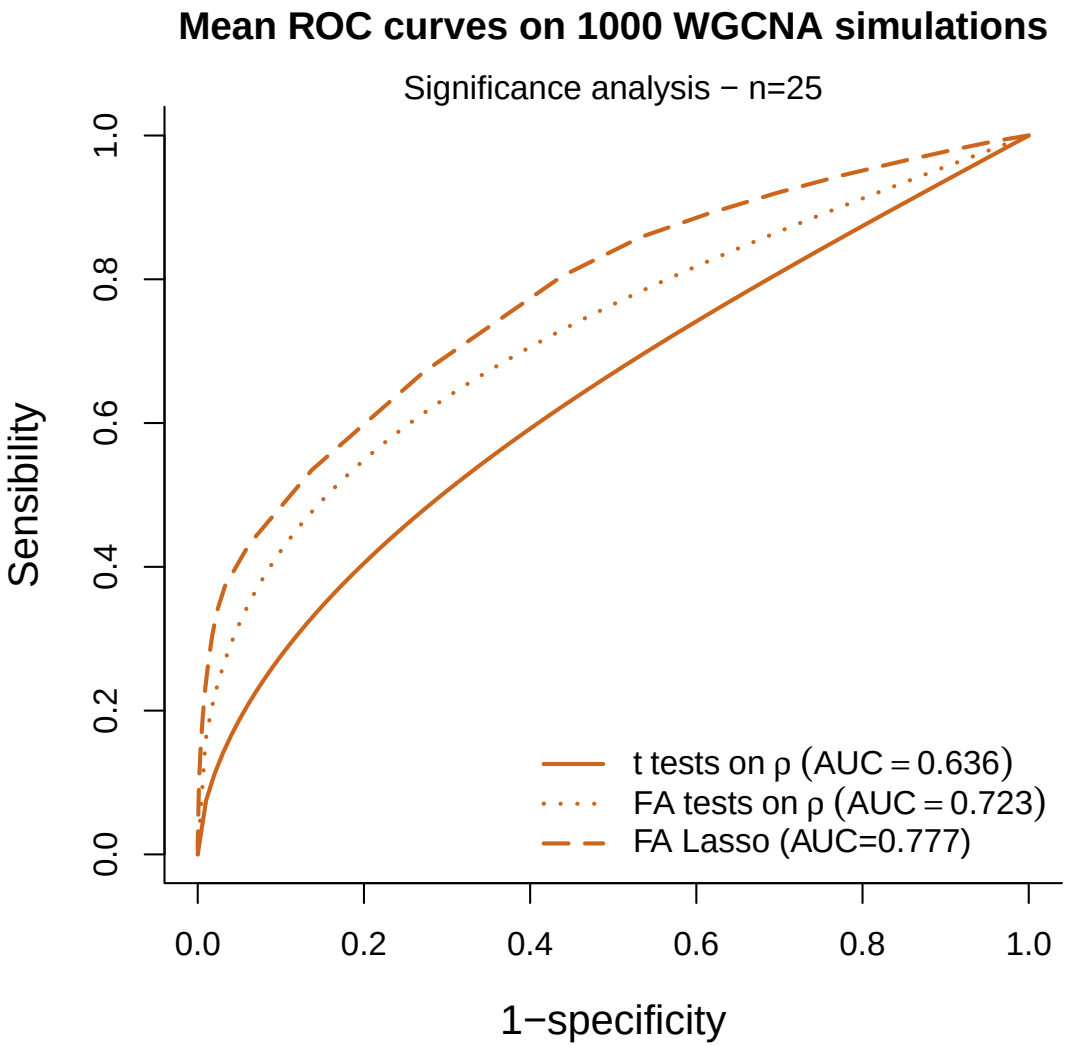
Simulation study



Simulation study



Simulation study

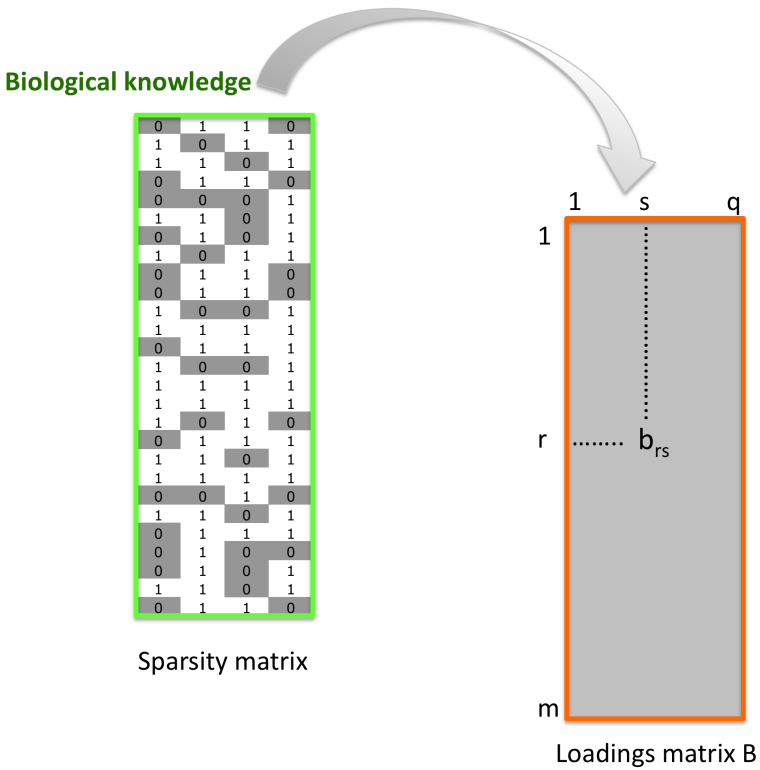


Sparse factor model using biological prior

EM algorithm for sparse factor structure:
M-step: minimizing the expected deviance

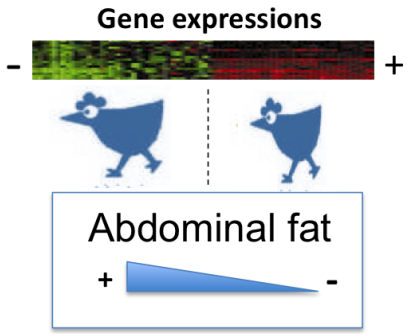
$$\mathcal{D}(\Psi, B; Z) + \lambda' R' vec(B)$$

where λ is the vector of Lagrange multipliers and R the constraints matrix deduced from the sparsity matrix.



