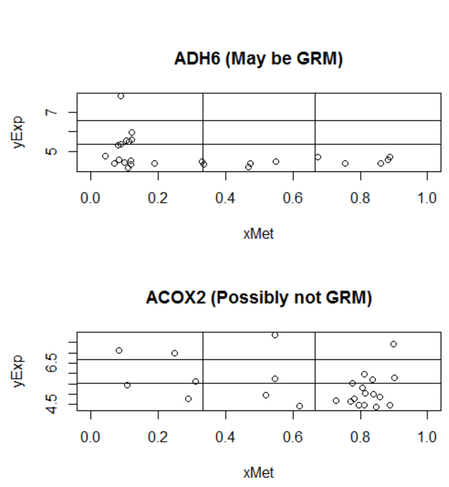
Integrative Analysis to Select Genes Regulated by Methylation in a Cancer Colon Study

Sánchez-Pla, Alex[1,3], Ruiz de Villa, M. Carme[1], Carmona, Francesc[1], Bazzoco, Sara[2], Arango, Diego[2]

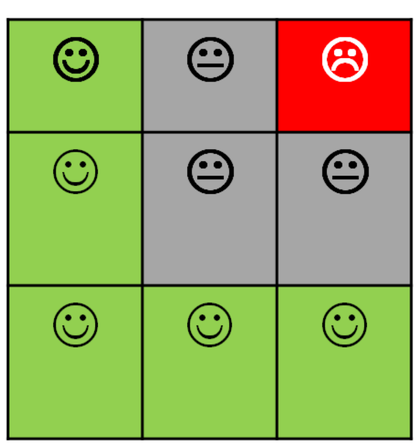
[1] Departament of Genetics Microbiology and Statistics, Universitad de Barcelona
[2] CIBBIM-Nanomedicine. Biomedical Research in Digestive Tumors, (VHIR), Barcelona
[3] Statistic and Bioinformatics Unit. Vall d'Hebron Research Institute. (VHIR). Barcelona

1 Introduction

- Methylation of genes involved in the oncogenic process is a key process contributing to tumor initiation and/or progression[4].
- Finding *Genes Regulated by Methylation* or GRM can lead to a better understanding and be a guide to finding new drug targets.
- This study originates in a work searching for colon cancer biomarkers [1]. Cell lines with increasing sensitivity to a chemotherapy drug, were analyzed with Expression and Methylation arrays. Finding GRM was used to search of candidate genes for new therapies.
- In cancer–related genes it is common to observe a decrease in gene expression associated with hypermethylation. Methylation is often described as a binary on-off signal ([2]) that is, when methylation is "off" the gene can express normally and its expression will be low or high, whereas when methylation is "on", the expression of the gene will be *repressed* and its values will tend to be low.
- As a consequence of this *high-methylation/low-expression* and *low-methylation/high-expression* relation plots depicting methylation and expression will show L–shape patterns so the strategy adopted will be to mine such plots and select those that have such a shape.



L-shape (GRM) vs non L-shape



Permitted, indistinct and forbidden regions
To score an L shape on a grid

2 Objectives

- To select an appropriate method to mine scatterplots extracted from a multiple high-throughput dataset formed by expression and methylation data and extract the desired patterns,
- To test the methods selected on a colon cancer dataset formed by a panel of cell lines derived from colorectal tumors.

3 Methods

3.1 Using 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.

3.2 Selection Based on Spline regression

- Spline regression [3] was also considered as a basis for scatterplot clustering applying the following algorithm:

1. Select genes with significant negative correlation.
2. For each selected gene fit a cubic splines regression model.
3. Obtain a distance matrix between all genes using the $1 - \rho$ distance computed on spline coefficients.
4. Perform a hierarchical clustering and
5. Select genes in the *L-shaped cluster(s)*.

3.3 Heuristic approach

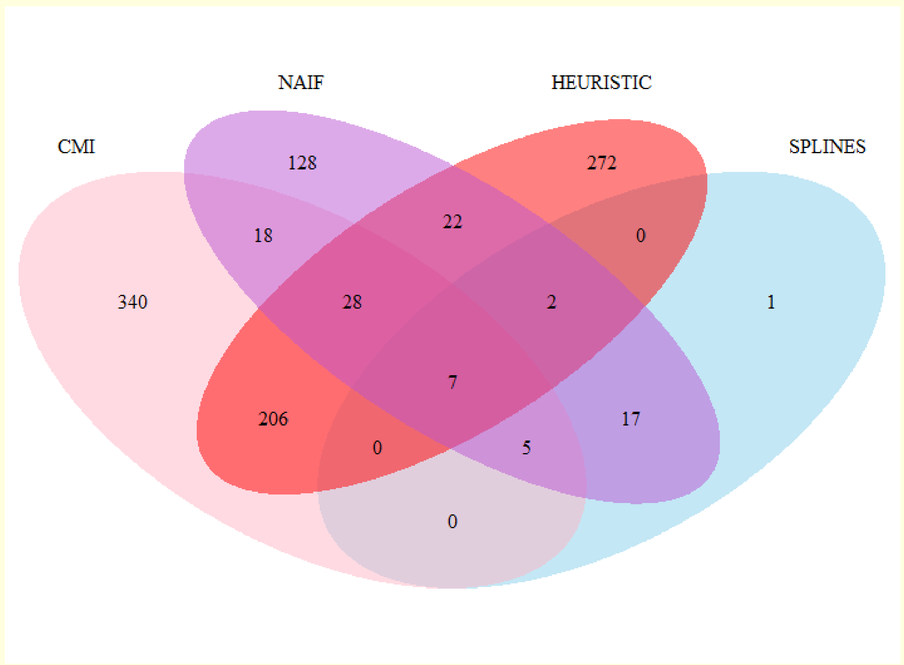
A heuristic method has been developed basing on imitating the visual selection of L-shapes (see figure)

- Overimpose a 3×3 grid on the scatterplot.
- Score points on each subgrid in such a way that
  - Points in permitted regions increase score
  - Points in non-desired regions decrease score
  - Points in non-allowed regions set score to $infty$.
- Use cross-validation to tune scoring parameters.

3.4 The Naive method

Traditionally selection has been based on searching for genes with negative correlation betweem expression and correlation. This "Naïve approach" is used to compare with other approaches.

4 Results and discussion



- All methods yield some L-shaped genes
- L-shape *is a surrogate for methylation*: Lack of empirical knowlege of methylated genes avoids, by now, to turn it into a supervised analysis problem.
- Candidate lists are being checked for biological significance related to methylation
- Most promising: Combine Naif + Heuristic.

[1] Sarah Bazzocco, Hafid Alazzouzi, M. Carme Ruiz de Villa, Alex Sanchez-Pla, John M. Mariadason, Diego Arango (2013) *Genome-Wide Analysis of DNA Methylation in Colorectal Cancer*. Submitted.

[2] Yihua Liu and Peng Qiu. (2012) *Integrative analysis of methylation and gene expression data in TCGA* IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)

[3] Jeffrey Racine. (2012) A primer on regression splines. `http://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf`

[4] B Sadikovic, K Al-Romaih, J.A Squire, and M Zielenska. Cause and Consequences of Genetic and Epigenetic Alterations in Human Cancer. *Current Genomics*, 9(6):394–408, September 2008.

Work partially supported by research project MTM2015-64465-C2-1-R (MINECO/FEDER) from Ministerio de Economía y Competitividad and by Grant 2014 SGR 464 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya.