A heuristic algorithm to select genes potentially regulated by methylation

Alex Sánchez, Berta Miró, Francesc Carmona,
Sarah Bazzoco and Diego Arango del Corro

July 09, 2018

Genetics, Microbiology and Statistics Department
Facultad de Biología, Universitat de Barcelona
Statistics and Bioinformatics Unit (UEB)
Department Molecular Oncology-CIBBIM
Vall Hebron Institut de Recerca

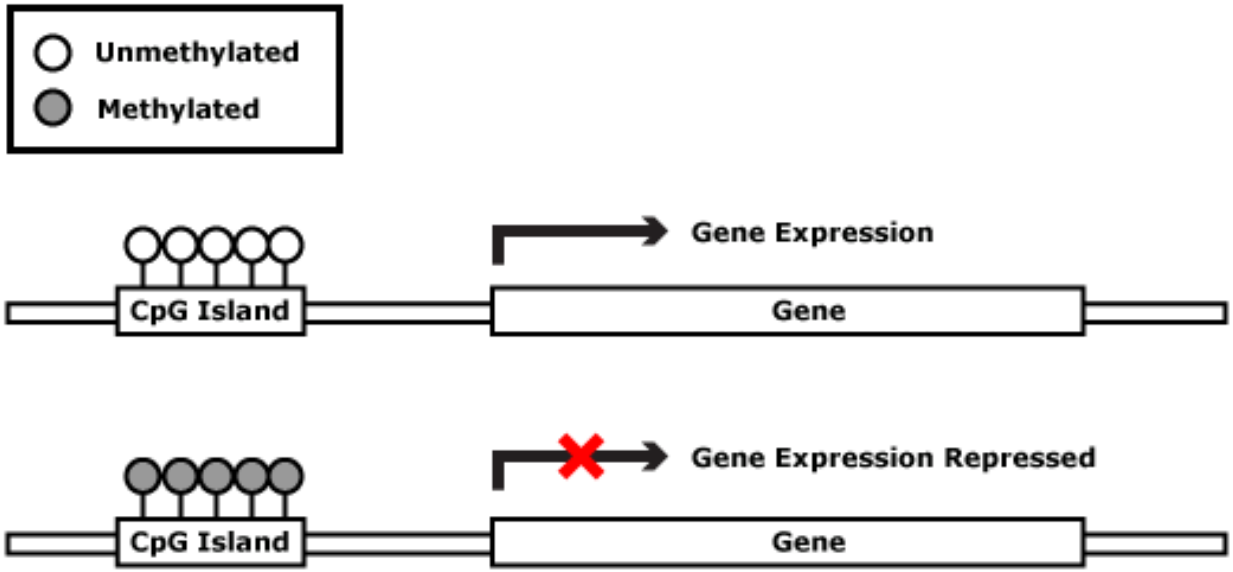
Table of Contents

Genome-wide analysis of colorectal cancer

- CRC is a serious public health problem (2.M diagnosed/year) but the number of therapies available is smaller than in other cancer types.
- Researcher's interest: identification of biomarkers for chemotherapy sensitivity in colorectal cancer (CRC).
- The researchers' approach was to look for *genes regulated by methylation* which could be considered possible therapeutic targets.

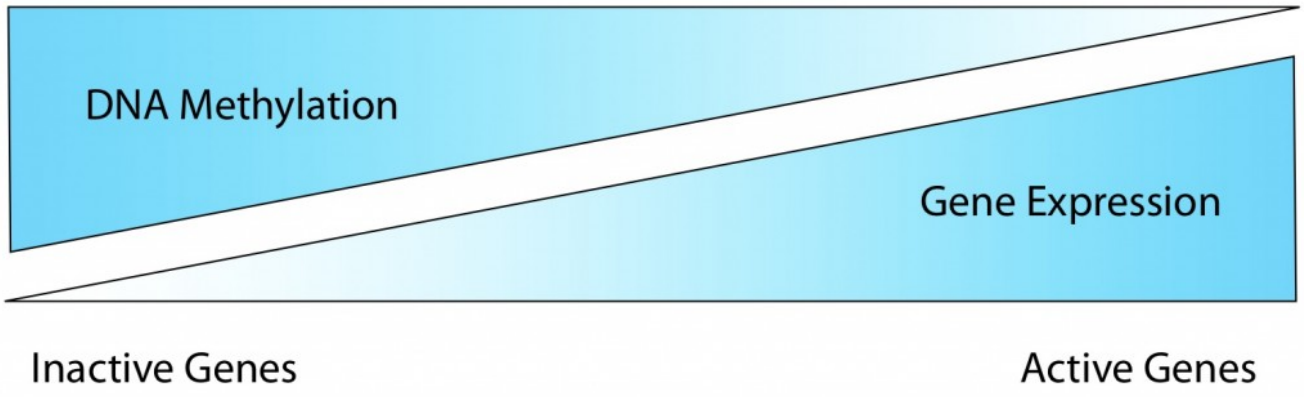
Methylation

- Methylation of CpG dinucleotides in the promoter of genes involved in the oncogenic process has been shown to be a key process contributing to tumor initiation and/or progression.
- Essentially (and especially in cancer) methylation acts by inhibiting gene expression that is, *the more methylated is a gene the more repressed is its expression*