Chicken dataset

- 338 annotated genes having their expression correlated to the abdominal fat weight (Blum *et al*, 2010)

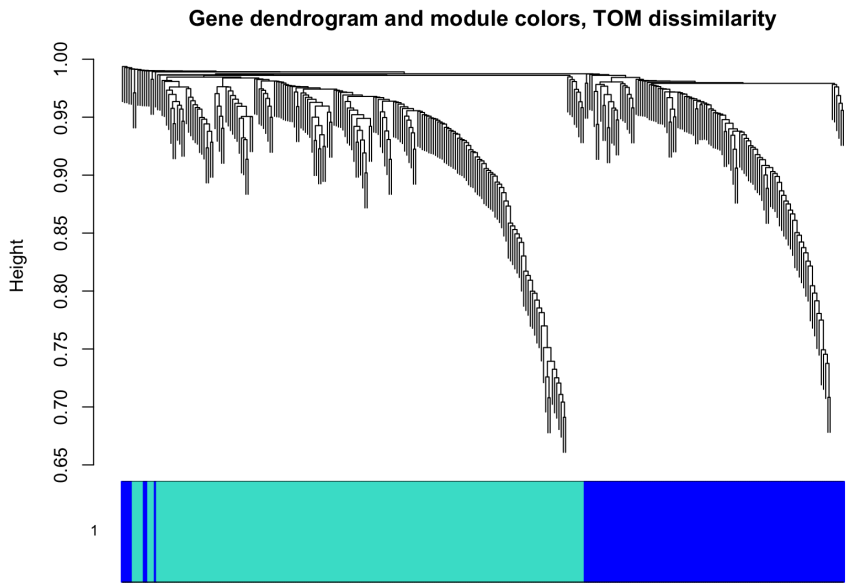


- synthesized GO bp terms as sparsity matrix

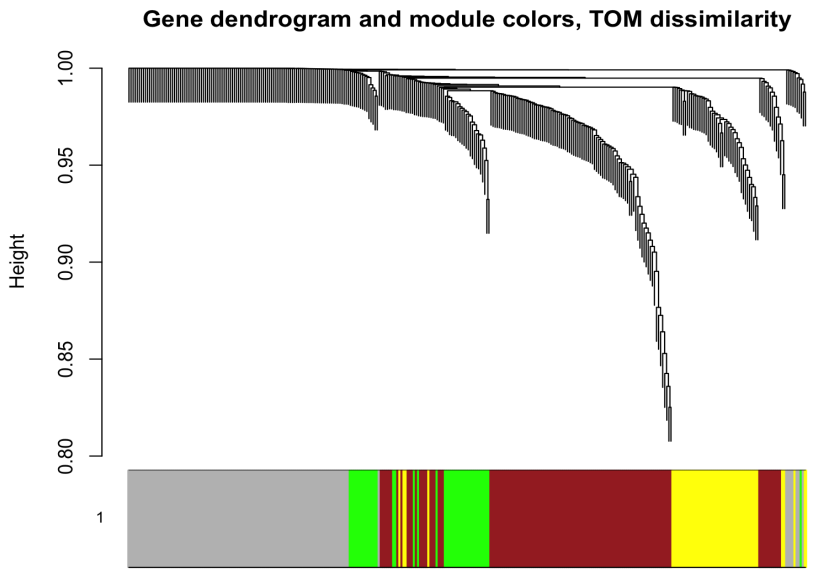
GO biological processes

	Term 1	Term 2	Term 3	...	Term X
gene 1	0	1	1	...	0
gene 2	1	0	0	...	0
gene 3	0	0	1	...	1
⋮	⋮	⋮	⋮	⋮	⋮
gene X	1	0	0	...	0

