Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach.

Marine Jeanmougin

JOBIM 2012, Rennes – July 4th, 2012

Outline

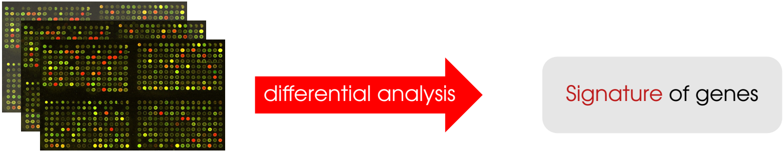
1. Introduction
   - Microarray experiments
   - Identification of molecular signatures
   - Motivations

2. DIsease Associated Module Selection (DiAMS)
   - Global approach
   - Local-score statistic for module ranking
   - Evaluation process

3. Results and application
   - Quantitative results
   - Application to Estrogen Receptor status in breast cancer

Microarray experiments

Objectives of microarray experiments



Expression level of thousands of transcripts

differential analysis

Signature of genes

Biological purpose

- Signature: genes involved in a phenotype of interest
- Medical applications: diagnosis, prognosis, treatment efficacy

Identification of molecular signatures

Differential analysis

Model

$X_{ig}^{(c)}$:**expression level** of the *i*th sample for gene *g* under condition *c* such as:

$$\mathbb{E}(X_{ig}^{(c)}) = \mu_g^{(c)}$$

Under the assumption of homoscedasticity between conditions:

$$\mathbb{V}(X_{ig}^{(c)}) = (\sigma_g)^2$$

Hypothesis testing strategy

For two conditions, the null hypothesis to test comes down to

$$\begin{cases} H_{0,g}: & \mu_g^{(1)} = \mu_g^{(2)} \\ H_{1,g}: & \mu_g^{(1)} \neq \mu_g^{(2)} \end{cases}$$

▷ Classical approach: *t*-statistic

Issues for gene-specific variance estimation

Identification of molecular signatures

Differential analysis

Model

$X_{ig}^{(c)}$:**expression level** of the *i*th sample for gene g under condition c such as:

$$\mathbb{E}(X_{ig}^{(c)}) = \mu_g^{(c)}$$

Under the assumption of homoscedasticity between conditions:

$$\mathbb{V}(X_{ig}^{(c)}) = (\sigma_g)^2$$

Hypothesis testing strategy

For two conditions, the null hypothesis to test comes down to

$$\begin{cases} H_{0,g} : & \mu_g^{(1)} = \mu_g^{(2)} \\ H_{1,g} : & \mu_g^{(1)} \neq \mu_g^{(2)} \end{cases}$$

▷ Classical approach: *t*-statistic

Issues for gene-specific variance estimation

Model

$X_{ig}^{(c)}$:**expression level** of the *i*th sample for gene g under condition c such as:

$$\mathbb{E}(X_{ig}^{(c)}) = \mu_g^{(c)}$$

Under the assumption of homoscedasticity between conditions:

$$\mathbb{V}(X_{ig}^{(c)}) = (\sigma_g)^2$$

Hypothesis testing strategy

For two conditions, the null hypothesis to test comes down to

$$\begin{cases} H_{0,g}: & \mu_g^{(1)} = \mu_g^{(2)} \\ H_{1,g}: & \mu_g^{(1)} \neq \mu_g^{(2)} \end{cases}$$

▷ Classical approach: *t*-statistic

Issues for gene-specific variance estimation

Identification of molecular signatures

Differential analysis

Limma: a shrinkage approach (Smyth, 2004)

Jeanmougin *et al.* 2010, *PLoS ONE*

Empirical Bayes variance estimate

$$S_g^{\text{limma}} = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g},$$

- S_0^2: *prior* variance from the scale-inverse-chi-square distribution
  $\leadsto$ fixed with an empirical Bayes approach
- S_g^2: usual unbiased estimator of the variance $(\sigma_g)^2$
- d_0, d_g: residual degrees of freedom for S_0^2 and for the linear model for gene g

Test statistic:

$$t_g^{\text{limma}} = \frac{\bar{x}_{\cdot g}^{(1)} - \bar{x}_{\cdot g}^{(2)}}{S_g^{\text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Motivations

Limitations of classical approaches

- Low reproducibility

  Ein-Dor *et al.* 2005, Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*

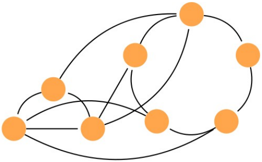
- Difficulty to achieve a clear biological interpretation

Improving gene signatures

- Genes causing the same phenotype are likely to interact together

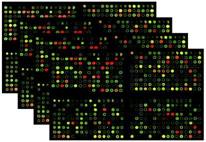
  Gandhi, T.K. *et al.* 2006, *Nature Genetics*

- Identification of genes that are functionally related (i.e. modules)



Functional relationship network

+



Expression data

Outline

1. Introduction
   - Microarray experiments
   - Identification of molecular signatures
   - Motivations

2. DIsease Associated Module Selection (DiAMS)
   - Global approach
   - Local-score statistic for module ranking
   - Evaluation process

3. Results and application
   - Quantitative results
   - Application to Estrogen Receptor status in breast cancer

Global approach

Goal

Select functional modules presenting unexpected accumulation of high-scoring genes

Input parameters

- PPI network (strong manifestation of functional relations)
- Gene scores from limma statistic

DiAMS: a 3-step process

1. Preprocessing
2. Local-score approach for module ranking
3. Selection of significant modules

Global approach

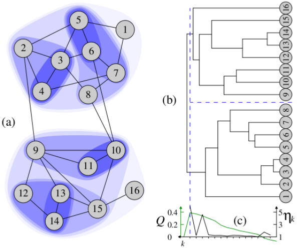
Step 1 - Preprocessing

High-dimensional network

- Impossibility of exploring the huge space of possible gene subnetworks

Hierarchical clustering

- Captures much information about network topology
- Enables to go easily through the structure
- Screen the entire network without constraints on module sizes



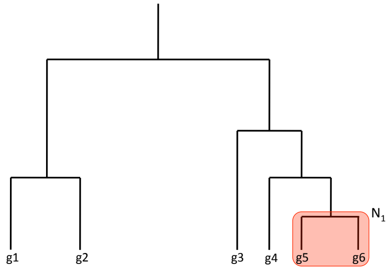
"Walktrap" approach

- Random walks strategy
- Distance (similarity measure of vertices)
- Ward's criterion

