XIV Conferencia Española de Biometría

Scatterplot clustering for the integrative analysis of expression and methylation data

M. Carme Ruiz de Villa, Francesc Carmona
Diego Arango del Corro, Josep Lluís Mosquera
Alex Sánchez

May 23, 2013

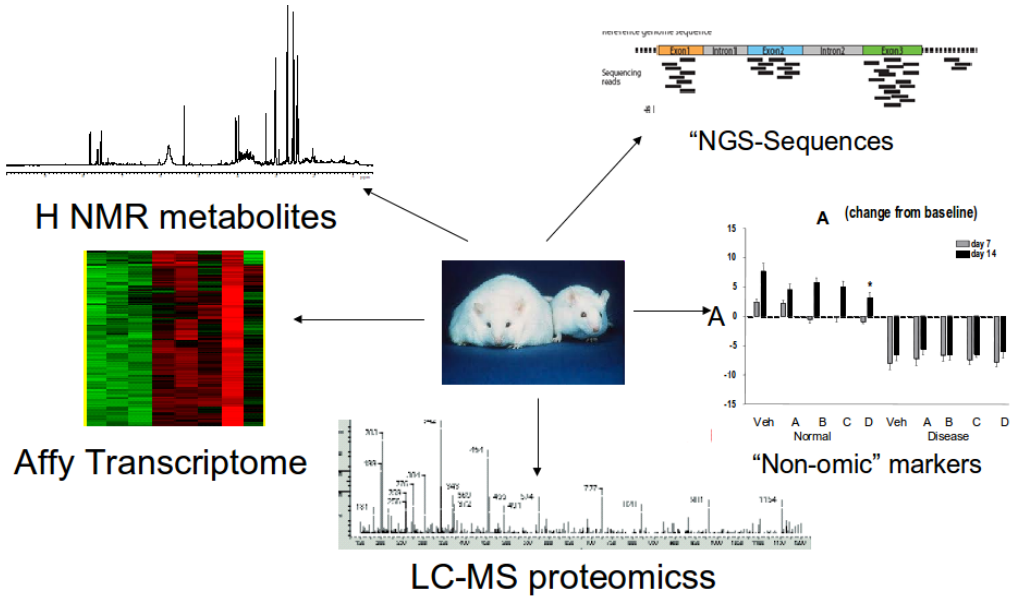
Departamento de Estadística
Facultad de Biología

Universitat de Barcelona

Table of Contents

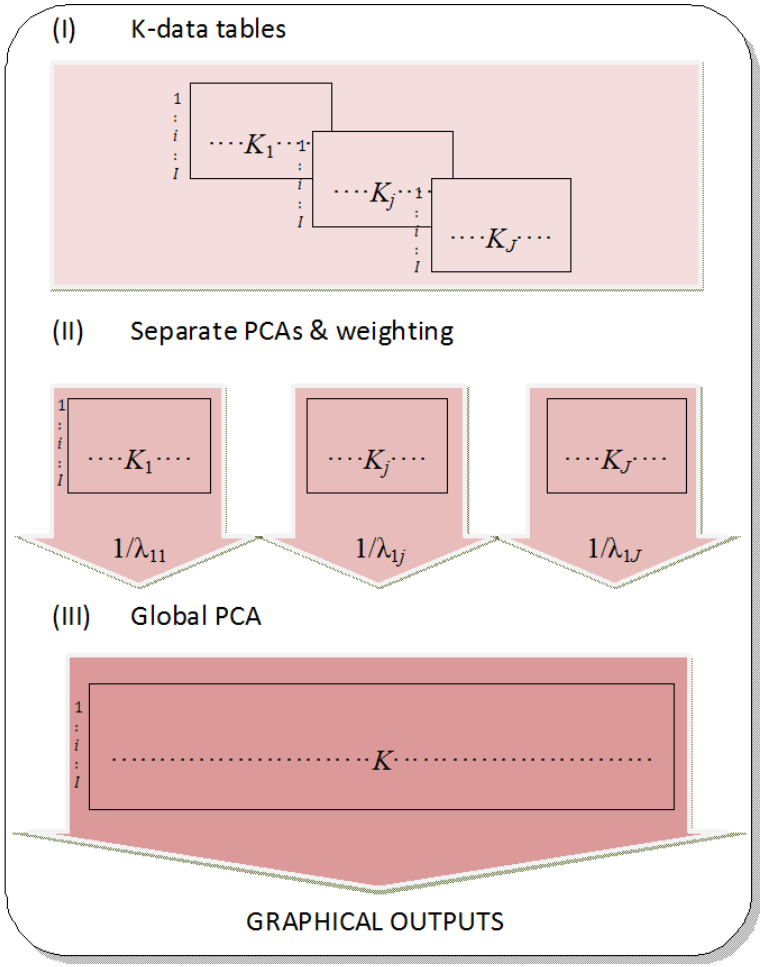
Post-genomics age: the next generation

- 'Omics' technologies are becoming increasingly important:
  - The advent of *next generation sequencing*, provides information on many type of *genomic*, *transcriptomic* or *epigenomic* data.
  - The generalization of high-throughput technologies allow to study biological processes at the different levels at which they happen.



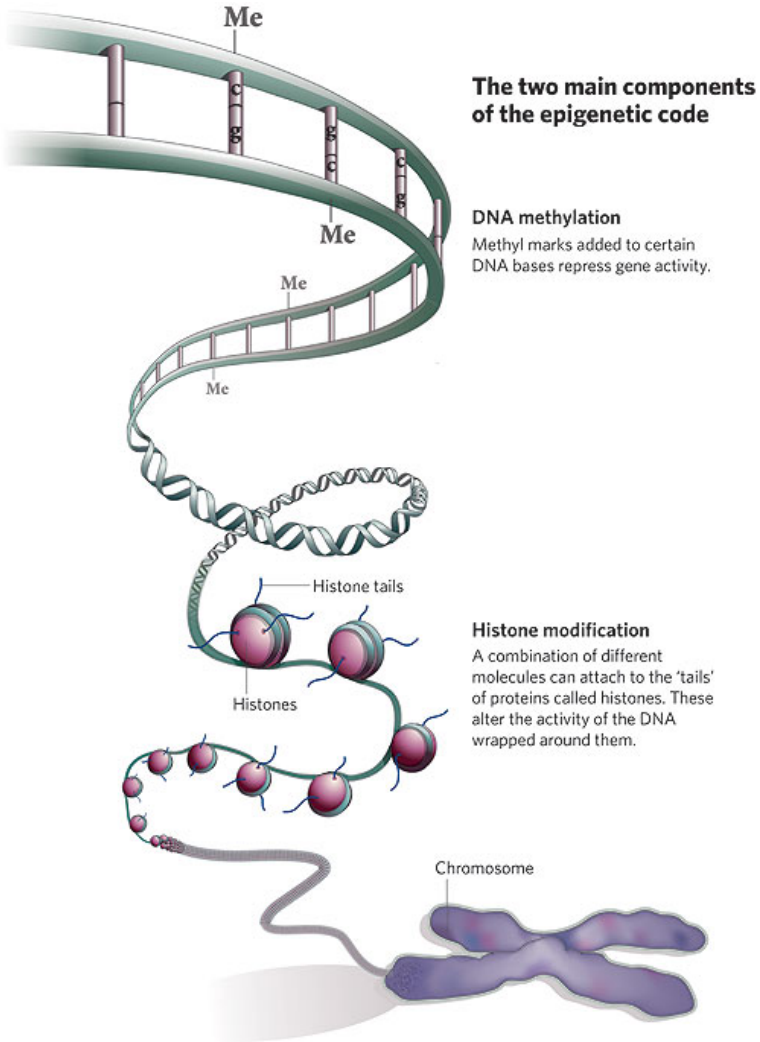
Data integration and systems biology -again

Statistics and bioinformatics are faced with the need for developing methods and tools for the integrative analysis of (big) data sets of different types and origins.



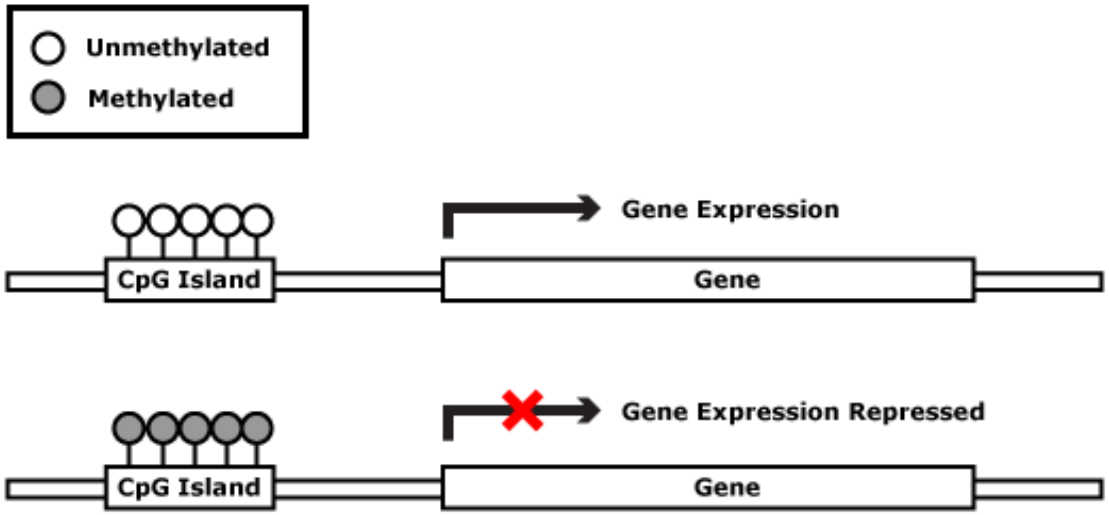
Epigenetics and epigenomics

- Epigenetics, *the study of environmental factors on gene expression in DNA*, is one of the disciplines that has experienced a renewed impetus:
- There is increasing evidence that many differentiation processes are triggered and maintained through epigenetic mechanisms such as *methylation* or *histone modifications*.



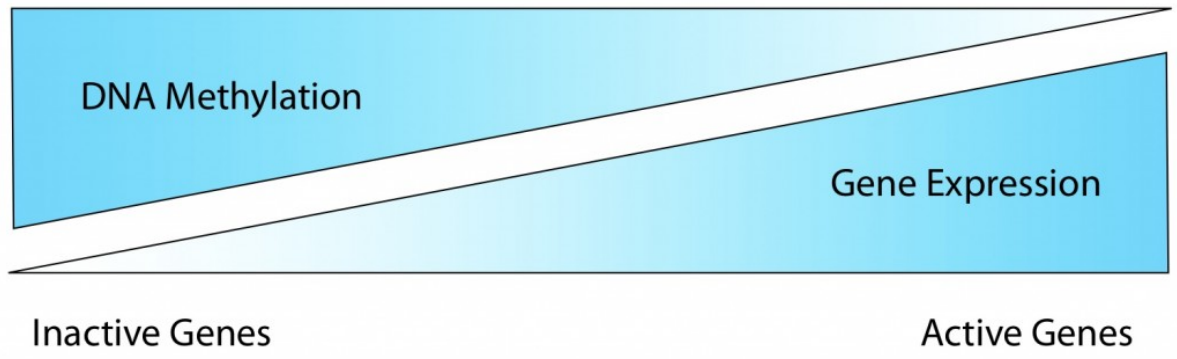
Methylation

- One main epigenetic regulatory mechanisms is methylation a process by which a gene's behavior is altered, but the gene itself isn't changed.
- Essentially methylation acts by inhibiting gene expression that is, the more methylated is a gene the more repressed is its expression



Methylation and gene expression

- Although the relation between methylation and gene expression is probably continuous ("*the more...the less...*"),

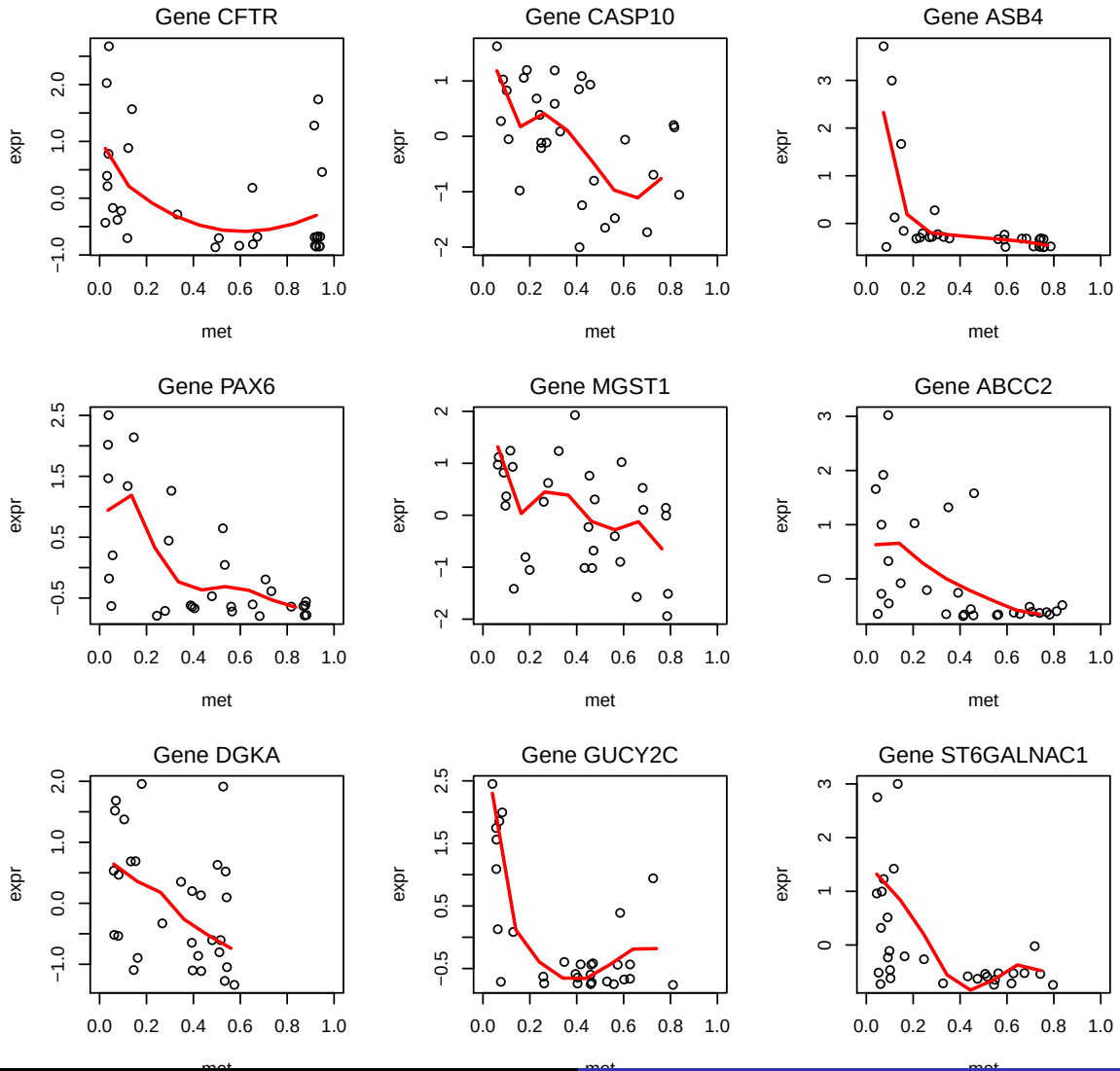


- methylation is, in practice, seen as a dual phenomenon
  - A methylated gene is "off"
  - An unmethlated gene is "on"
- Practical problem: **at which methylation level a gene is seen as "methylated" (that is, it is "turned off")?**

A colon cancer study

- This study originates in a work searching for colon cancer biomarkers.
- 30 cell lines characterized by increasing sensitivity to a drug were analyzed using several high-throughput methods: *transcriptomics*, *methylation*, *miRNAs*, *SNPs*, and *proteomics*.
- In this work we consider the problem of stablishing which genes were regulated by methylation.
- For each gene/methylation locus one has 30 points and a scatterplot showing the relation so we need methods to find patterns of scatterplots

Scatterplot patterns



Previous work

- Since measurements for methylation and expression are both continuous, scatterplots of these signals exhibit an L-shape pattern.
- Assuming that methylation is truly binary, there are two implications:
  1. the reflection point of the L-shape is an appropriate choice to binarize methylation data, and
  2. conditioning on the binarized on-off methylation status, the continuous valued methylation data and expression data should be independent,
- This motivates Liu(2012) to quantify the L-shape pattern using conditional mutual information (MI).