Gene modules detection using WGCNA



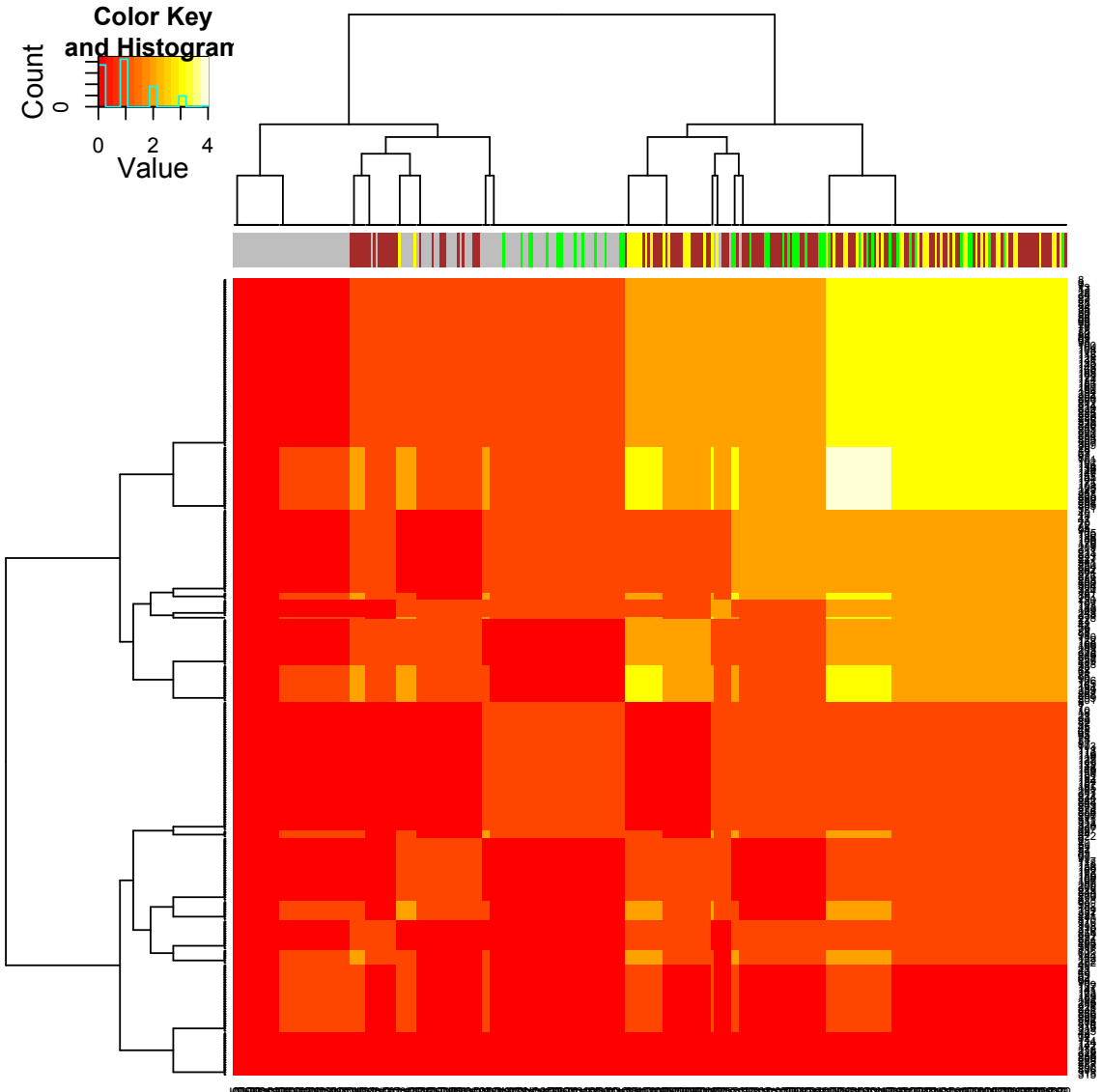
Unrestricted Factor Model



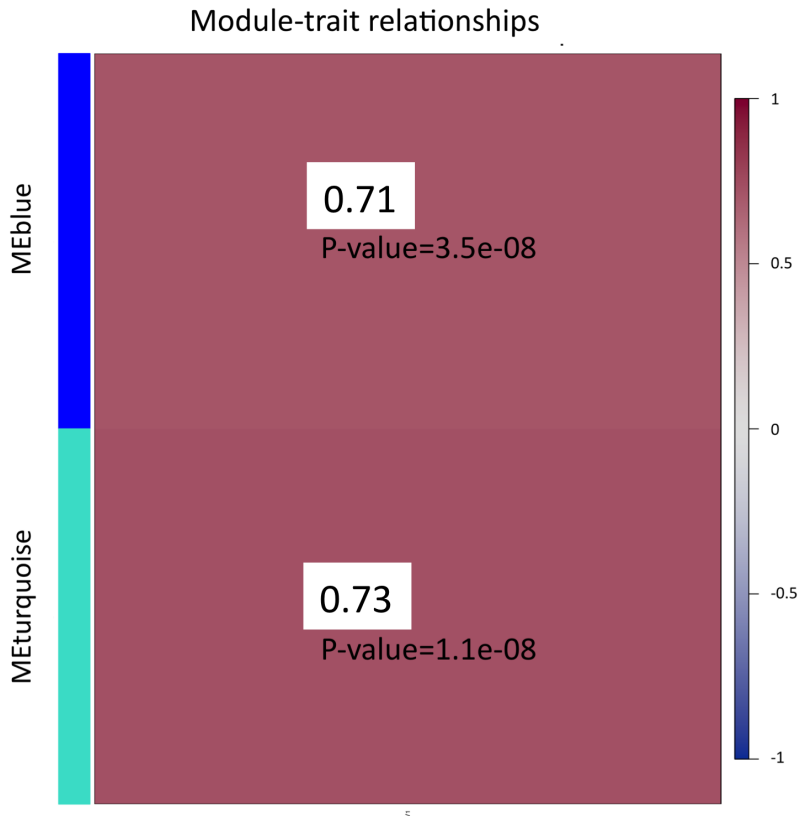
Sparse Factor Model using biological prior

Gene modules detection using WGCNA

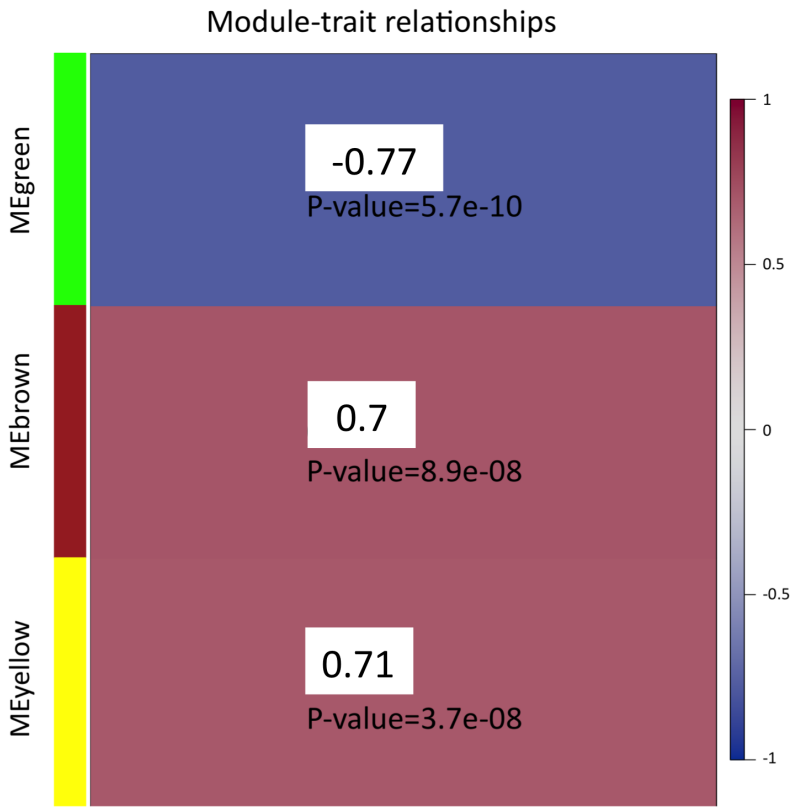
Biological information only.