Methylation and gene expression

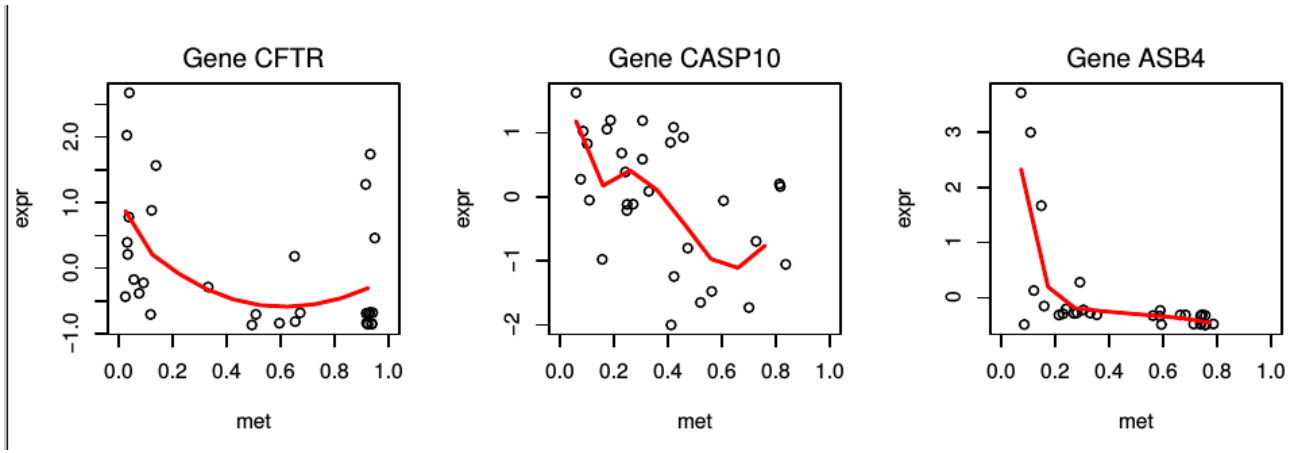
- Although the relation between methylation and gene expression is probably continuous ("*the more...the less...*"),



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

Patterns of (negative) association

- Considering the relation between methylation and expression in cancer (the higher methylation the lower the expression...)
- leads to expecting that scatterplots depicting the relation between methylation and expression show a negative correlation.
- This is usually the case and, indeed, *genes known to be regulated by methylation often show* **an L-shape pattern** *in these plots*.



Selecting genes by mining scatterplots

- Assuming the relation described above is true...
- Finding genes regulated by methylation is equivalent to finding genes whose methylation–expression scatterplot has an L–shape.
- There is a scatterplot *per* gene and thousands of genes: *Automatic methods for selecting interesting genes through their scatterplots are required*.
- There exist methods that add on the correlation coefficient but they are not very successful.

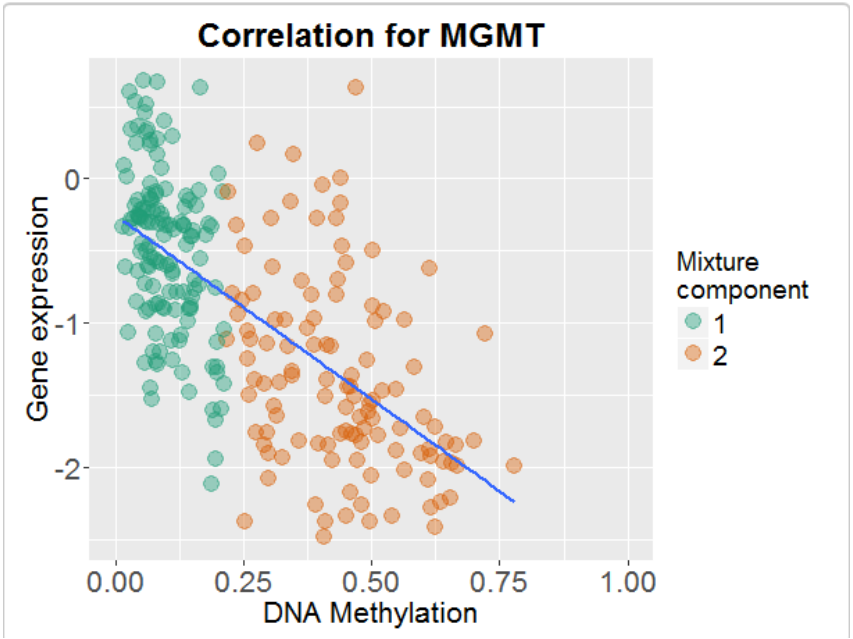
Objectives

The main objectives of this work are:

1. To introduce a new method to select genes showing an L-shape
2. To compare it with previously available methods,
3. To apply the selected methods on a specific CRC dataset and validate the findings based on their biological relevance.