Objectives

- Study how gene expression is regulated by methylation in a set of colon cancer cell lines.
- Set up a method to detect the level of methylation at which a gene can be considered regulated by methylation (to be "on").
- Compare this method with other existing that have been developed to
  - detect methylation thresholds
  - detect patterns in scatterplots

Conditional Mutual Information

When studying methylation we are faced with two main questions:

- Which genes exhibit an L-shape, and
- what is the optimal threshold for binarizing methylation data for each L-shape gene.

The key

To determine whether methylation and expression of a gene exhibit an L-shape, compute the conditional Mutual Information (MI) for different choices of threshold to binarize the methylation data.

- If we consider the continuous valued methylation and expression data as two random variables X and Y, and denote a nominal threshold as t, the conditional MI can be written as a weighted sum of MIs on the two sides of the threshold.

$$cMI(t) = I(X, Y|X > t)P(X > t) + I(X, Y|X \leq t)P(X \leq t)$$

- When t is 0 or 1, cMI equals to the mutual information derived from all data points.
- For an L-shape gene, as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, and its value approaches zero when t coincides with the reflection point.

Optimal threshold

Therefore,

Optimal threshold

The ratio $r = \frac{\min\{cMI(t)\}}{cMI(0)}$ for an L-shape gene is small, and $t^* = \operatorname{argmin}\{cMI(t)\}$ is the optimal threshold for dichotomizing the methylation data of this gene.

Joint distribution estimator

- To estimate the MI terms we use a kernel-based estimator, which
  - constructs a joint probability distribution by applying a Gaussian kernel to each data point,
  - and estimates the MI based on the joint distribution.
- The estimator is as follows:

$$I(X,Y) = \frac{1}{M}\sum_{i=1}^{M} \log \frac{M \sum_{j=1}^{M} e^{-\frac{1}{2h^2}((x_i-x_j)^2+(y_i-y_j)^2)}}{\sum_{j=1}^{M} e^{-\frac{1}{2h^2}(x_i-x_j)^2} \sum_{j=1}^{M} e^{-\frac{1}{2h^2}(y_i-y_j)^2}}$$

where h is a tuning parameter for the kernel width and empirically set $h = 0.3$.

Clustering using Spline regression

We implemented regression based on B-splines because they are particularly efficient due to the block-diagonal basis matrices that result.
Let

- $\varsigma = \{t_1 < \cdots < t_N\}$ non decreasing knot sequence
- $[t_m, t_{m+1})$ half open interval
- B_{mp} p-th order polynomial (degree $p-1$) with finite support over the interval and 0 everywhere else so that $\sum_{m=1}^{N-p} B_{mp}(x) = 1$
- then $s(x) = \sum_{m=1}^{N-p} B_{mp}(x) c_m$

Clustering using Spline regression (2)

To represent the curve we set:

$$y_{ij} = s(x_{ij})$$

So

$$\mathbf{y}_i = \mathbf{B}_i \mathbf{c}$$

with

- $\mathbf{B}_i = [B_{1p}\mathbf{x}_i, B_{2p}\mathbf{x}_i, \ldots, B_{Lp}\mathbf{x}_i]$ the spline basis matrix
- $\mathbf{c}$ the vector of spline coefficients.

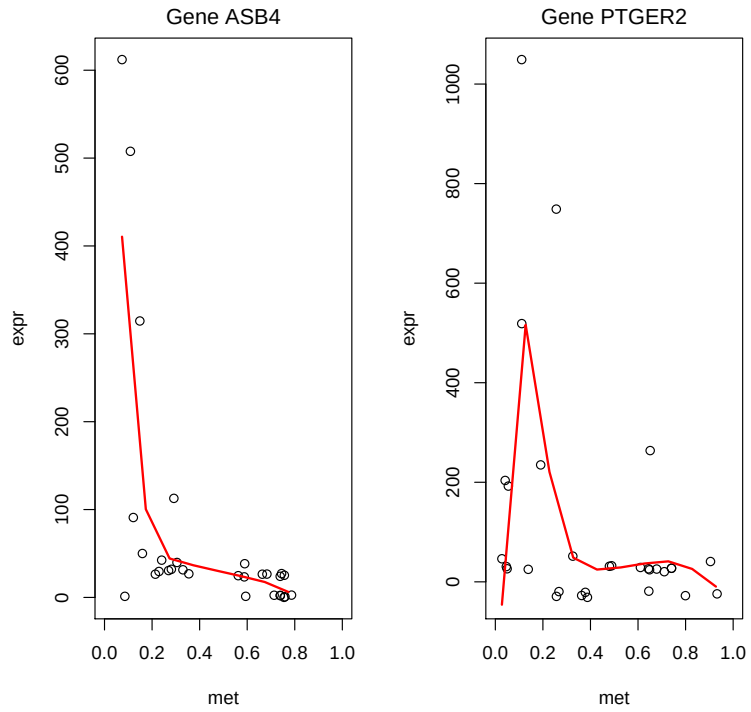
Clustering using Spline regression (3)

Algorithm

1. Selection of the genes with a negative significant correlation
2. Fit cubic regression splines
3. Data to cluster: splines coefficients
4. Calculation of a distance matrix between genes as $1 - \rho$
5. Hierarchical clustering

Results (1) Splines–based regression

- After the previous selection of genes we worked with 191 genes
- We decided to choose 5 clusters
- The 2 first clusters included the genes with an L-shape



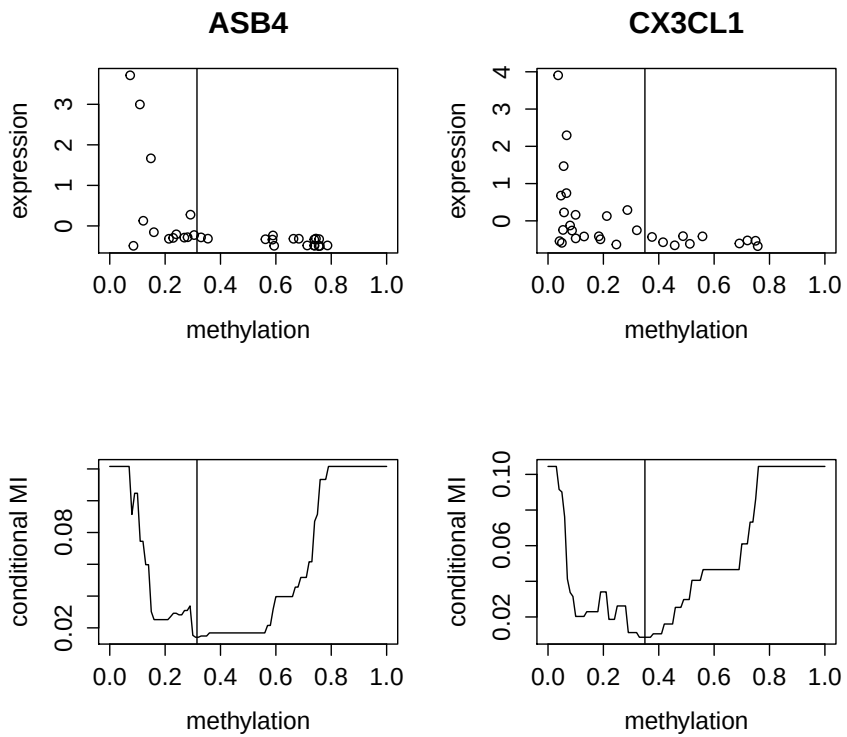
Results (2) Conditional Mutual Information

- No previous selection of the genes was needed
- We filtered for L-shapes using a combination of three criteria:
  - the ratio $r < 0.25$
  - unconditioned MI $cMI(0) > 0.1$
  - the median expression on the left side of the optimal threshold t^* is higher than the median expression on the right side.
- The parameters are chosen according to a random permutation test (see Liu(2012)).
- According to the above criteria, a total of 641 genes are selected to be L-shape genes.

Results (3) Comparison between the methods

The results of both methods that can be summarized in the following table:

Initial selection	191	641
Cluster	Splines	cMI
1	140	102
2	22	16
Total	162	118



Conclusions

- We have found similar results between both methods.
- Biological interpretation is still being done by biological researchers although results are consistent with the hypothesis (we have found genes regulated by methylation).
- Sample size is a limiting factor: cMI works better with hundreds of samples but one may have a very small number (real cases: 30, 12)

Acknowledgments

- Members of the research group *Estadística i Bioinformàtica* at the Statistics department at the University of Barcelona.
- Members of the *Statistics and Bioinformatics Unit (UEB)* and the *High technology Unit (UAT)* at the *Vall d'Henron Research Institute (VHIR)* in Barcelona.