Gene modules analysis

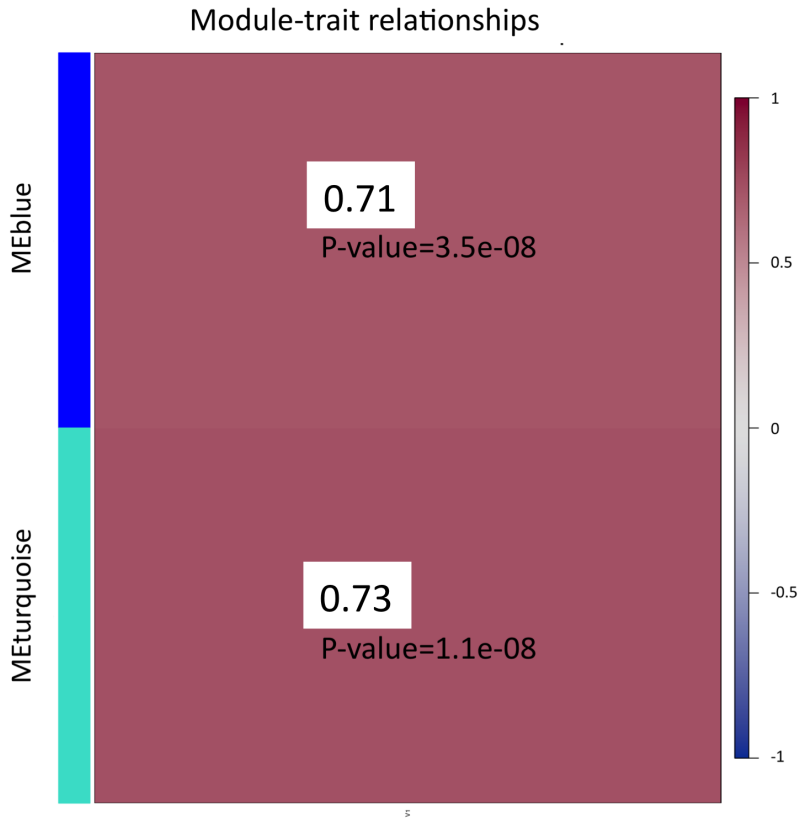


Unrestricted Factor Model

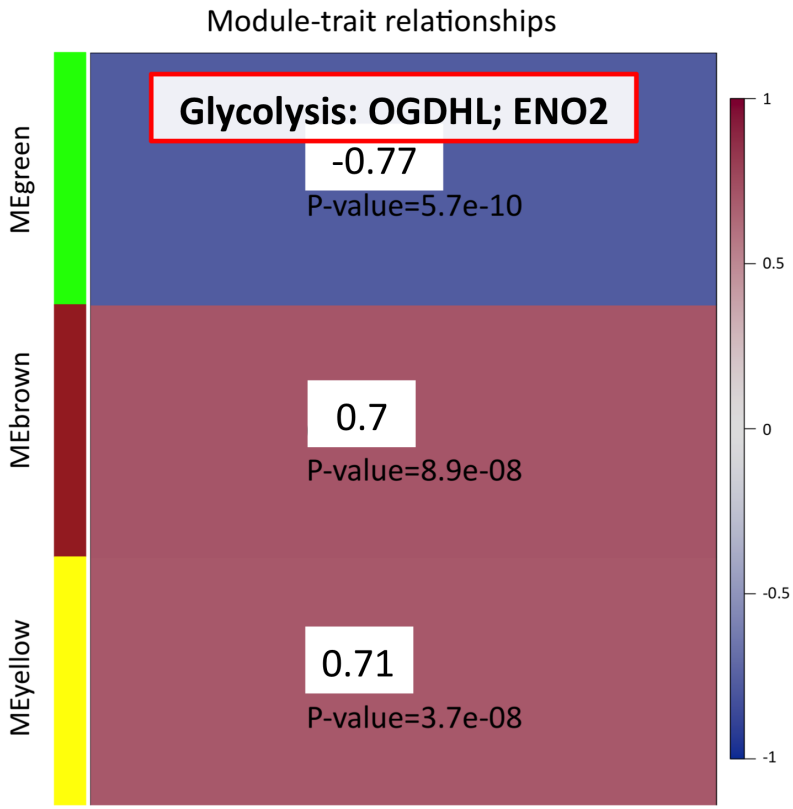


Sparse Factor Model using biological prior

Gene modules analysis



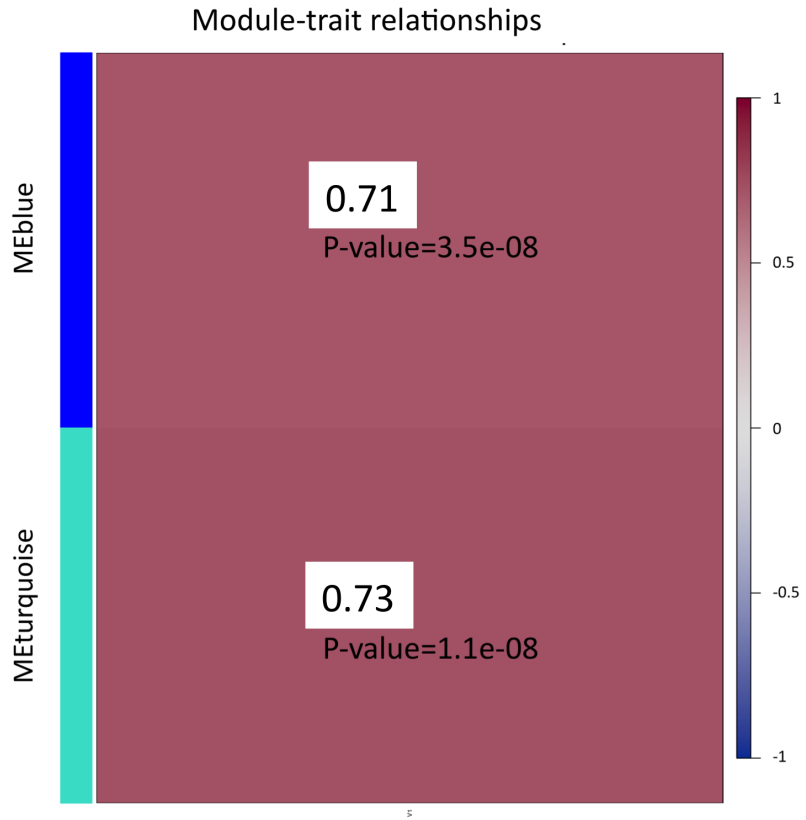
Unrestricted Factor Model



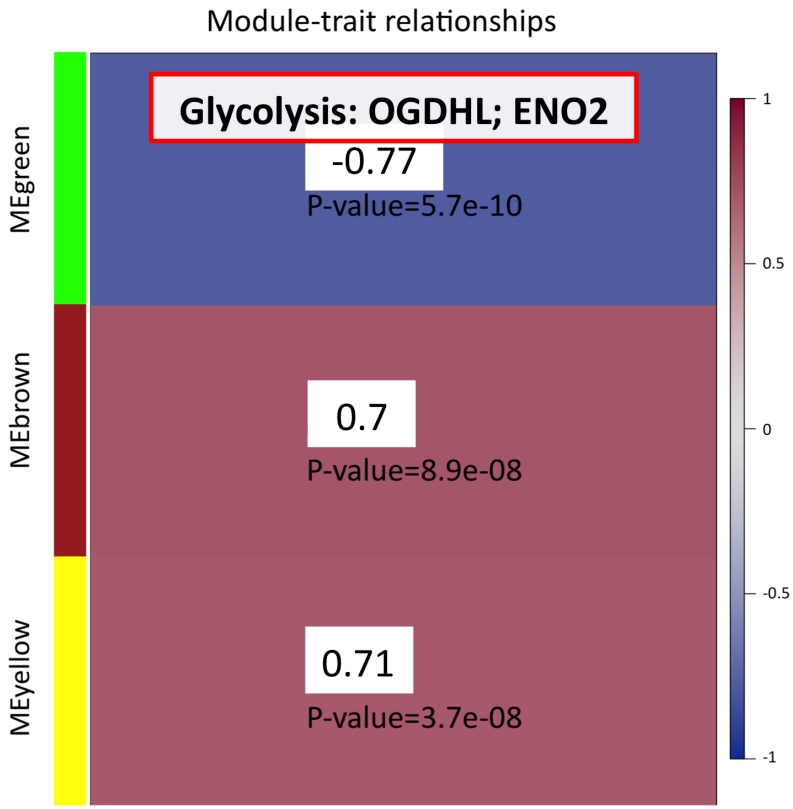
Sparse Factor Model using biological prior

Enhancing hepatic glycolysis reduces obesity (Wu *et al.*, 2005).

Gene modules analysis



Unrestricted Factor Model



Sparse Factor Model using biological prior

Enhancing hepatic glycolysis reduces obesity (Wu *et al.*, 2005).

$\Rightarrow$ biological process and possible key regulators for abdominal fat

Outline

Conclusion/Perspectives

Promising results: improvement of modules detection using a factor model for correlations and introducing sparsity.