Naïve method

- Simplest approach: Call a gene potentially regulated by methylation if a negative correlation between expression and methylation is observed.
- Most prevalent approach. (eg: Bioc. `MethylMix` package).
- Drawbacks:
  - Lack of power.
  - Easy to miss "sharp" L shapes.

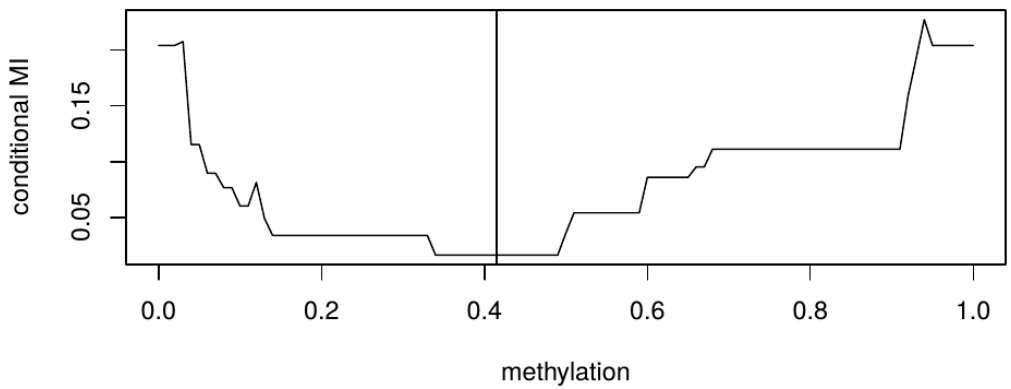


Conditional Mutual Information

- Following [2] in order to determine whether methylation X and expression Y of a gene exhibit an L–shape, the conditional Mutual Information $cMI(t)$ for different choices of threshold t is computed.

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

- If the relation between methylation and expression shows an L-shape as t moves from 0 to 1, $cMI(t)$ first decreases and then increases, its value approaching zero when t coincides with the reflection point.



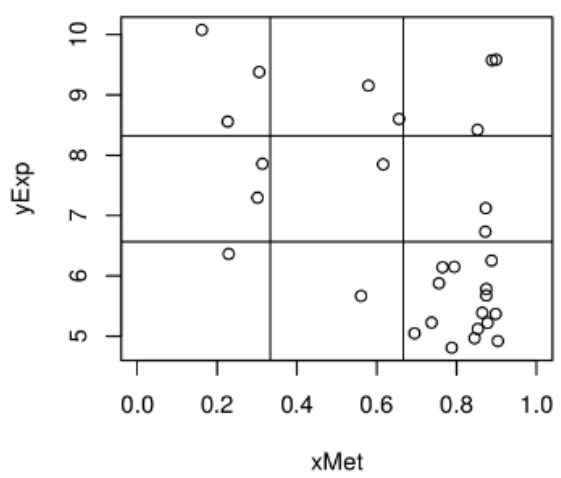
- Genes whose cMI go below a threshold can be considered L-shaped.

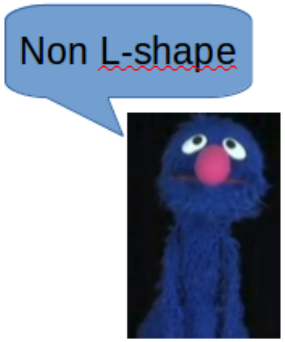
Some previously applied methods

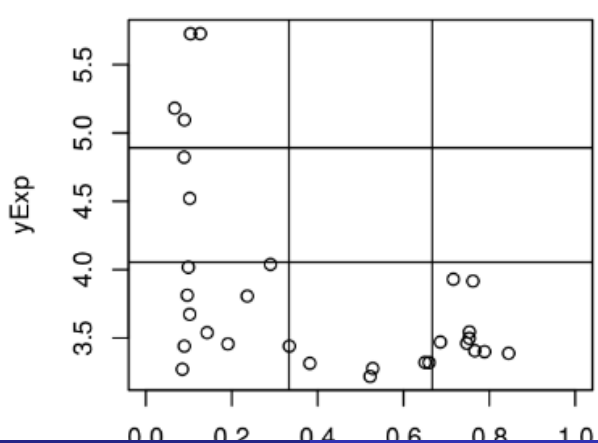
- Naive method: Call regulated by methylation genes depictinhg negative correlation between expression and methylation
- Study variation of conditional mutual information along different methylation values.[2].
- Use regression splines to fit a curve to the scatterplot and use clustering to group patterns. [3].
- Analyze scatterplots characteristics with Tukey's Scagnostics method [4]

What is an L-shape, whatsoever

- After trying different approaches to detect L-shapes, one comes back to a naive approach like
- *"L-shaped" genes should show an L shape in the scatterplot*;
  - The more L-shaped a scatterplot the more its values scatter near the vertical and horizontal axes,
  - The more these values move away move from these positions the least L-shaped the gene is.







