A co-expression network analysis reveals homogeneous functional clusters and important genes related to a phenotype of interest

Nathalie Villa-Vialaneix
http://www.nathalievilla.org

with **L. Liaubet, T. Laurent, A. Gamot, P. Cherel & M. SanCristobal**

IUT de Carcassonne (UPVD)

& Institut de Mathématiques de Toulouse

JdS, Gammarth, Tunisie - 26 mai 2011

Outline

Outline

Dataset description

- For 57 pigs,
  1. **2 464 transcript levels** collected by microarray;
  2. **a phenotype of interest: muscle pH** measured 24h post-mortem, related to meat quality.

Dataset description

- For 57 pigs,
  1. **2 464 transcript levels** collected by microarray;
  2. **a phenotype of interest: muscle pH** measured 24h post-mortem, related to meat quality.

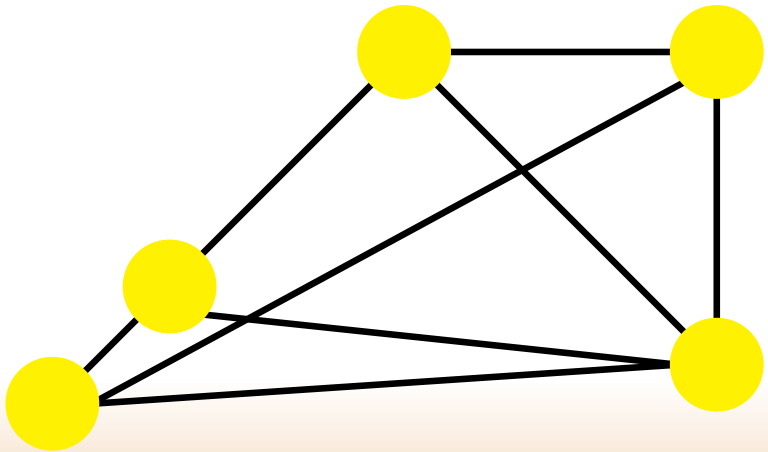
- **272 genes whose expression is (partially) under genetic control** were selected. Among them, only 2 are differentially expressed for pH.

Outline

From genes to genes network

Purpose: Detect and analyze the biological process in its whole.

What is a gene co-expression network?



nodes: genes

edges: “significant” correlation between gene expressions

Correlation and partial correlation

Main issue: Direct calculation of correlations between gene expressions can be biologically irrelevant due to shared correlations.

Correlation and partial correlation

Main issue: Direct calculation of correlations between gene expressions can be biologically irrelevant due to shared correlations.

Use of partial correlations, estimated in a Graphical Gaussian Model framework:

- **H**: gene expressions, X, are distributed as $\mathcal{N}(\mu, \Sigma)$;
- **Quantity to estimate**: Partial correlations, i.e., $\pi_{ij} = \mathrm{Cor}(X^i, X^j | (X^k)_{k \neq i,j})$;

Correlation and partial correlation

Main issue: Direct calculation of correlations between gene expressions can be biologically irrelevant due to shared correlations.

Use of partial correlations, estimated in a Graphical Gaussian Model framework:

- **H**: gene expressions, X, are distributed as $\mathcal{N}(\mu, \Sigma)$;
- **Quantity to estimate**: Partial correlations, i.e., $\pi_{ij} = \mathrm{Cor}(X^i, X^j | (X^k)_{k \neq i,j})$;
- Under **H**, $\pi_{ij} = \frac{-w_{ij}}{\sqrt{w_{ii} w_{jj}}}$ with $\Sigma^{-1} = (w_{ij})_{i,j}$.

Correlation and partial correlation

Main issue: Direct calculation of correlations between gene expressions can be biologically irrelevant due to shared correlations.

Use of partial correlations, estimated in a Graphical Gaussian Model framework:

- **H**: gene expressions, X, are distributed as $\mathcal{N}(\mu, \Sigma)$;
- **Quantity to estimate**: Partial correlations, i.e., $\pi_{ij} = \mathrm{Cor}(X^i, X^j | (X^k)_{k \neq i,j})$;
- Under **H**, $\pi_{ij} = \frac{-w_{ij}}{\sqrt{w_{ii} w_{jj}}}$ with $\Sigma^{-1} = (w_{ij})_{i,j}$.

Another issue: Estimation and inversion of Σ!

Correlation and partial correlation

Main issue: Direct calculation of correlations between gene expressions can be biologically irrelevant due to shared correlations.

