

# A factor model to analyze heterogeneity in gene expression in a context of QTL mapping

Yuna Blum, Sandrine Lagarrigue & David Causeur

*UMR598 Animal Genetics, Applied Mathematics Departement,  
Agrocampus Ouest, Rennes  
IRMAR UMR6625 CNRS*



January 2010

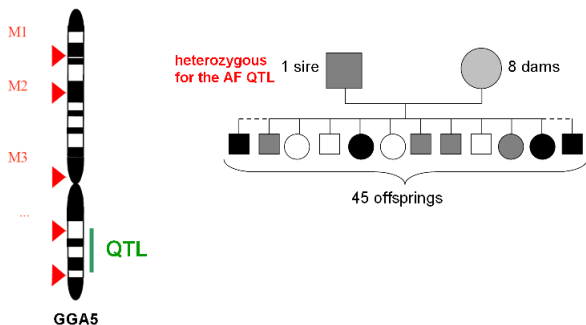
Workshop on **Statistical Methods for Post-Genomic Data**

# Outline

- 1 Background
- 2 The FAMT method
- 3 Results
  - Functional characterization
  - QTL characterization
  - Heterogeneity analysis
- 4 Concluding comments

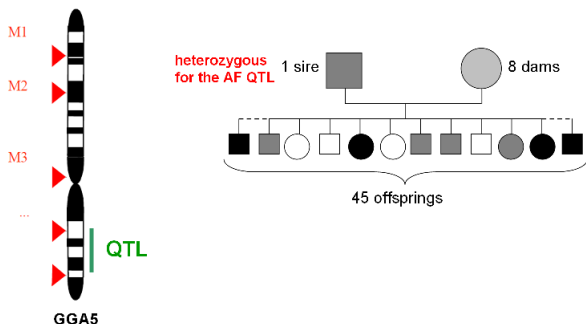
# QTL analysis using transcriptome profiles

**Context:** mapping QTL for **abdominal fatness** (AF) in chickens. One QTL has been previously detected around 175cM on the GGA5 chromosome (Le Mignon *et al*, 2009).



# QTL analysis using transcriptome profiles

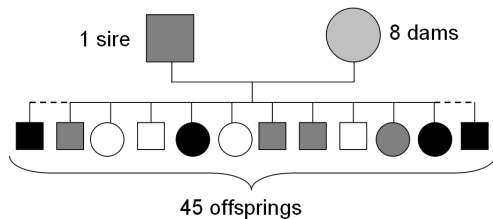
**Context:** mapping QTL for **abdominal fatness** (AF) in chickens. One QTL has been previously detected around 175cM on the GGA5 chromosome (Le Mignon *et al*, 2009).



**Aim:** a **better characterization** of the AF QTL on the GGA5 using transcriptomic data.

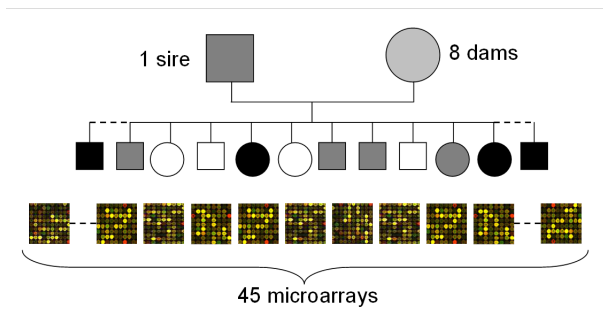
# Transcriptomic data

**Dataset:** hepatic transcriptome profiles for 11213 genes of the 45 half sib male chickens.



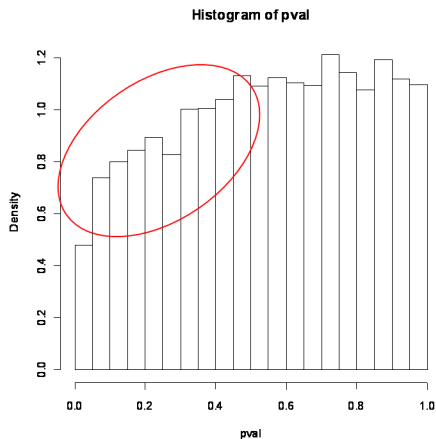
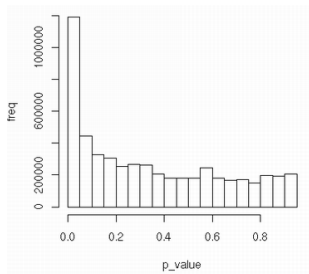
# Transcriptomic data