In progress:

- deeper investigations on biological prior knowledge integration
- R package implementation

- Gaussian Graphical model: sparse factor model for partial correlation estimation.

New parameterization of the factor model: $\Sigma^{-1} = \varphi(I_m - \theta\theta')\varphi$

$$\begin{aligned} \varphi &= \Psi^{-\frac{1}{2}}, \\ \theta &= \Psi^{-\frac{1}{2}} B (I + B'\Psi^{-1}B)^{-\frac{1}{2}} \end{aligned}$$

ML estimation of (φ, θ) using ML estimation of Ψ and B.

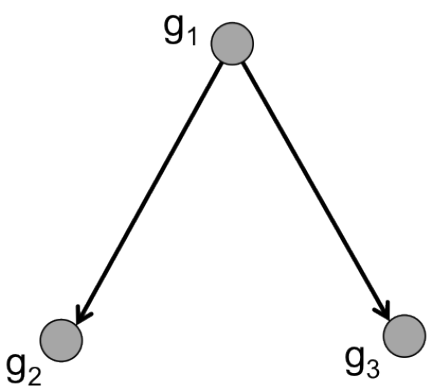
Gaussian Graphical model

Measure of the link:

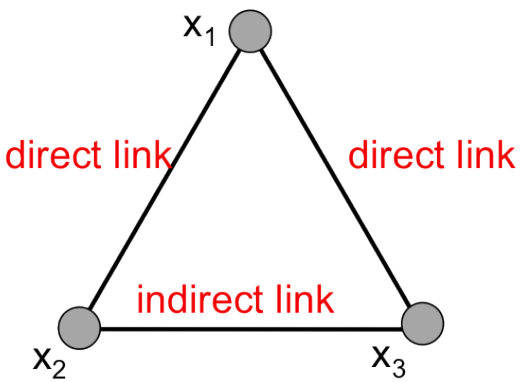
Partial correlation

$$corr(y_i, y_j | y_{\setminus i,j})$$

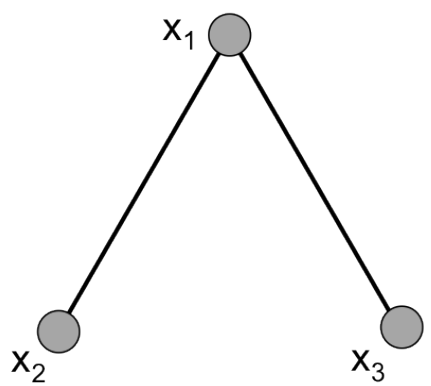
$\Rightarrow$ allows highlighting direct links only