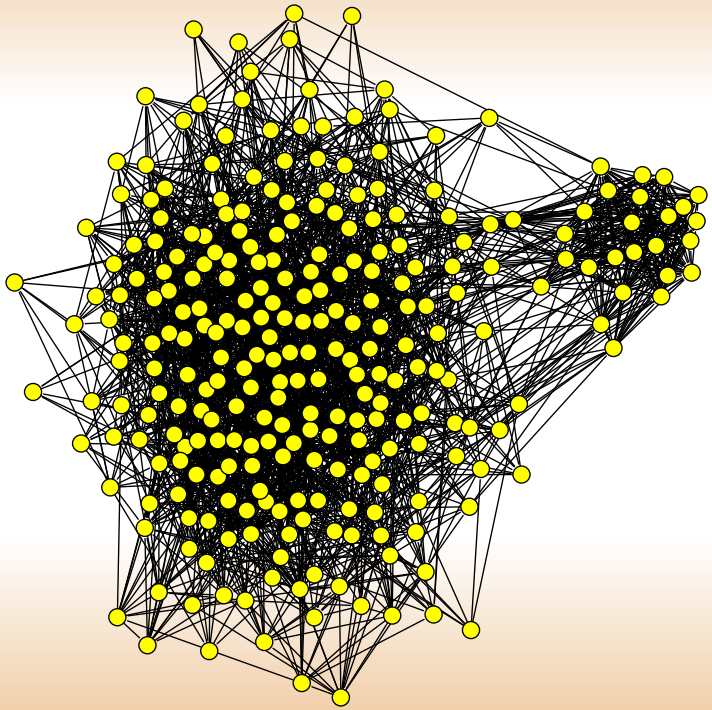
Use of partial correlations, estimated in a Graphical Gaussian Model framework:

- **H**: gene expressions, X, are distributed as $\mathcal{N}(\mu, \Sigma)$;
- **Quantity to estimate**: Partial correlations, i.e., $\pi_{ij} = \mathrm{Cor}(X^i, X^j | (X^k)_{k \neq i,j})$;
- Under **H**, $\pi_{ij} = \frac{-w_{ij}}{\sqrt{w_{ii} w_{jj}}}$ with $\Sigma^{-1} = (w_{ij})_{i,j}$.

Another issue: Estimation and inversion of Σ!

Use of "GeneNet" R package [Schäfer and Strimmer, 2005]: bootstrap estimation and Bayesian approach to test the significativity of the partial correlation.

Basic description of the co-expression network



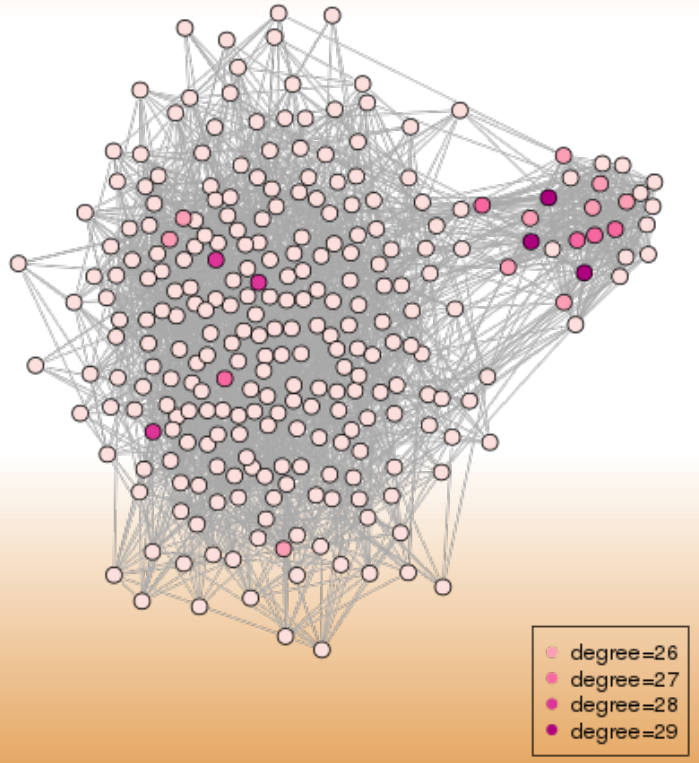
272 nodes (connected graph), density: 6.4 %, transitivity: 25.4 % (hence, probably a modular structure)

Outline

Key genes

Nodes degree: Number of links connected to a given node.