**Dataset:** hepatic transcriptome profiles for 11213 genes of the 45 half sib male chickens.



**First step:** identification of a list of genes correlated to the AF trait.

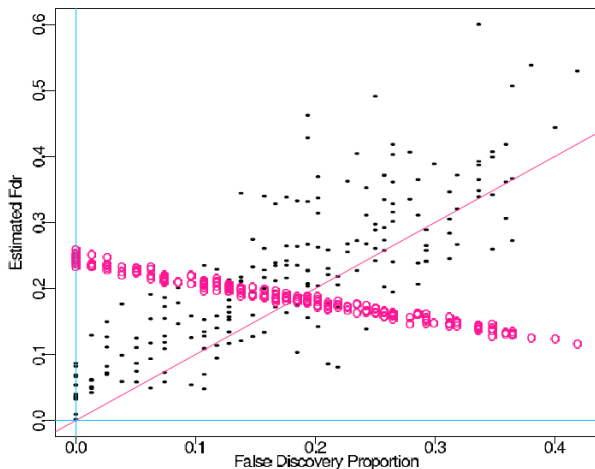
# Histogram of p-values



*Correlation and Large-Scale Simultaneous Significance Testing, B.Efron, 2007.*

# Impact of dependence in multiple testing

*Correlation and Large-Scale Simultaneous Significance Testing,*  
*B.Efron, 2007.*



# Outline

- 1 Background
- 2 The FAMT method
- 3 Results
  - Functional characterization
  - QTL characterization
  - Heterogeneity analysis
- 4 Concluding comments

# Factor Analysis for Multiple Testing

The common information shared by all the variables ( $m$ ) is modeled by a factor analysis structure.

The **common factors**  $Z$  : small number ( $q \ll m$ ) of latent variables (Friguet *et al.*, 2009, *JASA*)

**Unconditional model:**

$$Y^{(k)} = \beta_0^{(k)} + \mathbf{x}'\beta^{(k)} + \epsilon^{(k)}$$

$$\text{Var}(\epsilon) = \Sigma$$

**FAMT model:**

$$Y^{(k)} = \beta_0^{(k)} + \mathbf{x}'\beta^{(k)} + \mathbf{b}'_k Z + \epsilon^{*(k)}$$

$$\text{Var}(\epsilon^*) = \Psi \quad \Sigma = \Psi + \mathbf{B}\mathbf{B}'$$

# Factor-adjusted test statistics

The adjusted test statistics are conditionally centered and scaled version of usual test statistics

## Factor adjusted test statistics

$$T_z^{(k)} = T^{(k)}(Y^{(k)} - \mathbf{b}'_k Z)$$

## Noncentrality parameter

$$ncp(T_z^{(k)}) > ncp(T^{(k)})$$

# Outline

- 1 Background
- 2 The FAMT method
- 3 Results**
  - Functional characterization
  - QTL characterization
  - Heterogeneity analysis
- 4 Concluding comments

# Multiple testing

## Classical method :

**287** genes were significantly correlated considering a significant threshold of 0.05 without any correction for multiple tests.