A weighting system

1. Overimpose a 3×3 grid on the scatterplot.
2. Classify the scatterplot as **"L" or "non-L"** based on a small set of conditions:
    1. There must be a *minimum* number of points in the upper-left (cell (1,1)) and lower right (cell (3,3)) corners of the grid.
    2. There must be a *maximum* number of points in the upper right (cell (1,3)) because points there mean hypermethylation and hyperexpression which is the opposite of what we are looking for.
    3. We will usually *not require to have a minimum of points in cell (3,1)* unless we are really willing to have an L-shape (in our setting we will also be happy tho recover diagonals, which also reflect a negative correlation!).

$$\mathbb{1}_L(X) = \bigwedge_{i,j} X \circ C \circ \left(mMP \times \sum_{i,j} x_{ij} \right),$$

A scoring system

1. Score points on each subgrid in such a way that
   1. Points in permitted regions (left-outer margin, i.e. cells: (1,1), (2,2), (3,1), (3,2), (3,3)) score positively if the scatterplot has been classified as L or zero if it has been classified as non-L.
   2. Points in non-desired regions (outer band. i.e. cells (1,2), (1,3), (2,3)) score negatively in all cases.
   3. Some regions may be declared neutral and not-score, such as cell (2,2).

$$S(X) = W_L \circ X \times \mathbb{1}_L(X) + W_{L^C} \circ X \times \mathbb{1}_{L^C}(X),$$

2. Use cross-validation to tune scoring parameters (*if a set of positive and negative L-shaped genes is available*).

An example

1. Min-Max Counts

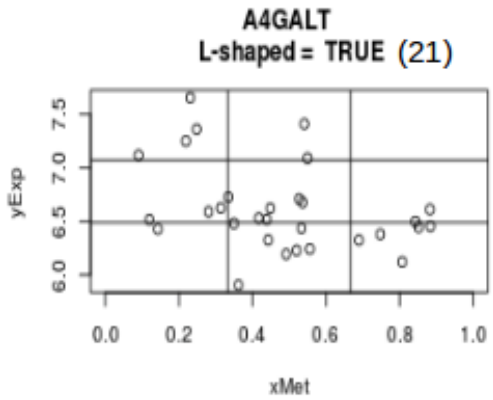
$$mMP = \begin{pmatrix} 10 & 20 & 0 \\ 5 & 0 & 20 \\ 0 & 5 & 5 \end{pmatrix}$$

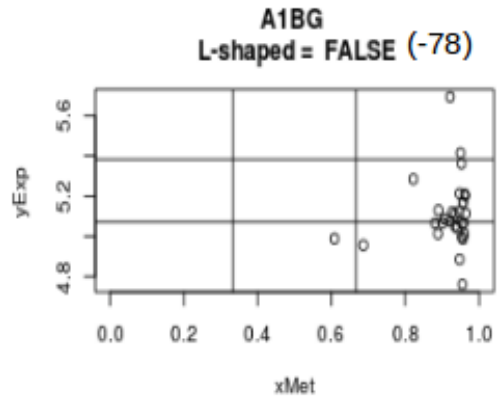
2. Matrix of weights for TRUE L scatterplots

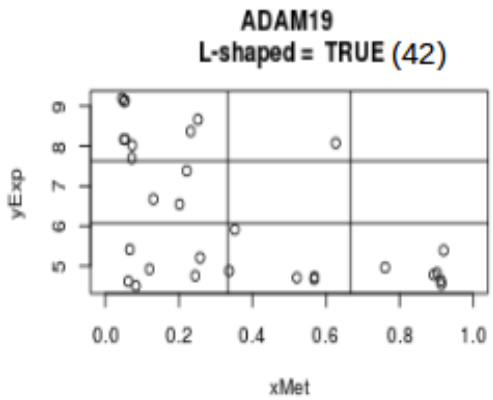
$$W_{TRUE-L} = \begin{pmatrix} 2 & -2 & -25 \\ 1 & 0 & -2 \\ 1 & 1 & 2 \end{pmatrix}$$

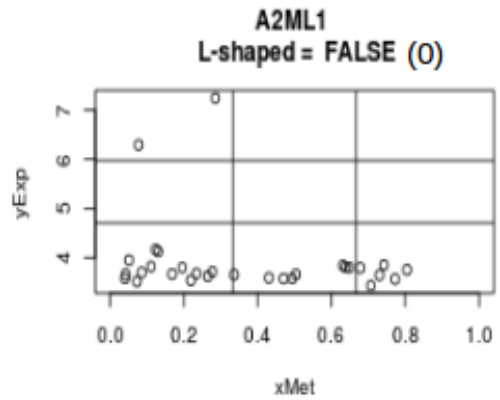
3. Matrix of weights for FALSE L scatterplots

$$W_{FALSE-L} = \begin{pmatrix} 0 & -2 & -25 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$







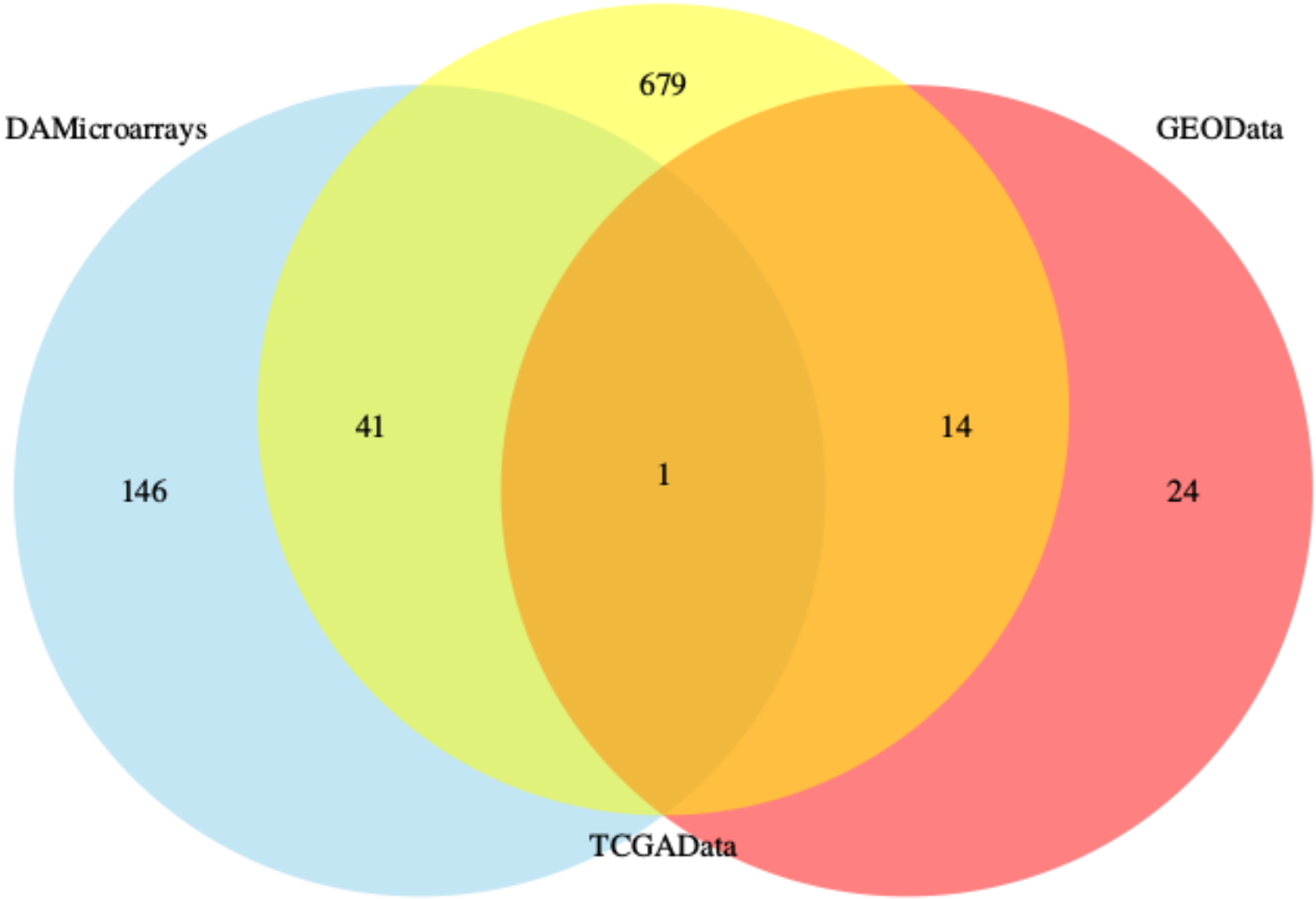


Data for the comparisons

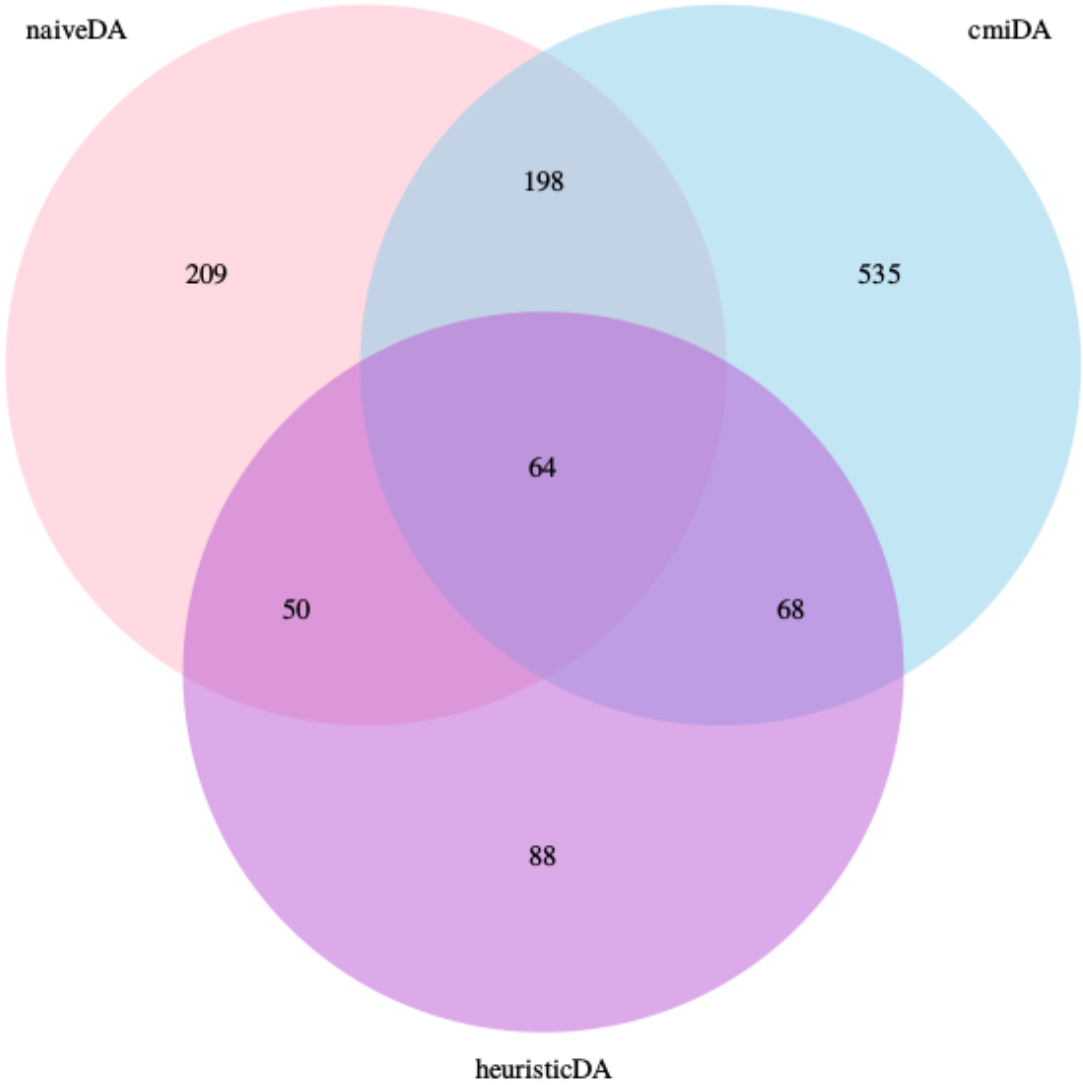
- The methods have been tested using three real and one simulated dataset.
- Distinct datasets were generated by similar but not identical technologies.
- Genes non common to the three datasets were removed from the analysis

Name	*Source*	*Genes*	*Samples*	*Arrays*	*Methylation*