True network

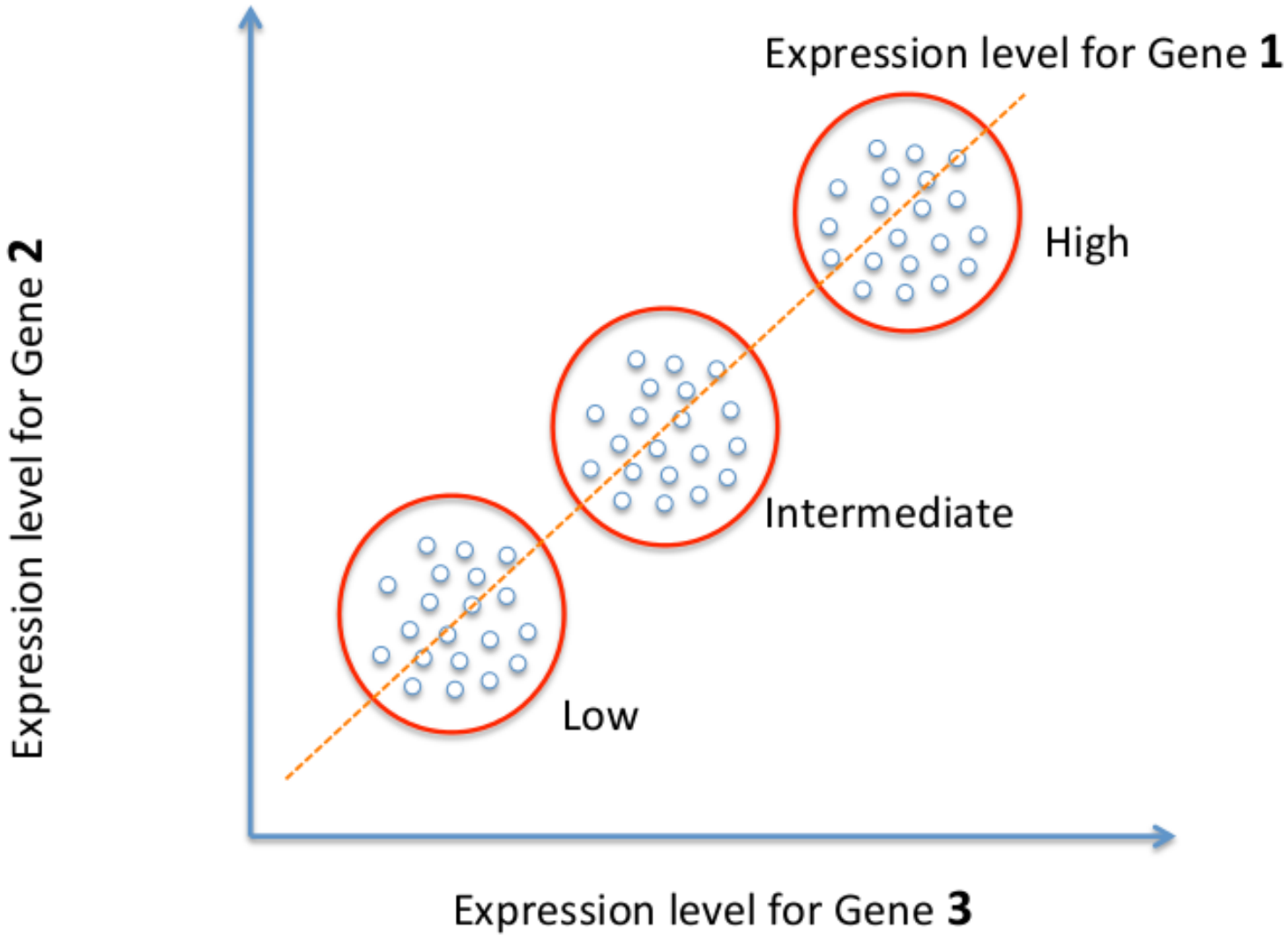


Relevance network



Gaussian Graphical Model

Gaussian Graphical model



Gaussian Graphical model

Y dataset with n rows (individuals) and p columns (genes).
$Y \sim \mathcal{N}_p(\mu, \Sigma)$

The partial correlation matrix $\Pi = (\pi_{i,j})$ is directly linked to the inverse of the variance-covariance matrix as follows:

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

with $\Sigma^{-1} = (\omega_{i,j})$ for $i, j \in [1, p]$

Gaussian Graphical model

Y dataset with n rows (individuals) and p columns (genes).
$Y \sim \mathcal{N}_p(\mu, \Sigma)$

The partial correlation matrix $\Pi = (\pi_{i,j})$ is directly linked to the inverse of the variance-covariance matrix as follows:

$$\pi_{i,j} = \frac{-\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}$$

with $\Sigma^{-1} = (\omega_{i,j})$ for $i, j \in [1, p]$

$\Rightarrow$ Estimation and inversion of Σ. Problem when $p > n$.

Existing methods

- Regularized estimation of the (inverse) covariance matrix: `GeneNet` (Schäfer and Strimmer, 2005)

$$\Sigma_{shrink.} = \lambda T + (1 - \lambda) S$$

where $\lambda \in [0, 1]$ is a tuning parameter, S is the empirical covariance matrix and T a *basic* model for Σ.

Sparsity: significance test on partial correlations.

- Regularized regressions: `SPACE` (Zhu *et al.*, 2009))

$$y_i = \sum_{j \neq i} \beta_{i,j} y_j + \epsilon_i \qquad \pi_{i,j} = sign(\beta_{i,j}) \sqrt{\beta_{i,j} \beta_{j,i}}$$

Sparsity: LASSO penalization

Our proposal

Factor model to estimate partial correlations.

- New parameterization of the factor model: $\Sigma^{-1} = \varphi(I_m - \theta\theta')\varphi$

$$\begin{aligned} \varphi &= \Psi^{-\frac{1}{2}}, \\ \theta &= \Psi^{-\frac{1}{2}} B (I + B'\Psi^{-1}B)^{-\frac{1}{2}} \end{aligned}$$

  ML estimation of (φ, θ) using ML estimation of Ψ and B.

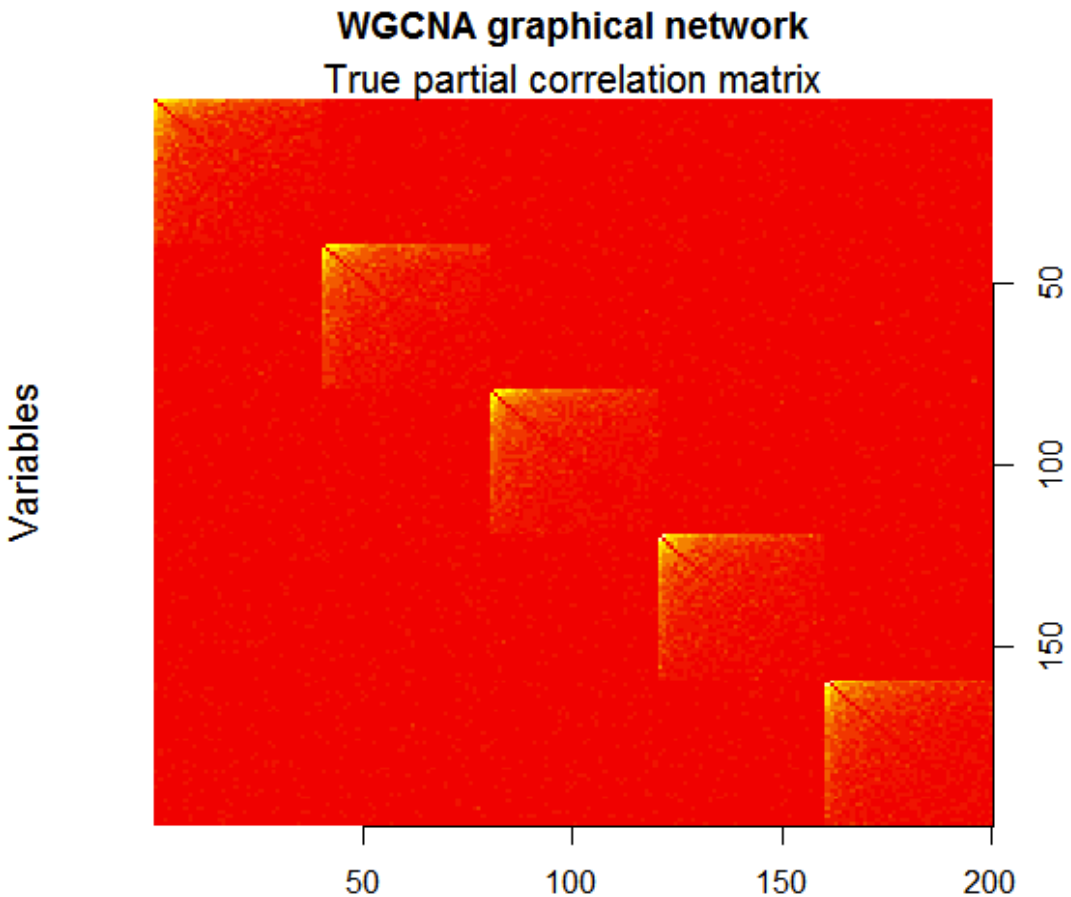
- Sparsity of Σ^{-1}:
  - **Significance testing**: $\theta_{rs} = 0,\ t_{rs} = \hat{\theta}_{rs}/\sqrt{\hat{v}_{rs}}$