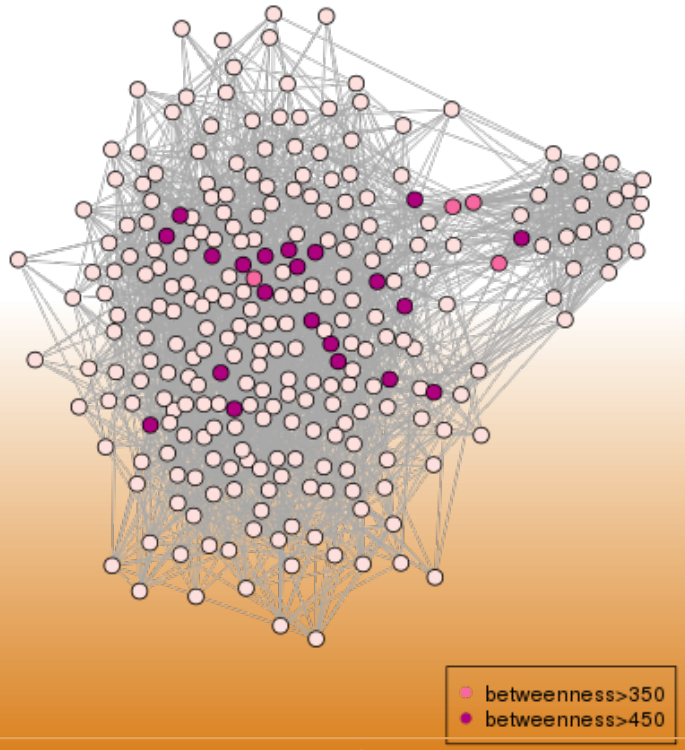
21 hubs (nodes with high degrees) identified



Key genes

Betweenness: Number of shortest paths between two nodes passing by a given node (measure of the importance of a given node to connect the graph)

25 genes with a high betweenness identified



Comments

From these two lists were found:

- **genes already known** to be involved in meat quality (biological validation);
- **annotated genes** that have never been related to meat quality (unexpected genes);
- **unknown genes** (not even annotated).

$\Rightarrow$ unexpected and unknown genes are good candidates for a deeper biological analysis.

Nodes clustering

Purpose: Find groups of genes highly connected to each others (modules) $\Rightarrow$ insights about biological functions and robustness compared to gene-by-gene analysis.

Nodes clustering

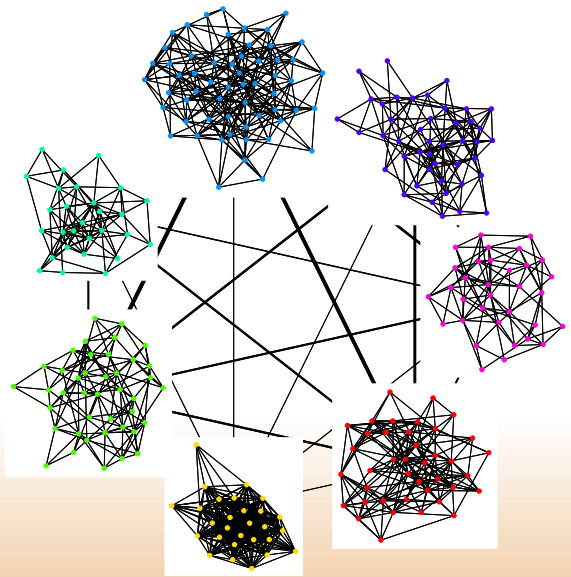
Purpose: Find groups of genes highly connected to each others (modules) $\Rightarrow$ insights about biological functions and robustness compared to gene-by-gene analysis.
Methodology: Modularity optimization

$$Q = \sum_{i,j \in C_k} \left(W_{ij} - \frac{d_i d_j}{2m} \right)$$

with $d_i = \sum_l W_{il}$ and $m = \frac{1}{2} \sum_i d_i$ **[Newman and Girvan, 2004]** by simulated annealing (after a previous comparison of several methods and parameters).

Results

- **7 clusters** with 28 to 58 nodes;



Results

- **7 clusters** with 28 to 58 nodes;
- For annotated genes, at least 68% (but often more than 80%) of a given cluster, **belong to a single IPA network** (literature validation) $\Rightarrow$ assumptions for functions of unknown genes.

