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

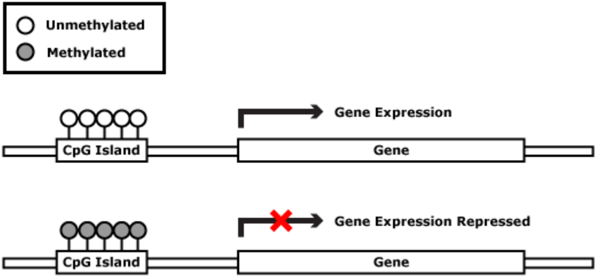
1. Introduction / Motivation
    - Genome-wide analysis of colorectal cancer
    - Objectives
2. Methods for selecting L-shaped patterns
    - A new algorithm
3. Results and Applications
4. Discussion and conclusions

Genome-wide analysis of colorectal cancer

- This work started as collaboration with a Molecular Oncology group working on Colorectal Cancer (CRC).
- 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.
- The researchers' approach was to look for *genes regulated by methylation* which could be considered possible therapeutic targets.

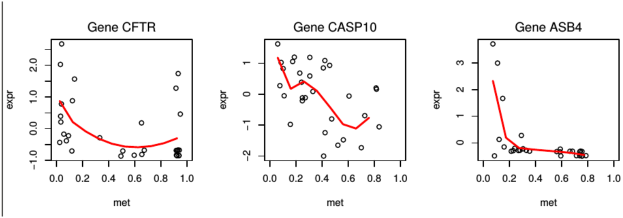
Methylation

- 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*



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

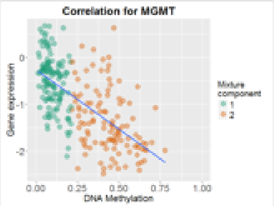
- Assuming the relation described above is true...
- Finding genes regulated by methylation is equivalent to finding genes whose methylation–expression scatterplot has an L–shape.
- There is a scatterplot *per* gene and thousands of genes: *Automatic methods for selecting interesting genes through their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

Objectives

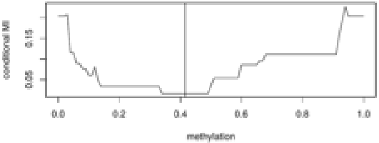
The main objectives of this work are:

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.

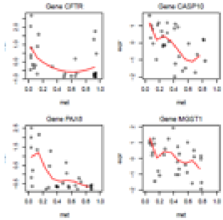
Previously applied methods



Naïve (r < 0)



Conditional Mutual Information



Splines regression+Clustering

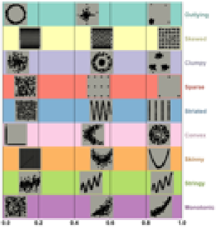
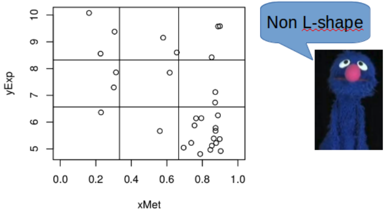


Figure 2: Example scatterplots and their scagnostics measures.

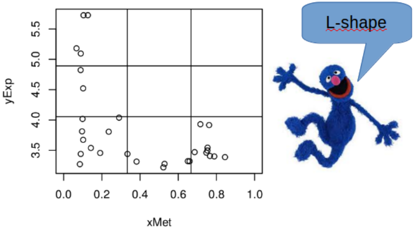
Scagnostics

What is an L-shape, whatsoever

- Go back to an intuïtive idea
- The more the values in the scatterplot move away from the axes the least L-shaped the gene is.



- The more the values cluster near the vertical and horizontal axes, the more L-shaped can be considered the scatterplot.



A penalization 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 left and lower cells of the grid.
    2. There must be a *maximum* number of points in the upper region (points there mean hypermethylation and hyperexpression, the opposite of what we are looking for).

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

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

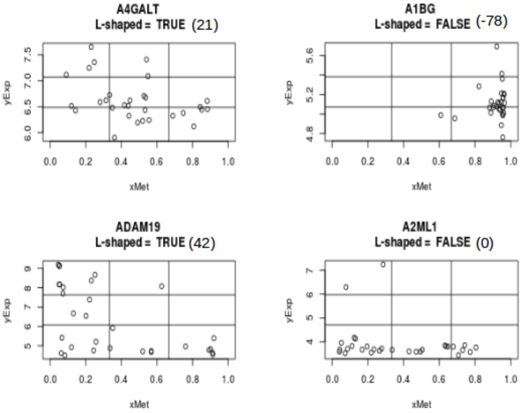
$$mMP = \begin{pmatrix} 10 & 20 & 0 \\ 5 & 30 & 20 \\ 0 & 5 & 5 \end{pmatrix}$$