Pons and Latapy 2006 *JGAA*

Global approach

Step 2 - Local-score approach for module ranking

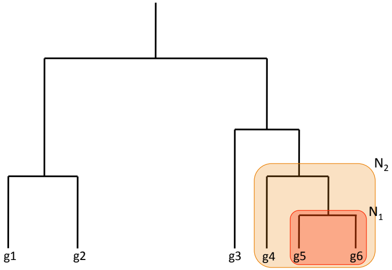


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

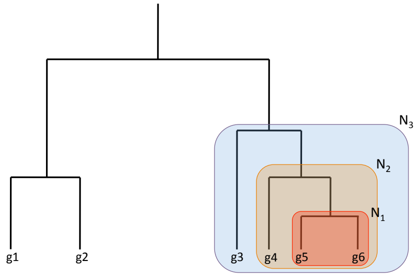


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

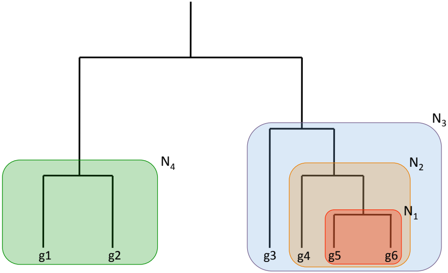


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

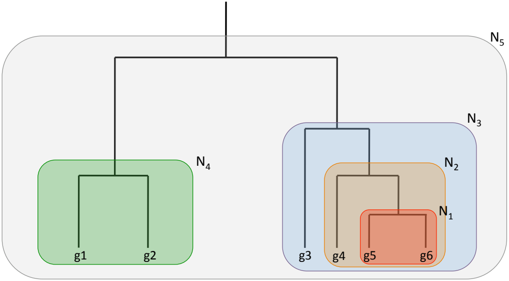


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

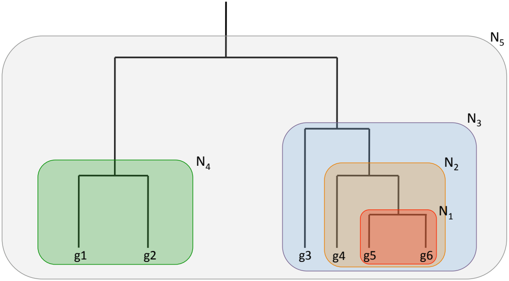


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

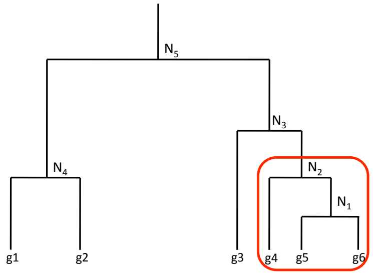


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking

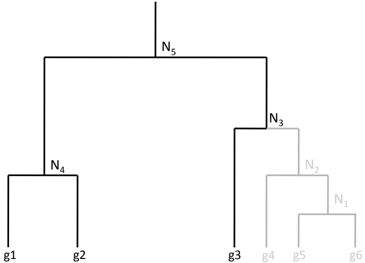


Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

Global approach

Step 2 - Local-score approach for module ranking



Iterative module ranking

1. Score each module N_k (by summing gene scores)
2. Identify the highest scoring module (local-score statistic)
3. Remove it
4. Repeat setps 1) to 3) until all disjoint modules have been enumerated

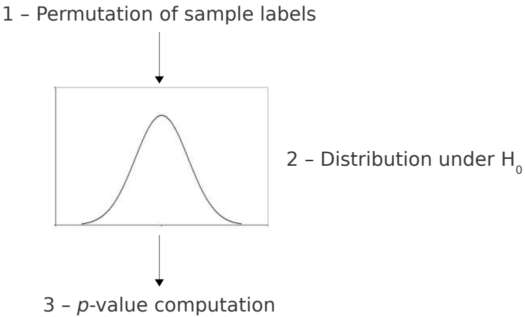
Global approach

Step 3 - Selection of significant modules

Goal

Assess the global significance of each module

Monte-Carlo approach



1 – Permutation of sample labels

2 – Distribution under H_0

3 – *p*-value computation

⇝ Selection of modules at 5% FDR level.

Outline

1. Introduction
   - Microarray experiments
   - Identification of molecular signatures
   - Motivations
2. DIsease Associated Module Selection (DiAMS)
   - Global approach
   - Local-score statistic for module ranking
   - Evaluation process
3. Results and application
   - Quantitative results
   - Application to Estrogen Receptor status in breast cancer

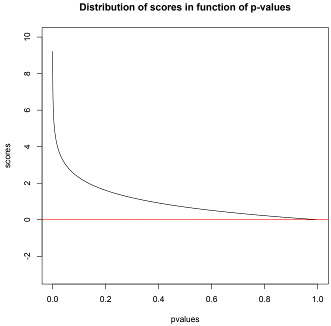
Module scoring

Individual gene scoring

The gene score is given by:

$$\nu_g = -\log(p_g) - \delta,$$

- p_g: gene p-value from limma,
- δ, a constant such as $\mathbb{E}(\nu_g) \leq 0$.



Local-score statistic

***Definition**: value of the highest-scoring module.*

Given $\mathcal{H}$, a hierarchical community structure, the local-score statistic is defined as:

$$L = \max_{H \subseteq \mathcal{H}} \left(\sum_{g \in H} \nu_g \right),$$

such as H is a subtree of $\mathcal{H}$.

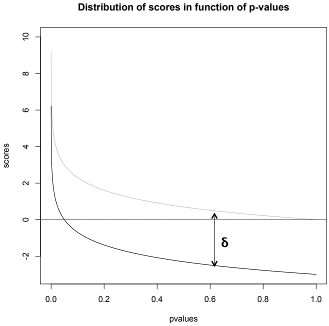
Module scoring

Individual gene scoring

The gene score is given by:

$$\nu_g = -\log(p_g) - \delta,$$

- p_g: gene p-value from limma,
- δ, a constant such as $\mathbb{E}\left(\nu_g\right) \leq 0$.


Distribution of scores in function of p-values

Local-score statistic

***Definition**: value of the highest-scoring module.*

Given $\mathcal{H}$, a hierarchical community structure, the local-score statistic is defined as:

$$L = \max_{H \subseteq \mathcal{H}} \left(\sum_{g \in H} \nu_g \right),$$

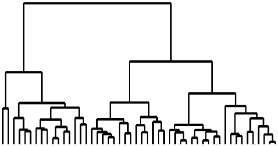
such as H is a subtree of $\mathcal{H}$.