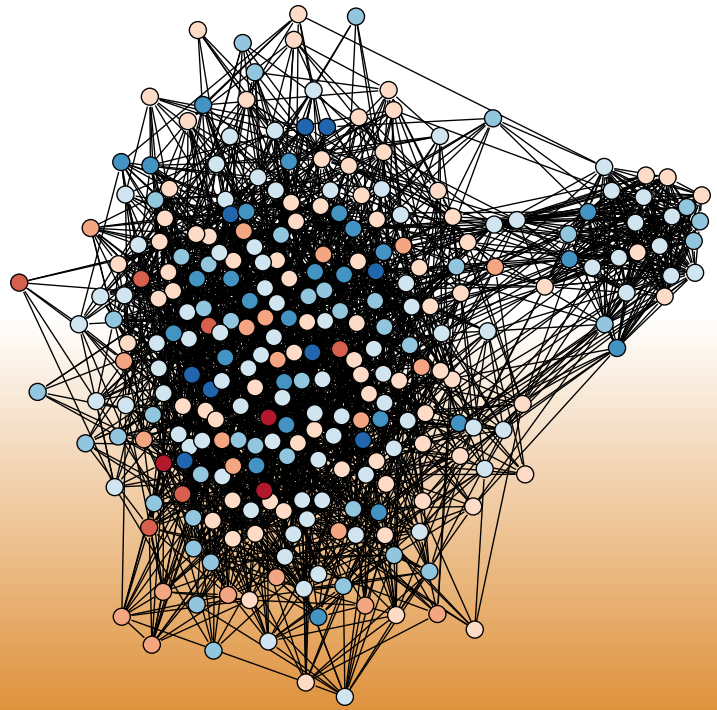
Superimposing additional information on nodes

Purpose: Study of the correlation between the network topology and a phenotype of interest.

Superimposing additional information on nodes

Purpose: Study of the correlation between the network topology and a phenotype of interest.

Data: Partial correlation between pH and gene expression.



Fist analysis of the relations between pH and network topology

[Laurent and Villa-Vialaneix, 2011]

- **Cluster 4** has a significatively higher correlation with pH than the other clusters;

Fist analysis of the relations between pH and network topology

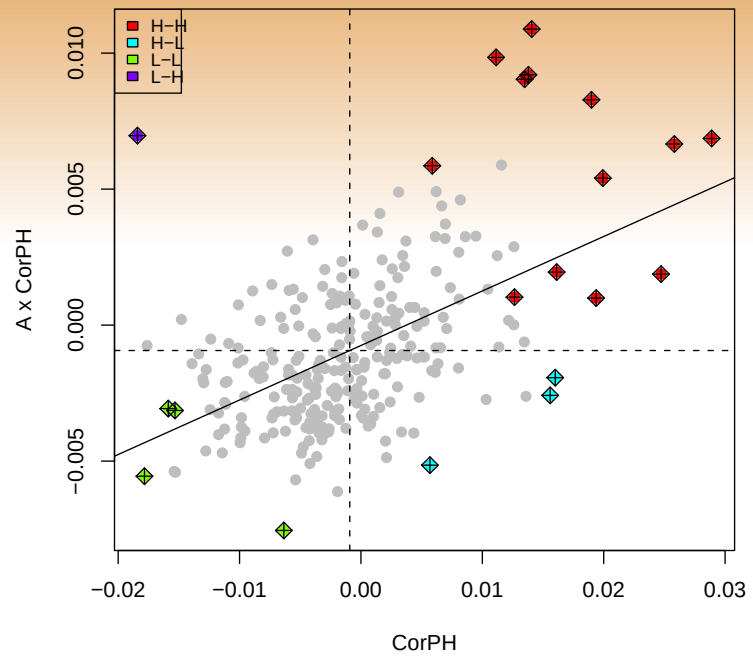
[Laurent and Villa-Vialaneix, 2011]

- **Cluster 4** has a significatively higher correlation with pH than the other clusters;
- Network auto-correlation of the labels can be assessed through Moran's I

$$I = \frac{\frac{1}{2m}\sum_{i\neq j} w_{ij}\bar{c}_i\bar{c}_j}{\frac{1}{n}\sum_i \bar{c}_i^2}$$