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

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

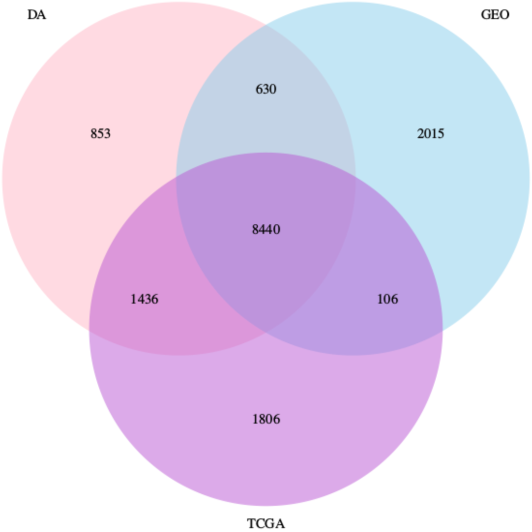


Data for the comparisons I

- 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	11788	223	Agilent	Illumina 27K
GEO	GSE25070	11191	25	Agilent	Illumina 27K
DA	Researcher's	11359	30	Affymetrix	Illumina 27K

Data for the comparisons II

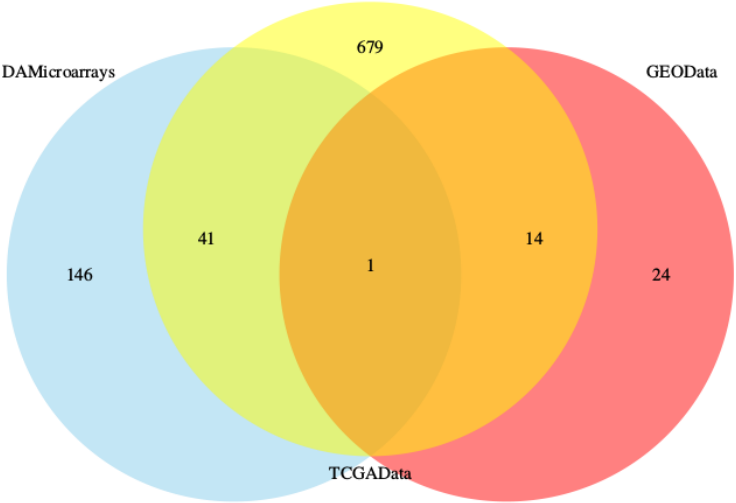


Results: Pathway Analysis of selected genes

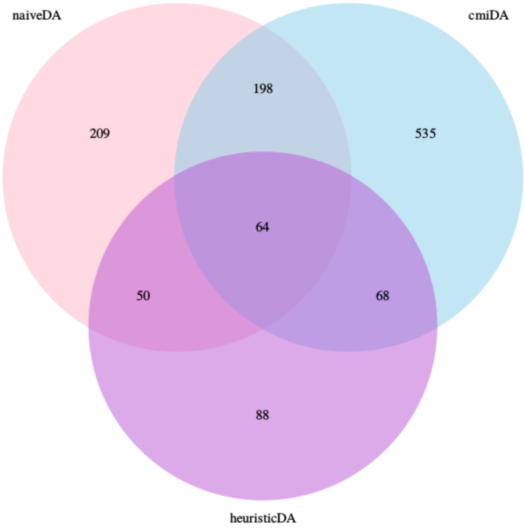
Two types of pathway analysis are undergoing

1. *Gene Enrichment Analysis* on each of the three gene lists analyzed. This analysis yields a set of functional categories
    - How similar are the gene sets obtained from the analysis?
    - *Which of these sets suggest there is regulation by methylation?*
2. *Pathway Equivalence Analysis* (`goProfiles`) of the three gene lists complemented with a three more random lists of same sizes.
    - Are the lists equivalent in their GO categories representation?
    - Are they more similar to each other than to random selected lists from the corresponding sets?

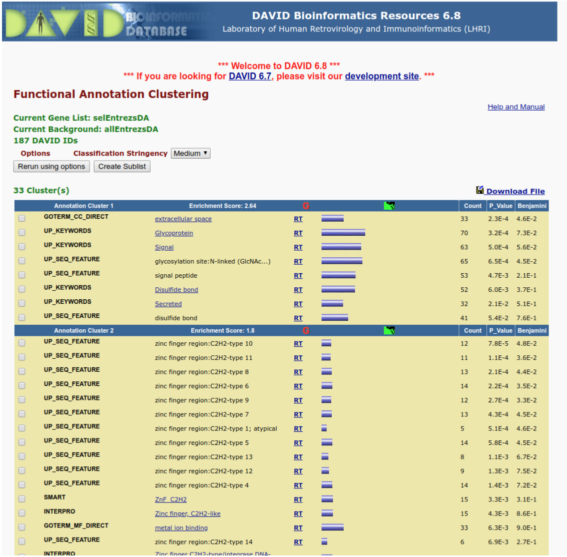
Results: Comparison between datasets



Results: Comparison between the methods



Results: Pathway Analysis



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID 6.8 ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Functional Annotation Clustering