$$\sqrt{n}(vec(\hat{\theta}) - vec(\theta)) \sim \mathcal{N}_p(0, V_\theta)$$

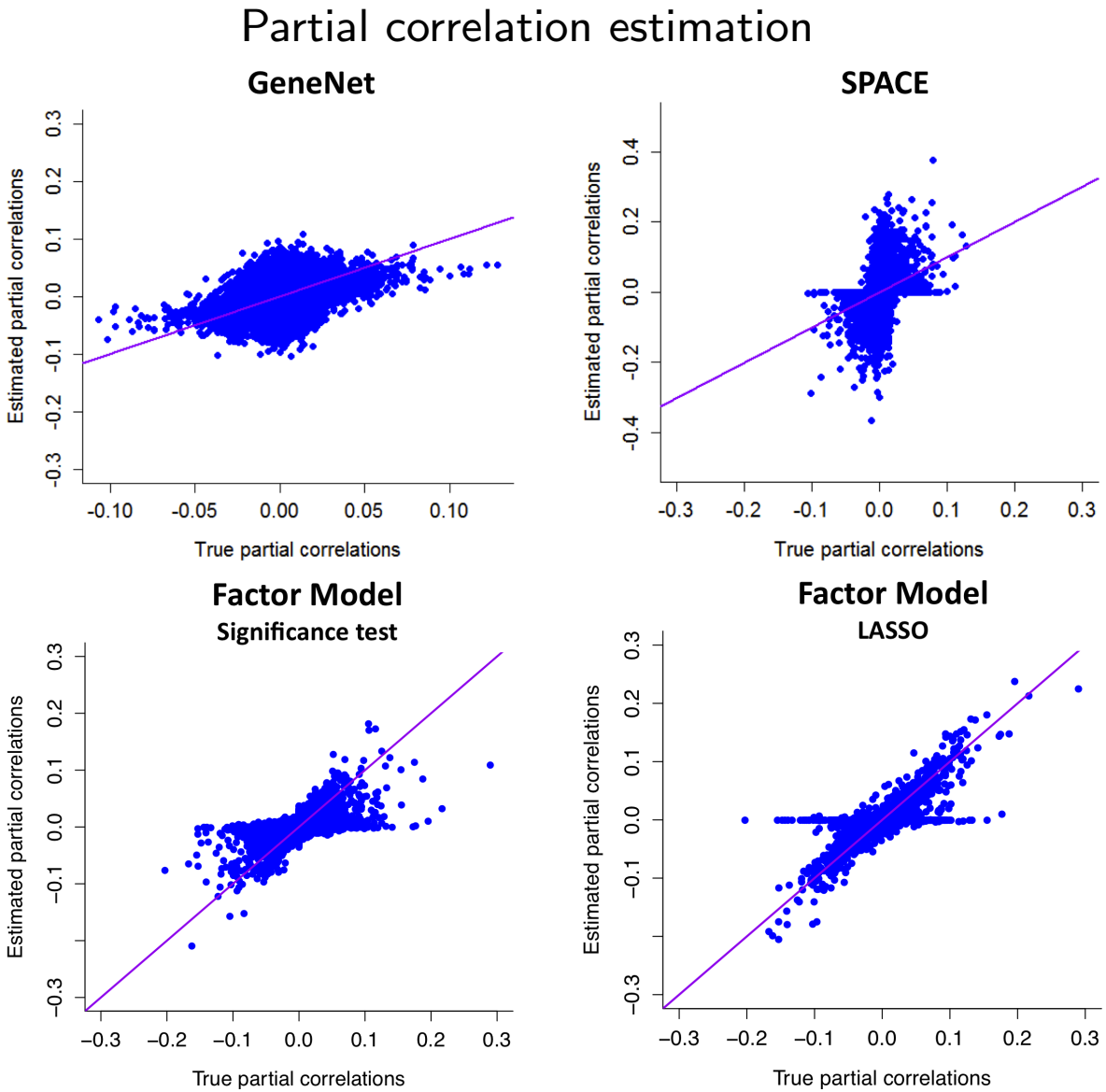
    where V_θ is calculated using the information matrix of the log-likelihood
  - **LASSO estimation**: $\mathcal{D}(\varphi, \theta, \lambda) = \mathcal{D}(\varphi, \theta) + \lambda \sum_{r=1}^{m} \sum_{s=1}^{q} |\theta_{rs}|$ using a CCD algorithm and BIC criteria for choosing λ

