Scatterplot analysis of expression and DNA methylation integration data

Berta Miro Cau

Director: Alex Sanchez Pla

Master en Bioinformatica i Bioestadistica

Area: Estadistica i Bioinformatica

Universitat Oberta de Catalunya

5 June 2018

.

FITXA DEL TREBALL FINAL	
Titol del treball:	Scatterplot analysis for the integrative analysis of expression and methylation data
Nom de l'autora:	Berta Miro Cau
Nom del consultor:	Alex Sanchez Pla
Nom del PRA:	Alex Sanchez Pla
Data d'entrega:	05/06/2018
Titulacio:	Master universitari en Bioinformatica i Bioestadistica
Area del treball final:	Estadistica i Bioinformatica
Idioma del treball:	English
Paraules clau	"L-shape pattern", methylation, "gene expression", integartion

Abstract:

Gene expression regulated by DNA methylation patterns has been long studied in relation to cancer. There exists a negative correlation between the expression of a gene and its methylation level. An integrative analysis of expression and methylation arrays was performed using three datasets for colorectal cancer: TCGA, GEO and own data. The datasets had over 11000 genes, 9000 of which were common.

Based on the preconception that methylation represses expression, we selected genes that showed an L-shaped expression and methylation scatterplot with 4 different methods. The first method used, naive, was based on a significant negative correlation. Another methodwas based on Conditional Mutual Information (CMI). A heuristic method was carried out by superimposing a grid on each scatterplot and weighing the cells according to an L-shape. Finally, a scagnostics selection analysis was based on 9 parameters defining the shape of scatterplots. The scagnostics needed to be used in conjunction with other methods for optimal results.

The accuracy, sensitivity and specificity were measured for the various methods and the one with the best diagnostic measures was the Heuristic, followed by the CMI and the naive. The one that fared lowest was the scagnostics.

The final gene list was obtained from a pool of all methodologies.It resulted in 179 target genes, mostly coding for ATP-binding, transcription and zinc finger proteins.

Aknowledgements

I would like to thank my supervisor for his long term suppport and guidance to get me were I am. I would also like to thank my family, specially during the very busy times. And finally, to my colleagues, that also became my friends. Thank you to all for seeing me through the good, the bad and the ugly. You're great!

Contents

List of Figures

List of Tables

1 Introduction

Cancer is an illness related to changes in the cell cycle, when cells go through uncontrolled cell division. Cancer cells behave differently than normal cells because:

NORMAL CELLS	CANCER CELLS
Controlled growth (contact inhibition)	Invasive
Repair or apoptosis when damaged or old	No apoptosis when damaged or old
Stick together in a group	Float away and metastasize
Mature	Immature
Functional	Malfunctioning (sometimes)
Controlled angiogenesis (growth, repair)	Continuous angiogenesis
Controlled by growth (tumor supressors)	Evade growth supressors
Energy source Krebs Cycle (oxygen)	Energy source Glycolysis (no oxygen)
Mortality	Immortality
	Ability to hide
Usual DNA and chromosome number	Abnormal chromosome number, mutated DNA

Table 1: Differences between normal and cancer cells

In the cell cycle regulation there are two types of genes: the oncogenes which are positive cell cycle regulators, and the tumor suppressors, which are negative regulators. Oncogenes are in the body as proto-oncogenes that become opartional if there is a mutation, or a change in expression, that makes them become overactive. Many proteins related to cell growth factors are proto-oncogenes.Tumor suppressors have the opposite effect, they reduce or even stop the activity of the cell cycle. Mutations in proteins related to reducing or stopping the cycle of an aberrant cell will not function, and the cell with damaged DNA will proliferate. The activation of proto-oncogenes or the inactivation of tumor suppressor genes is what will produce cancerous cells (**??**).