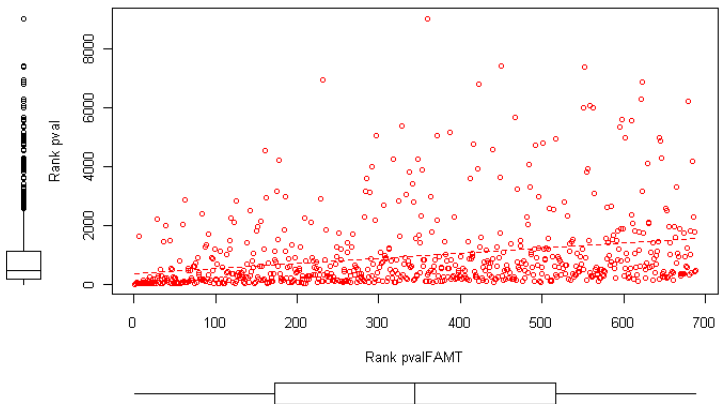
## FAMT :

**6** factors containing a common information shared by all genes and independent from the variable of interest.

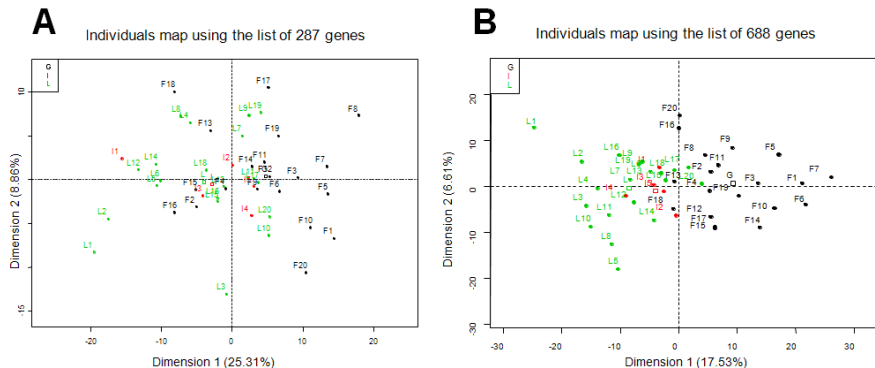
**688** genes which expressions were significantly correlated to the AF trait.

This suggests that correlation between many gene expressions and the variable of interest is underestimated due to gene dependence.

# Multiple testing



# Principal component analysis



The PCA generated with the 688 genes discriminates much more the **lean** and the **fat** chickens.

# Enrichment tests

## LIST OF 287 GENES

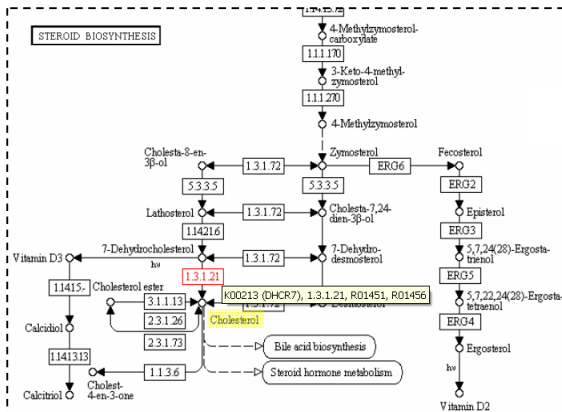
GOID	GO Term	Size	Count	Pvalue
GO.0006470	<i>protein amino acid dephosphorylation</i>	56	5	0.015
GO.0006725	<i>cellular aromatic compound metabolic process</i>	38	4	0.017
GO.0007259	<i>JAK STAT cascade</i>	9	2	0.022
GO.0043543	protein amino acid acylation	9	2	0.022
GO.0044259	multicellular macromolecule metabolic process	10	2	0.027
GO.0008033	tRNA processing	26	3	0.0296
GO.0033002	muscle cell proliferation	11	2	0.032
GO.0050730	regulation of peptidyl tyrosine phosphorylation	12	2	0.038
Kegg ID	Kegg pathway	Size	Count	Pvalue
map04320	Dorso ventral axis formation	9	3	2.38E-03