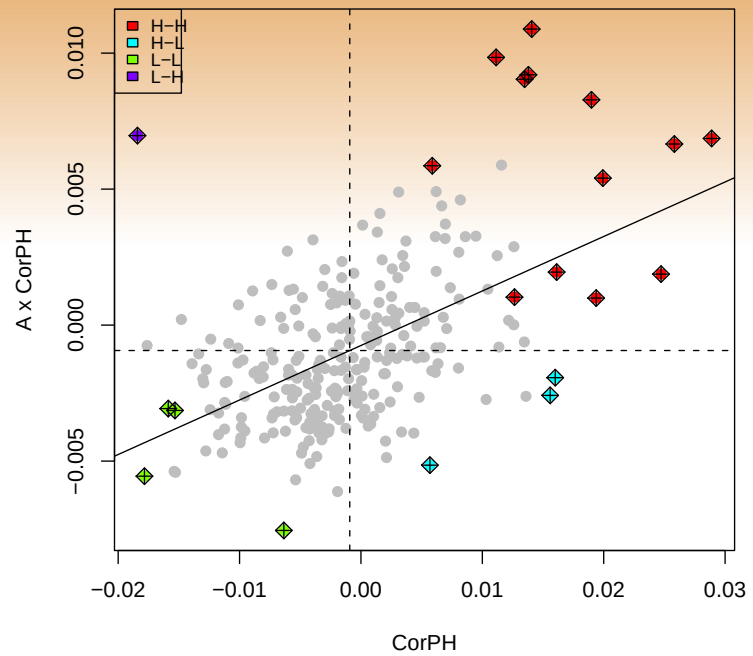
where $m = \frac{1}{2}\sum_{i\neq j} w_{ij}$, c_i is the label of node i (partial correlation with pH) and $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$.
Moran's I is significatively large: **network topology is related to partial correlation with pH**.

Moran's plot



Average values for partial correlation with pH in the neighborhood of a node in function of the partial correlation with pH for this node.

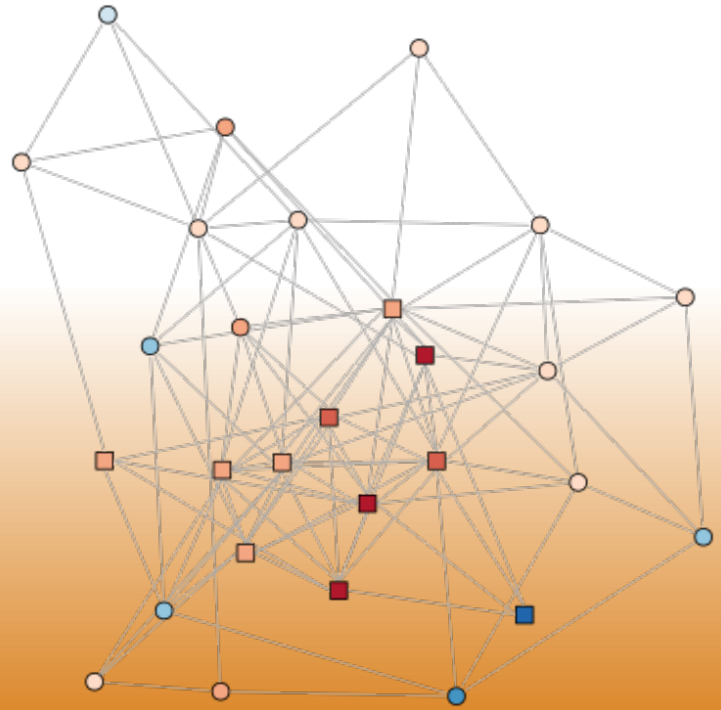
Moran's plot



Average values for partial correlation with pH in the neighborhood of a node in function of the partial correlation with pH for this node.
Influential nodes can be extracted: most belong to cluster 4 and some of them are already known to be involved in meat quality (e.g., one gene involved in glycolysis and gluconeogenesis).

Moran's plot

Influential nodes can be extracted: most belong to cluster 4 and some of them are already known to be involved in meat quality (e.g., one gene involved in glycolysis and gluconeogenesis).



Conclusion

Statistical approaches were **validated by biological knowledge**:

- clusters are consistent with the literature (homogeneous biological functions);
- some key genes extracted from the analysis of the partial correlation with pH are already known to be involved in meat quality.

Conclusion

Statistical approaches were **validated by biological knowledge**:

- clusters are consistent with the literature (homogeneous biological functions);
- some key genes extracted from the analysis of the partial correlation with pH are already known to be involved in meat quality.

Hence, **unexpected statistical results could provide insights about unknown genes** (their function, that they are possibly involved in meat quality...)

Laurent, T. and Villa-Vialaneix, N. (2011).
Using spatial indexes for labeled network analysis.
Information, Interaction, Intelligence (i3).
Under revision.

Newman, M. and Girvan, M. (2004).
Finding and evaluating community structure in networks.
Physical Review, E, 69:026113.

Schäfer, J. and Strimmer, K. (2005).
An empirical bayes approach to inferring large-scale gene association networks.
Bioinformatics, 21(6):754–764.