TCGA	Nature 2012	1788	223		
GEO	GSE25070	1191	25	Bead	25K
DA	Reseracher's	11359	30	Affy	25k

Results: Comparison between datasets



Results: Comparison between the methods

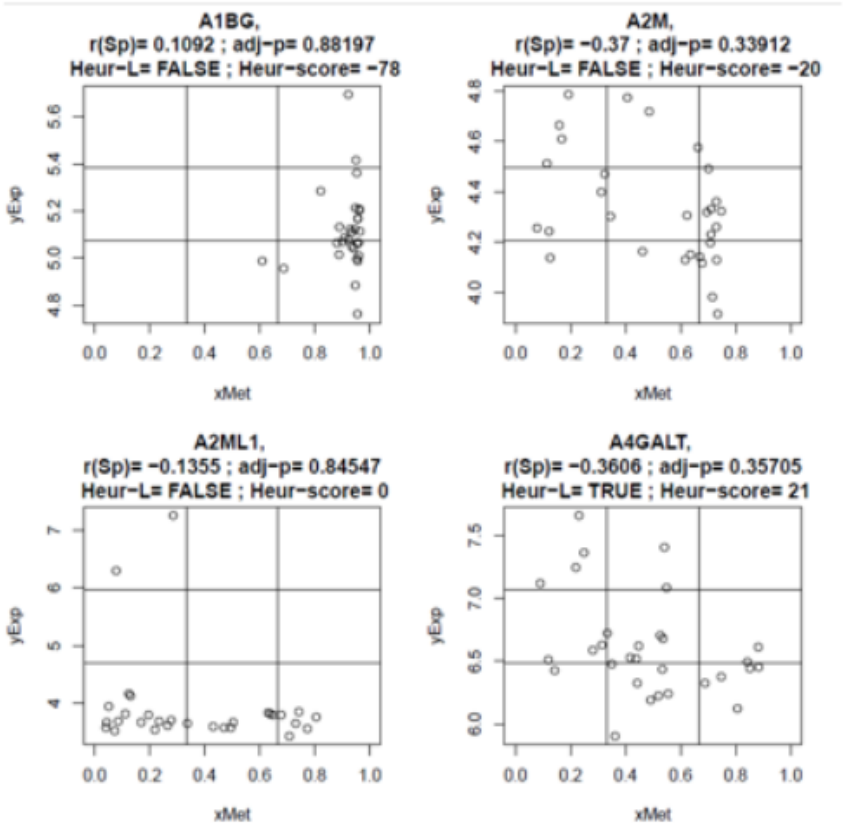


Software I

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Selection of genes potentially regulated by Methylation