## LIST OF 688 GENES

GOID	GO Term	Size	Count	Pvalue
GO.0016311	<i>protein amino acid dephosphorylation</i>	60	11	8.52E-04
GO.0046483	heterocycle metabolic process	33	7	3.21E-03
GO.0051186	cofactor metabolic process	64	10	4.97E-03
GO.0007259	<i>JAK STAT cascade</i>	9	3	0.014
GO.0006534	cysteine metabolic process	4	2	0.021
GO.0006725	<i>cellular aromatic compound metabolic process</i>	38	6	0.026
GO.0007185	transmembrane receptor tyrosine phosphatase signaling	5	2	0.033
GO.0000097	sulfur amino acid biosynthetic process	5	2	0.033
GO.0006700	<b>C21 steroid hormone biosynthetic process</b>	5	2	0.033
GO.0006787	porphyrin catabolic process	5	2	0.033
GO.0001764	neuron migration	12	3	0.033
GO.0008211	glucocorticoid metabolic process	6	2	0.048
Kegg ID	Kegg pathway	Size	Count	Pvalue
map00630	Glyoxylate and dicarboxylate metabolism	9	4	1.87E-03
map00140	<b>C21 Steroid hormone metabolism</b>	6	3	5.11E-03
map04320	Dorso ventral axis formation	9	3	0.018

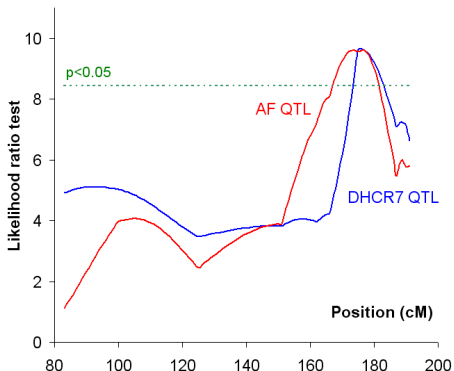
# QTL characterization

*Steroid metabolism*: STAR, **DHCR7** (not in the list of 287 genes), HSD11B1, CYP17A1 are in the list of 688 genes (FAMT).



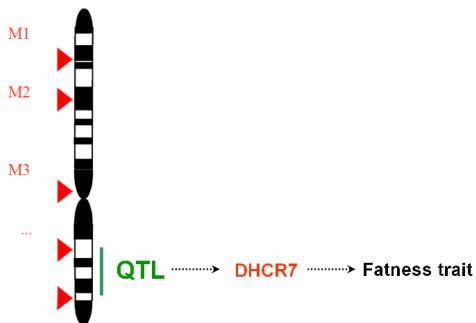
# QTL characterization

**Results:** DHCR7 finding through FAMT is controlled by the QTL located around 175 cM. **The causal mutation might be involved in the cholesterol metabolism.**



# QTL characterization

**Results:** DHCR7 finding through FAMT is controlled by the QTL located around 175 cM. **The causal mutation might be involved in the cholesterol metabolism.**



## Dissection of the complex trait

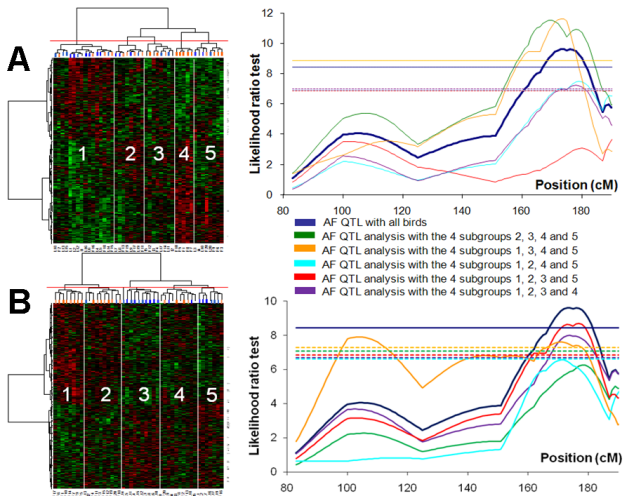
The variation of AF trait is due to variation of multiple biological pathways reflecting **numerous mutations**.

**Strategy**: dissection of the complex trait by grouping the offsprings according to their partial transcriptome profile based on a specific geneset correlated to the trait of interest.

This strategy allows in some cases to highlight **new QTL** which are unobserved at the family level (Schadt *et al*, 2003, Le Mignon *et al*, 2009).