Outline

1. Introduction
   - Microarray experiments
   - Identification of molecular signatures
   - Motivations
2. DIsease Associated Module Selection (DiAMS)
   - Global approach
   - Local-score statistic for module ranking
   - Evaluation process
3. Results and application
   - Quantitative results
   - Application to Estrogen Receptor status in breast cancer

Evaluation process

Power and false-postive rate study

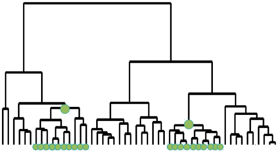


Tree structure

Evaluation process

Power and false-postive rate study

1. Simulation of significant nodes

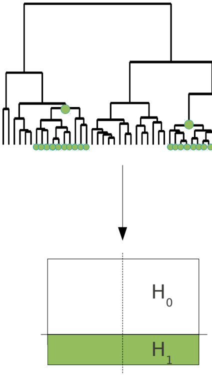


Tree structure

Evaluation process

Power and false-postive rate study

1. Simulation of significant nodes



Tree structure

Gene expression matrix

2. Simulation of the gene expression matrix

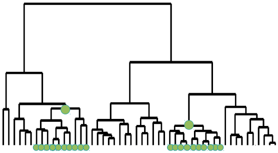
$$\begin{cases} X_{ig}^{(1)} & \sim \mathcal{N}(\mu_g^{(1)}, \sigma^2) \\ X_{ig}^{(2)} & \sim \mathcal{N}(\mu_g^{(2)}, \sigma^2) \end{cases}$$

$$\begin{aligned} H_{0,g}: &\quad \mu_g^{(1)} = \mu_g^{(2)} \\ H_{1,g}: &\quad \mu_g^{(2)} = \mu_g^{(1)} + \Delta, \text{ with } \Delta \text{ in } \{0.5, ..., 3\} \end{aligned}$$

Evaluation process

Power and false-postive rate study

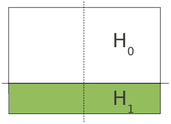
1. Simulation of significant nodes



Tree structure

2. Simulation of the gene expression matrix

$$\begin{cases} X_{ig}^{(1)} & \sim \mathcal{N}(\mu_g^{(1)}, \sigma^2) \\ X_{ig}^{(2)} & \sim \mathcal{N}(\mu_g^{(2)}, \sigma^2) \end{cases}$$



Gene expression matrix

$$\begin{aligned} H_{0,g}: &\quad \mu_g^{(1)} = \mu_g^{(2)} \\ H_{1,g}: &\quad \mu_g^{(2)} = \mu_g^{(1)} + \Delta, \text{ with } \Delta \text{ in } \{0.5, ..., 3\} \end{aligned}$$

3. Power and False-Positive (FP) rate evaluation

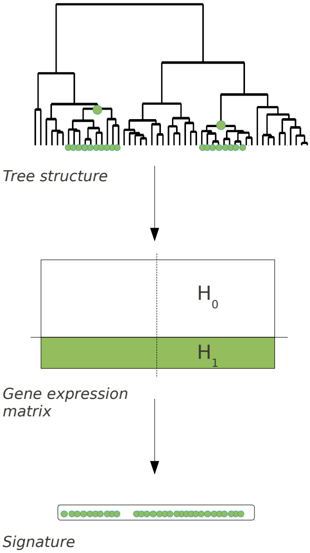
Signature

$$\text{Power}_\alpha = \mathbb{P}_{H_1}(H_0 \text{ rejected at the } \alpha \text{ level})$$

$$\mathbb{P}(\text{FP})_\alpha = \mathbb{P}_{H_0}(H_0 \text{ rejected at the } \alpha \text{ level})$$