Vall d'Hebron Institut de Recerca

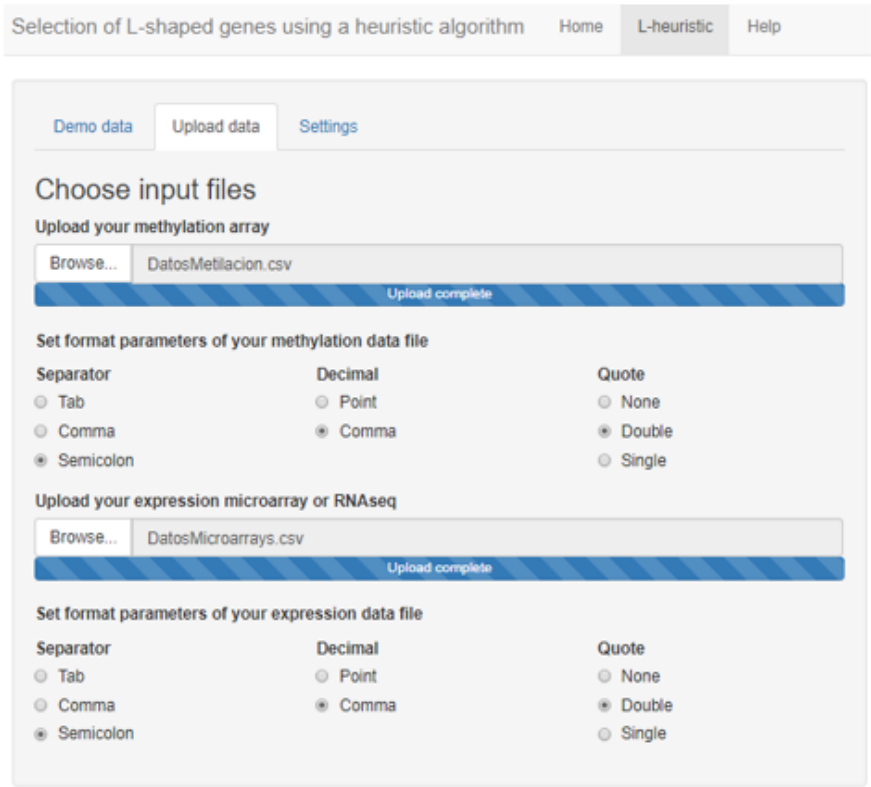


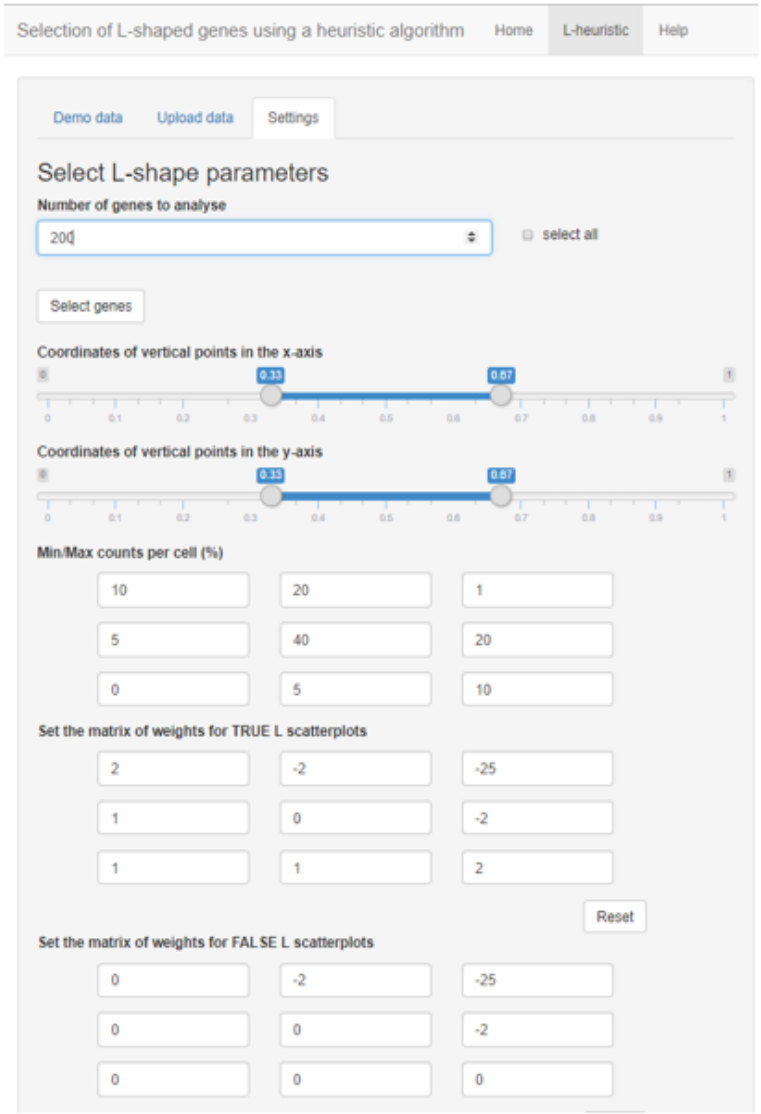
UNIVERSITAT DE BARCELONA

GRUP DE RECERCA EN ESTADÍSTICA I BIOINFORMÀTICA

GRBIO

Software II


Selection of L-shaped genes using a heuristic algorithm
Home
L-heuristic
Help
Demo data
Upload data
Settings
Choose input files
Upload your methylation array
Browse...
DatosMetilacion.csv
Upload complete
Set format parameters of your methylation data file
Separator
Tab
Comma
Semicolon
Decimal
Point
Comma
Quote
None
Double
Single
Upload your expression microarray or RNAseq
Browse...
DatosMicroarrays.csv
Upload complete
Set format parameters of your expression data file
Separator
Tab
Comma
Semicolon
Decimal
Point
Comma
Quote
None
Double
Single


Selection of L-shaped genes using a heuristic algorithm
Home
L-heuristic
Help
Demo data
Upload data
Settings
Select L-shape parameters
Number of genes to analyse
200
select all
Select genes
Coordinates of vertical points in the x-axis
0.33
0.67
Coordinates of vertical points in the y-axis
0.33
0.67
Min/Max counts per cell (%)
10
20
1
5
40
20
0
5
10
Set the matrix of weights for TRUE L scatterplots
2
-2
-25
1
0
-2
1
1
2
Reset
Set the matrix of weights for FALSE L scatterplots
0
-2
-25
0
0
-2
0
0
0

Software III

All Genes | L-shaped Genes | Non L-shaped Genes

L-shaped Genes: 4
Non L-shaped Genes: 196

Min/Max counts per cell