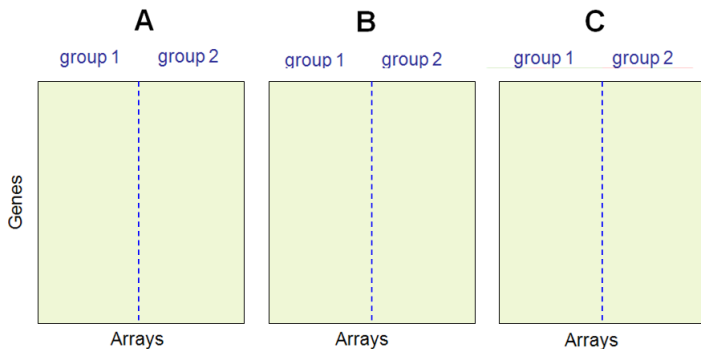
# Dissection of the complex trait

Two-way hierarchical cluster analysis: **(A)** using the list of the **287** genes (classical approach), **(B)** using the list of **688** genes (FAMT).



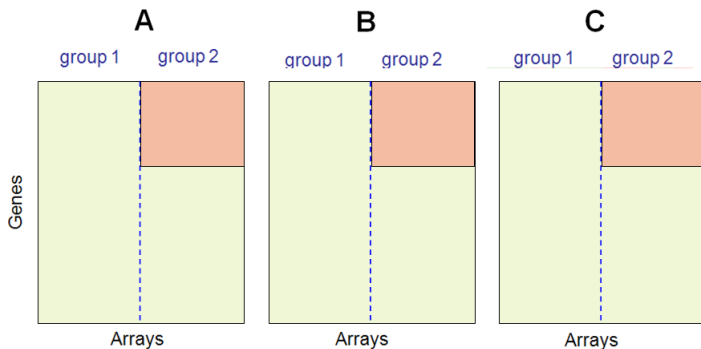
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



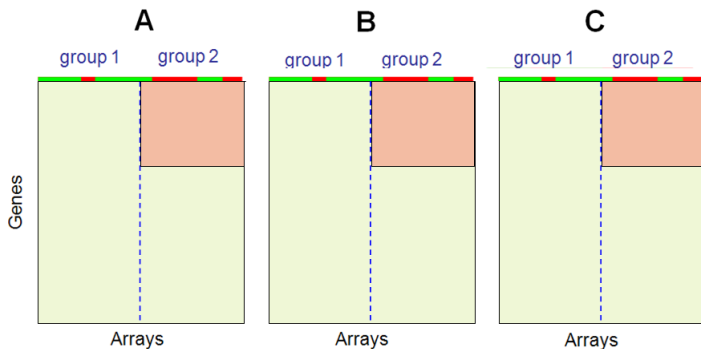
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



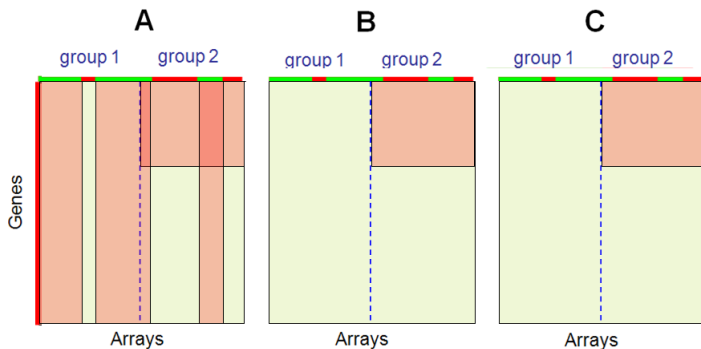
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



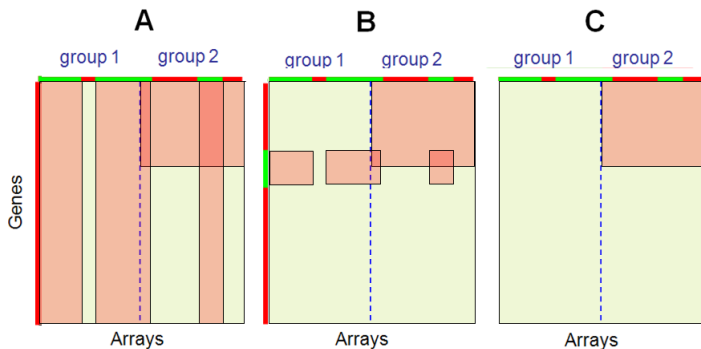
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