Help and Manual

Current Gene List: selEntrezsDA
Current Background: allEntrezsDA
187 DAVID IDs

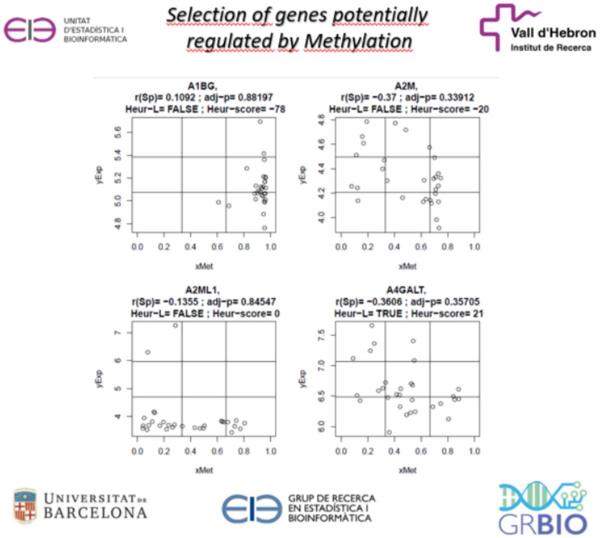
Options **Classification Stringency** Medium ▾

33 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 2.64	G		Count	P_Value	Benjamini
□	GOTERM_CC_DIRECT	extracellular space	RT		33	2.3E-4	4.6E-2
□	UP_KEYWORDS	Glycoprotein	RT		70	3.2E-4	7.3E-2
□	UP_KEYWORDS	Signal	RT		63	5.0E-4	5.6E-2
□	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		65	6.5E-4	4.5E-2
□	UP_SEQ_FEATURE	signal peptide	RT		53	4.7E-3	2.1E-1
□	UP_KEYWORDS	Disulfide bond	RT		52	6.0E-3	3.7E-1
□	UP_KEYWORDS	Secreted	RT		32	2.1E-2	5.1E-1
□	UP_SEQ_FEATURE	disulfide bond	RT		41	5.4E-2	7.6E-1
	Annotation Cluster 2	**Enrichment Score: 1.8**	**G**		**Count**	**P_Value**	**Benjamini**
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 10	RT		12	7.8E-5	4.8E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 11	RT		11	1.1E-4	3.6E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 8	RT		13	2.1E-4	4.4E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 6	RT		14	2.2E-4	3.5E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 9	RT		12	2.7E-4	3.3E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 7	RT		13	4.3E-4	4.5E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 1; atypical	RT		5	5.1E-4	4.6E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 5	RT		14	5.8E-4	4.5E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 13	RT		8	1.1E-3	6.7E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 12	RT		9	1.3E-3	7.5E-2
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 4	RT		14	1.4E-3	7.2E-2
□	SMART	ZnF_C2H2	RT		15	3.3E-3	3.1E-1
□	INTERPRO	Zinc finger, C2H2-like	RT		15	4.3E-3	8.6E-1
□	GOTERM_MF_DIRECT	metal ion binding	RT		33	6.3E-3	9.0E-1
□	UP_SEQ_FEATURE	zinc finger region:C2H2-type 14	RT		6	6.9E-3	2.7E-1
□	INTERPRO	Zinc finger C2H2-type/integrase DNA-					

Software I



UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Selection of genes potentially regulated by Methylation

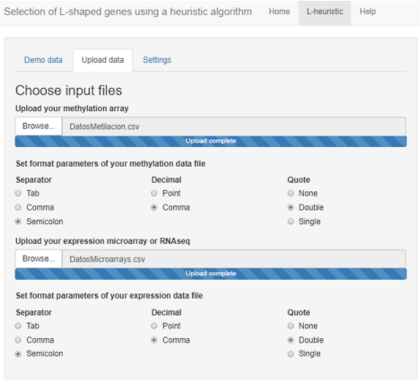
Vall d'Hebron Institut de Recerca

UNIVERSITAT DE BARCELONA

GRUP DE RECERCA EN ESTADÍSTICA I BIOINFORMÀTICA

GRBIO

Software II



Choose input files

Upload your methylation array

Browse... DatosMetilacion.csv

Upload complete

Set format parameters of your methylation data file

Separator	Decimal	Quote
○ Tab	○ Point	○ None
○ Comma	● Comma	● Double
● Semicolon		○ Single

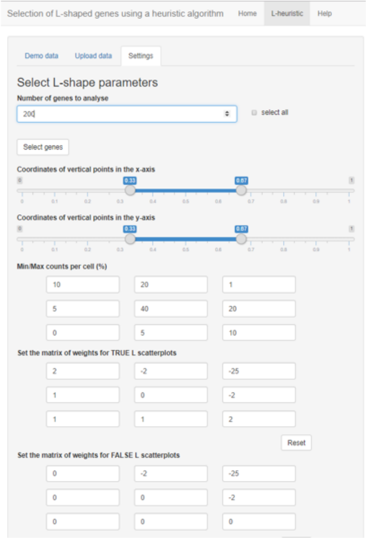
Upload your expression microarray or RNAseq

Browse... DatosMicroarrays.csv

Upload complete

Set format parameters of your expression data file

Separator	Decimal	Quote
○ Tab	○ Point	○ None
○ Comma	● Comma	● Double
● Semicolon		○ Single



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

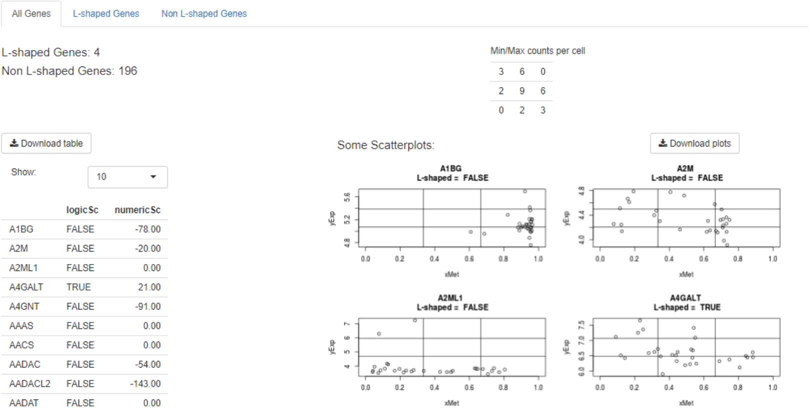
2	-2	-25
1	0	-2
1	1	2

Reset

Set the matrix of weights for FALSE L scatterplots

0	-2	-25
0	0	-2
0	0	0

Software III



L-shaped Genes: 4

Non L-shaped Genes: 196

Min/Max counts per cell

3	6	0
2	9	6
0	2	3

	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:

Summary of results

- The heuristic method is an intuitive approach to select L-shape genes.
  - It can be tuned easily so that different quantities of genes and L-shapes of distinct severity can be selected.
  - Indeed the method can be easily extended to detect other patterns such as vertical or horizontals clouds depicting distinct biological behavior.
- Right now the recommended approach is to apply the Naive and the Heuristic method and select the union of both sets.
- A tool for using the method is available at `http://cinna.upc.edu:3838/alex/Lheuristic/`.

Discussion I

- The standard approach is a poor one: The naive method looks for correlation (may miss L's) and relies on significance p-values for ρ, *that only tests if this is 0 or not*: **Improving this approach through, for instance, an appropriate scoring scheme is worthwile**.
- We look for a statistical approach (select lists of genes ...) but most of the times researchers end up looking at a few. In this case there was only **one gene** selected for further study, which showed a clear negative correlation, but not L-shape.
- The method shows several weaknessess
  1. The scoring system is ubiquous
  2. It is difficult to validate

Some Limitations I

- There is no TRUE/FALSE positive dataset, or at least it is very hard to consensuate one that is a list of genes related to CRC (or other diseases) and known to be regulated by methylation.
  - Sensitivity and specificity cannot be computed
  - The methods cannot be compared.
- TRUE and FALSE lists can be built manually or by simulation, but *How can we trust them?*
- Validation based on pathway analysis has come to be harder than expected because the number of GRM is often small, and belonging to distinct pathways (need very strong biological knowledge?)

Some Limitations II

- The scoring method works conditionally on the decision that the scatterplot is declared to have an L-shape
    - It is intuitive, easy to tune and easy to extend, because it allows to combine several scoring schemes.
    - What if the decision is wrong?
    - Doing inference based on the scores' distribution under H_0 is not possible (it should be if there were no binarization).
- The number of parameters is high
    - The method is flexible: easy to define which pattern (only L by now) and with which stringency (how restrictive) one wants to select.
    - May be hard to optimize even if TRUE positives/negatives were available.
    - Also hard to think of simulation scenarios.
- Altogether makes that, right now the methd can only be considered to be a descriptive tool.

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.