3	6	0
2	9	6
0	2	3

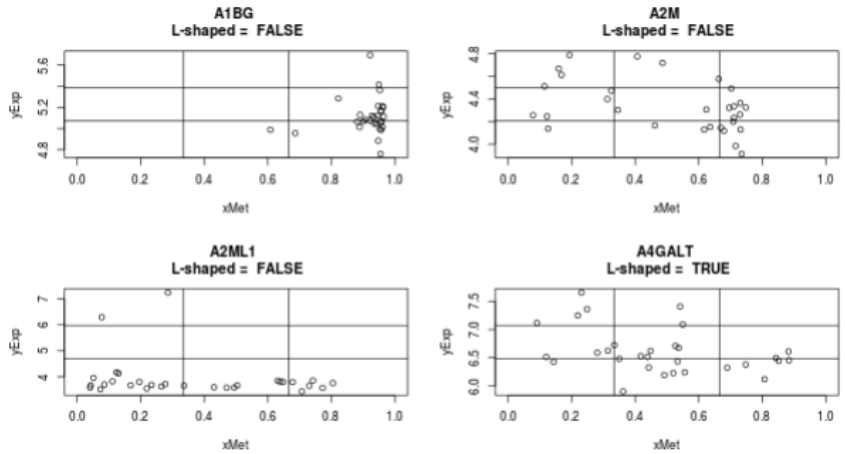
Download table

Show: 10

	logicSc	numericSc
A1BG	FALSE	-78.00
A2M	FALSE	-20.00
A2ML1	FALSE	0.00
A4GALT	TRUE	21.00
A4GNT	FALSE	-91.00
AAAS	FALSE	0.00
AACS	FALSE	0.00
AADAC	FALSE	-54.00
AADACL2	FALSE	-143.00
AADAT	FALSE	0.00

Some Scatterplots:

Download plots



Summary of results

Limitations

Conclusions

Acknowledgments

1. My group ESTBIOINFO at the GME department at the University of Barcelona and the research group GRBio, led by Guadalupe Gómez.
2. The Statistics and Bioinformatics Unit at Vall d'Hebron Institut de Recerca (VHIR).
3. The Nanomedicine and Molecular Oncology group at VHIR, led by Dr. Diego Arango.

Thanks for your attention!

References

- Bazzocco, Sara. *et al.* (2015) *Highly Expressed Genes in Rapidly Proliferating Tumor Cells as New Targets for Colorectal Cancer Treatment*. DOI: 10.1158/1078-0432.CCR-14-2457
- Liu, Y. and Qiu, P. (2012) *Integrative analysis of methylation and gene expression data in TCGA* IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)
- Sánchez-Pla, A., Ruiz de Vila, M.C:, Carmona, F., Bazzoco, S. and Arango del Corro, D. (2017). *Integrative analysis to select genes regulated by methylation in a cancer colon study* Trends in Mathematics 2017. DOI: 10.1007/978-3-319-55639-09
- Wilkinson, Leland, and Graham Wills. (2017). *"Scagnostics Distributions"*, JCGS.