Simulated example

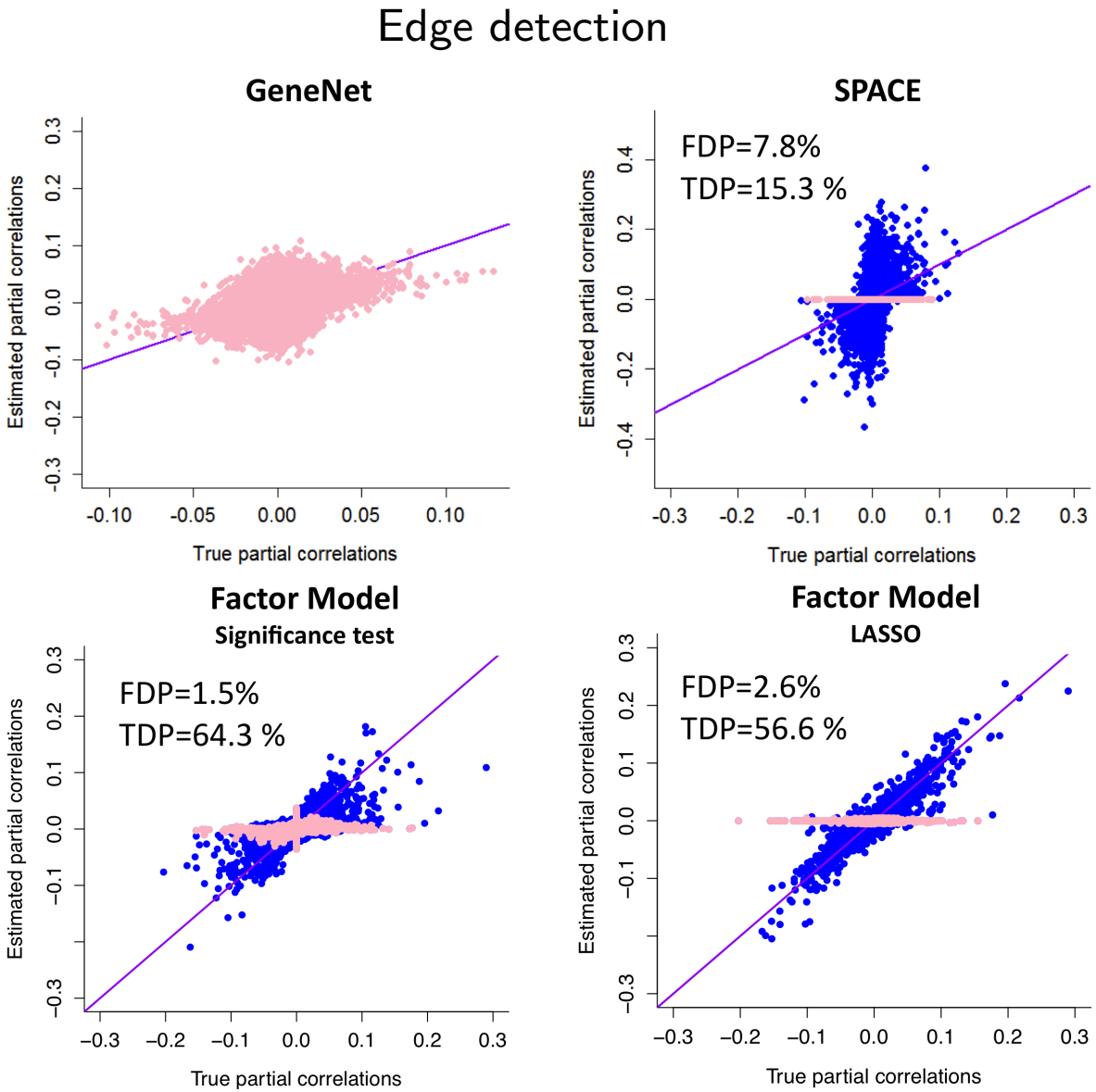
Dataset simulated using the WGCNA package with $m = 200$ and $n = 50$
5 equal modules



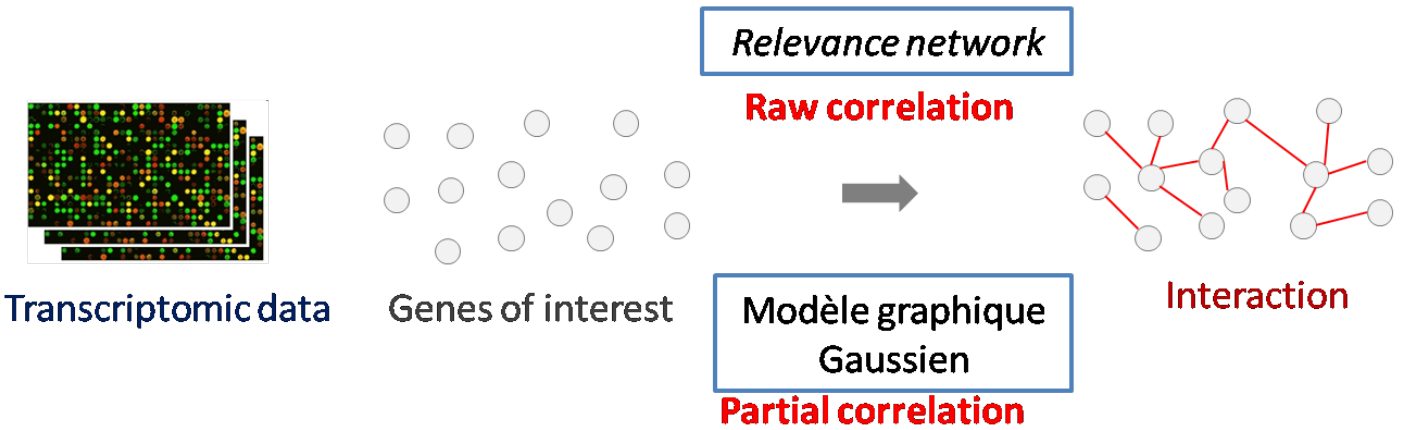
Comparison with GeneNet and SPACE



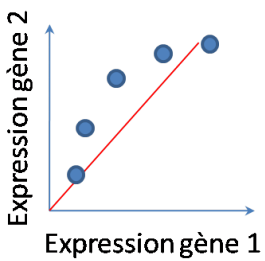
Comparison with GeneNet and SPACE



Concluding comments



Which kind of measure to choose between raw and partial correlations? Linear or non-linear dependence measures ?



Comparative study: by Allen et al. 2012 PLoS ONE (WGCNA/GeneNet/ARACNE/BN)

Comparing Statistical Methods for Constructing LargeScale Gene Networks