Evaluation process

Reproducibility



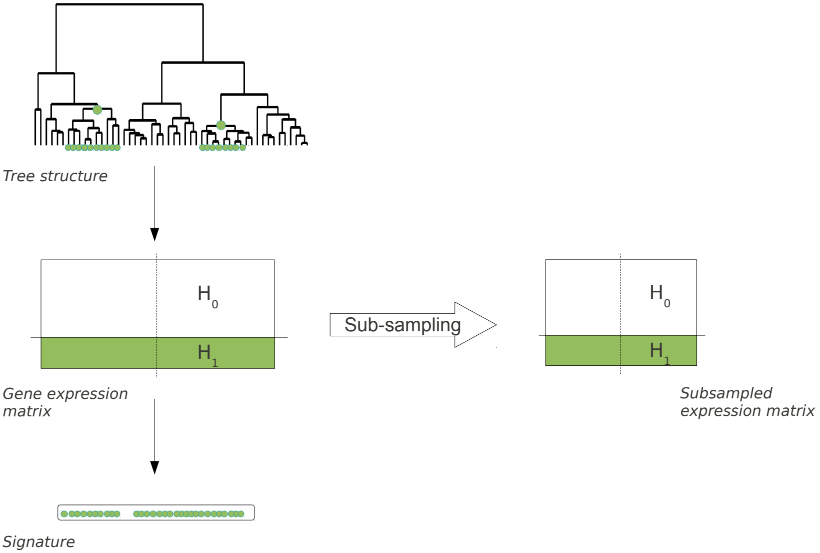
Tree structure

Gene expression matrix

Signature

Evaluation process

Reproducibility



Tree structure

Gene expression matrix

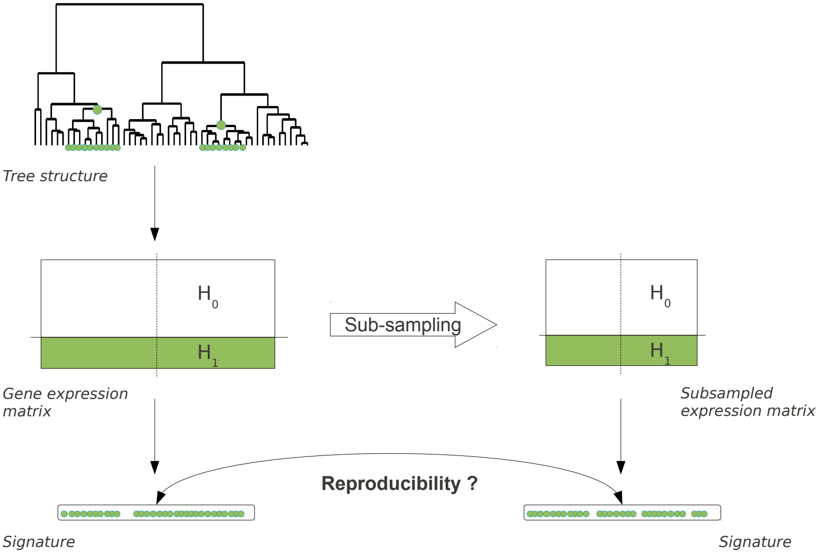
Sub-sampling

Subsampled expression matrix

Signature

Evaluation process

Reproducibility



Tree structure

H_0

H_1

Gene expression matrix

Sub-sampling

H_0

H_1

Subsampled expression matrix

Reproducibility ?

Signature

Signature

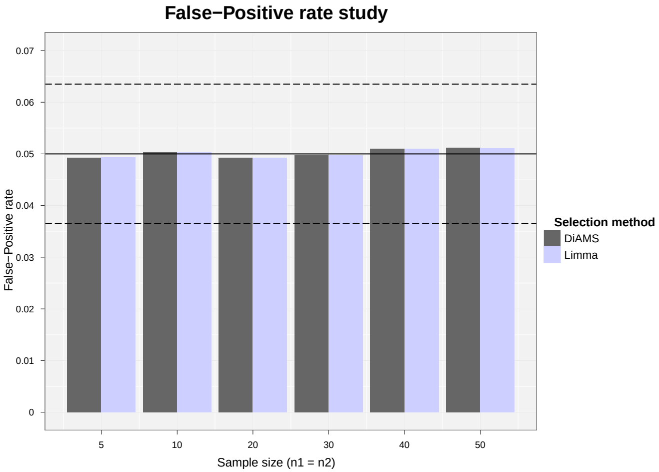
Outline

1. Introduction
   - Microarray experiments
   - Identification of molecular signatures
   - Motivations

2. DIsease Associated Module Selection (DiAMS)
   - Global approach
   - Local-score statistic for module ranking
   - Evaluation process

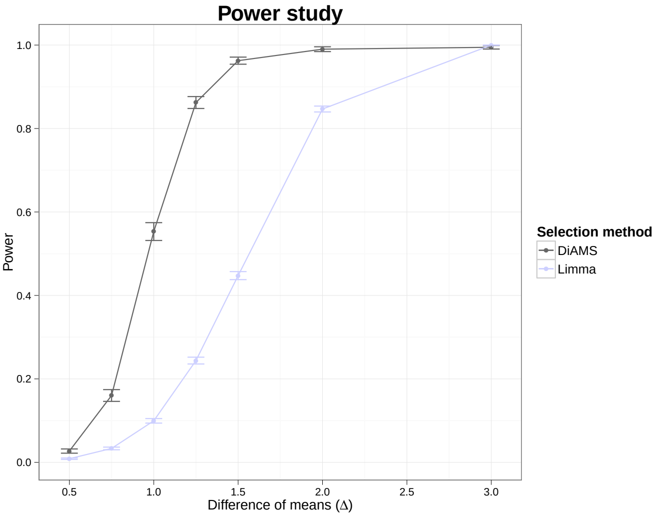
3. Results and application
   - Quantitative results
   - Application to Estrogen Receptor status in breast cancer