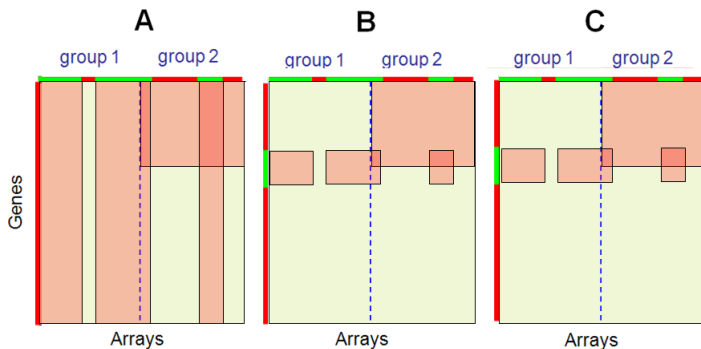
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



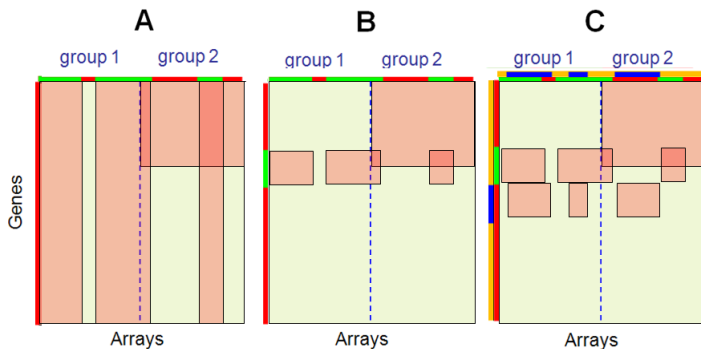
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



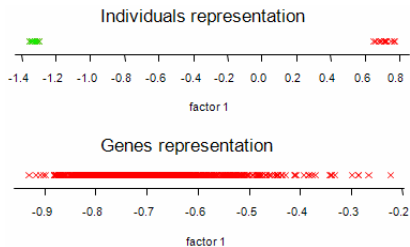
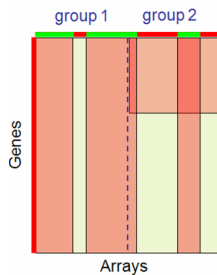
# Illustrative examples

Simulation of independent expressions for 1000 genes on 20 arrays. 3 simple situations of heterogeneity:



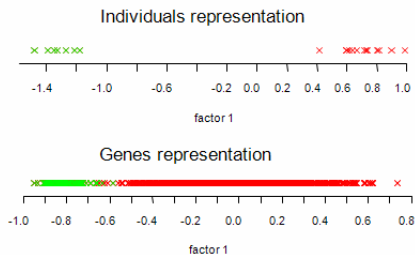
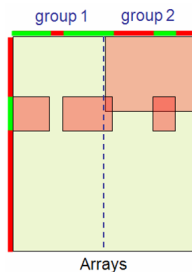
# Illustrative examples

## Case A: One independent variable affecting **all** genes



# Illustrative examples

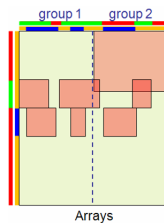
## Case B: One independent variable affecting a **set** of genes



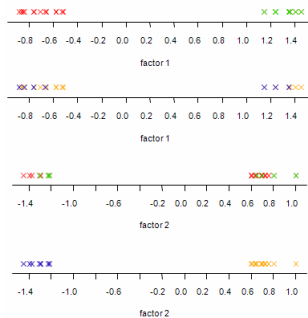
# Illustrative examples

## Case C:

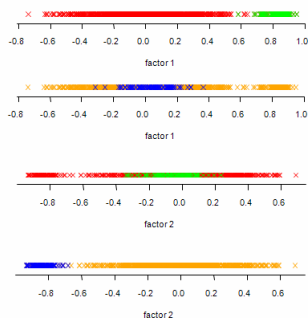
Two independent variables affecting **two different sets** of genes



Individuals representation



Genes representation



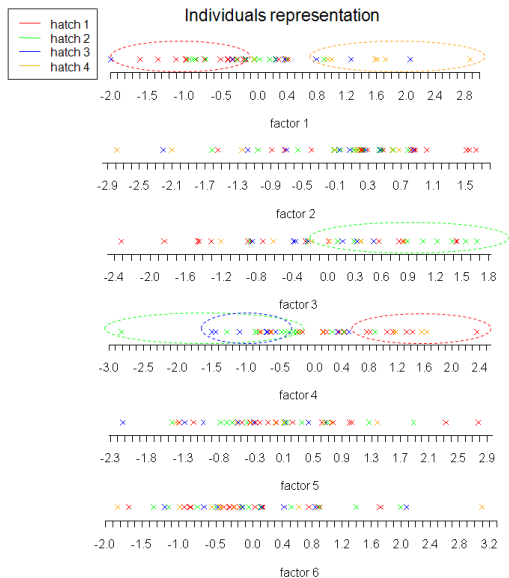
# Expression data set in chickens

Using :

**external information** on the experimental design such as the hatch, the body weight and the dam.

**gene information** such as functional categories, oligonucleotide size and location on the microarray (block, row, column).

# Interpretation of the factors






# Interpretation of the factors

	Individual information			Gene information			
	hatch	dam	weight	oligo size	chip block	chip row	chip column
Factor 1	<b>8.92E-05</b>	0.139	0.129	<b>2.20E-16</b>	<b>2.20E-16</b>	0.074	0.179
Factor 2	0.074	0.913	<b>4.70E-03</b>	<b>2.20E-16</b>	<b>2.20E-16</b>	0.041	0.857
Factor 3	<b>1.90E-02</b>	0.848	0.489	<b>2.55E-14</b>	<b>2.20E-16</b>	0.716	0.376
Factor 4	<b>6.00E-03</b>	0.127	0.959	<b>1.41E-07</b>	<b>2.20E-16</b>	0.707	0.167
Factor 5	0.435	0.217	0.884	0.529	<b>2.20E-16</b>	<b>4.97E-03</b>	<b>9.99E-05</b>
Factor 6	0.946	0.412	0.615	<b>1.79E-07</b>	<b>2.20E-16</b>	0.876	<b>5.11E-07</b>

# Outline

- 1 Background
- 2 The FAMT method
- 3 Results
  - Functional characterization
  - QTL characterization
  - Heterogeneity analysis
- 4 Concluding comments

## Concluding comments

- **FAMT procedure**: large improvements in multiple testing procedures comparing to the classical approach. List of genes more related to the trait of interest.
- **QTL context**: the list of genes found by FAMT allows a functional characterization of a known QTL and the detection of another QTL.
- **Heterogeneity analysis**: extraction of information from what was before simply considered as statistical noise.
- FAMT  package available at <http://www.agrocampus-ouest.fr/math/FAMT>

# Bibliography

- **Blum Y *et al.***: *A factor model to analyze heterogeneity in gene expression*. BMC Bioinformatics, submitted.
- **Friguet C *et al.***: *A Factor Model Approach to Multiple Testing Under Dependence*. Journal of the American Statistical Association 104:488, 1406-1415, 2009.
- **Leek J, Storey J**: *Capturing heterogeneity in gene expression studies by surrogate variable analysis*. PLoS Genetics 2007, 3(9).
- **Le Mignon G *et al.***: *Using transcriptome profiling to refine QTL regions on chicken chromosome 5*. BMC Genomics, 10-575, 2009.
- **Schadt E.E. *et al.***: *Genetics of gene expression surveyed in maize, mouse and man*, Nature, 297-302, 2003.