Quantitative results



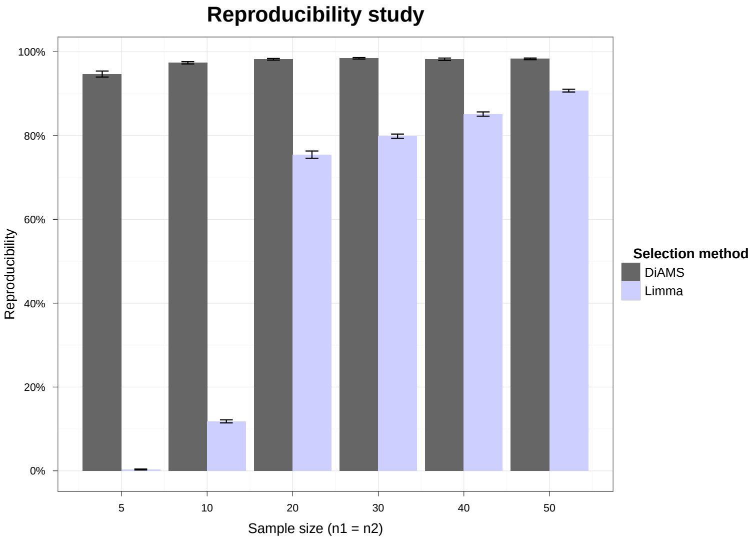
False-positive rate study - Estimated false-positive rate over the 1,000 simulations. Plain black line: the 5% level. The dashed black lines: 95% confidence intervals.

Quantitative results



Power study - The mean of power values over the 1,000 simulations are calculated at a 0.05 FDR level.

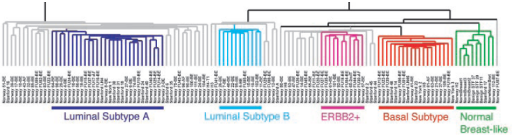
Quantitative results



Application

Breast cancer in a few words

- An heterogeneous disease (5 subtypes)



- Presence (ER+)/absence (ER-) of Estrogen Receptors: an essential parameter of tumor characterization.

Data

Affymetrix U133-Plus2.0 arrays:

- 537 patients (446 ER^{+} vs. 91 ER^{-})
- 54,675 probes

Topological data

PPI network from HPRD and String:

- 13,611 proteins
- $\sim 600,000$ interactions

Application

Results

- 27 221 initial modules
- 14 significant modules (FDR 1%)
- 159 genes

Interpretation

Module	Size	Molecular / cellular function
1	38	**Amino-acid metabolism**
2	1 (GATA3)	**Strong association with ER status** (Voduc et al. 2008)
3	35	**Breast cancer regulation by Stathmin1*** (*oncoprotein which takes part in the preventive progression of ER$^+$ tumors)
4	1 (AGR3)	**Involved in ER-responsive breast tumors** (Fletcher et al. 2002)
5	7	**PI3K/AKT signaling** (cell death and cellular growth) **Aryl Hydrocarbon Receptor signaling** (*AHR represses ER)

Discussion

Summary

- DiAMS: local-score approach for the selection of disease associated modules of genes
- Proved quantitative results on:
  - power gains,
  - reproducibility improvements,

  in comparison to the classical approach.
- Limitation: coverage and quality of PPI databases

Perspectives

- Investigate the predictive performance of DiAMS
- Assess the reproducibility on real datasets.

Acknowledgements

Statistic & Genome Laboratory

Christophe Ambroise

Michèle, Carène, Claudine, Catherine, Camille, Etienne, Pierre, Gilles, Cecile, Maurice, Marie-Luce, Anne-Sophie, Cyril, Justin, Van-Hanh, Yolande, Sarah, Marius, Bernard et Julien.

Jan, Caroline, Fabrice, Mickaël, Matthieu, Jonas, Sory.

Pharnext

Mickaël Guedj

Serguei Nabirotchkin

Ilya Chumakov