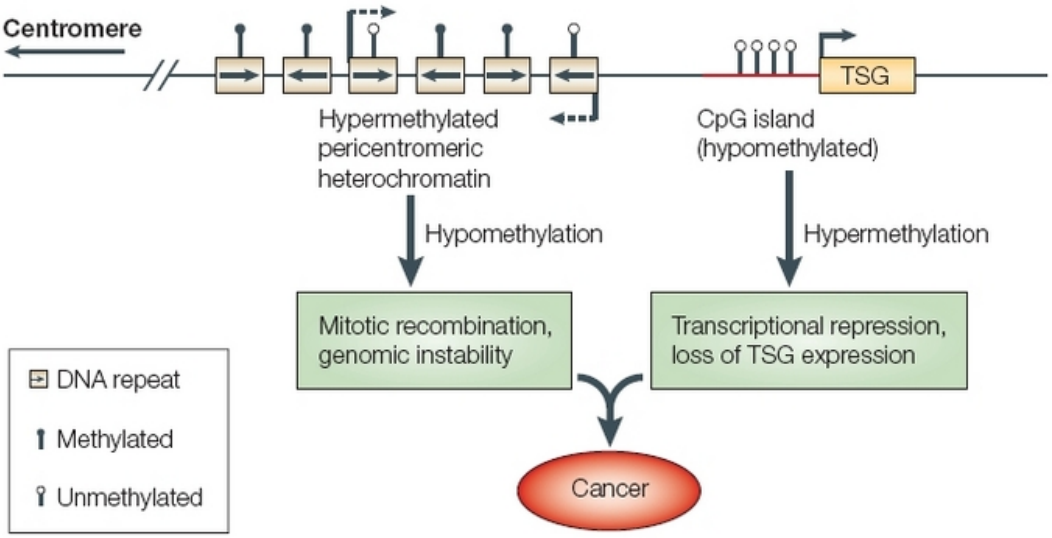

Figure 1: Diagram representing a DNA region in a normal cell, with hypermethylated and hypomethilated regions and their reationship to cancer ([?]).

There are many ways in which gene expression is regulated in eukaryotes. DNA methylation occurs in CpG dinucleotides and is one of several epigenetic mechanisms that cells use to control gene expression. It has been long known the involvement of methylation in numerous cellular processes ([?]; review in [?]), including embryonic development ([?]), X-chromosome inactivation and preservation of chromosome stability ([?]). Methylation has a critical role in gene expression and cell differentiation, and most research has been focused on tumor repressor genes, which are often silenced in cancer cells due to hypermethylation ([?]). However, cancerous cells tend to have hypomethylated genomes, but show hypermethylation in certain genes related to processes such as cell cycle regulation, tumor cell invasion, and DNA repair among others; when compared to normal cells ([?]). The generation of large amounts of data has posed a challenge, which has resulted in various methodologies being developed for the analysis and interpretation of such data.

Colorectal cancer (CRC) is the third most common cancer in the world in men, and the second one in women. It is a type of cancer that is related with development and income ([?]). Various studies have identified methylated genes ([?], [?]) and hypermethylation of promoters ([?], [?], [?]). Changes in methylation profile of the CpG islands are a very powerful tool that can be used for biomarker identification ([?], [?], [?], [?]). In CRC, it has been found that both methylation and demethylation vary along the cancer progression. Methylation is progressive, but also, genes change their methylation status at different stages, going from methylated to demethylated and viceversa([?]). That implies that methylation is a dynamic process in cancer.

One of the aims of using epigenetic alterations in cancers it to develop specific therapies to combat cancer in a les invasive manner ([?]). In addition to that, the identification of genes highly regulated by methylation, would be a useful tool for early detection ([?]).

1.1 Project context and justification

Tradicional cancer detection techniques are not accurate enough to predict the occurrence of cancer, and it relies on the doctor's oppinion to decide on the treatment. Nowadays, there are various high-throughput techniques available to study cancer by analyzing gene expression and methylation patterns, i.e. CpG islands analysis through microarray technology, Reduced Representation Bisulfite Sequencing. These techniques can provide biomarkers that may be able to detect and predict cancer with better results. However, the large numbers produced in the data still pose an analysis and interpretation challenge ([**?**]).
As new large datasets will continue to be produced, the development of new analysis and visualization techniques will keep developing. There are various softwares available for the analysis of gene expression in relation with methylation: *PiiL*, *MEXPRESS*, *methylPipe*, among others. *PiiL* is a genome web browser that allows for the visualization of methylation and gene expression at various levels. It allows the selection of genes or CpG island, single or multiple samples. It then draws pathways, selects genes by methylation and expression pattern ([**?**]). *MEXPRESS* ([**?**]) is a package that has a web interface that allows the user to visualize expression and methylation data from genes in the TCGA data. The visualization collocates for each selected gene, CpG islands, with transcripts expression together with other clinical values such as gender and age. The tool also generates p-values in relation to the variables specified. Another application developed in R language is *methylPipe* ([**?**]. This software package works together with *compEpiTools* that allow integrative analysis of diverse epigenomics datsets. The first one performs an analysis of the methylation high-throughput data, and the second allows for the integration of these data with other sources. They work both together and independently.

There is an direct association between expression an methylation that varies in different regions of a gene. High methylation is associated with high expression inside the gene, however, this correlation is inverse in the gene promoters (**??** which are the regions upstream of a gene involved in regulation of expression)([**?**]). In addition to that, CpGs shores, found in close proximity to CpG islands, are associated with cancer specific methylated regions ([**?**]). A recent paper studied correlations between gene expression and methylation in cancerous and non cancerous tissues ([**?**]). In this paper they observed that there are correlation diffferences between the the cancer and normal tissues in breast cancer. In addition to that, the found out that the genes involved in the differentially significant correlations were not necessarily neither differentially expressed nor differentially methylated. Therefore, they concluded that a gene selection by correlation may detect a different set of genes for biomarker generation that would not have been found in usual differential studies.

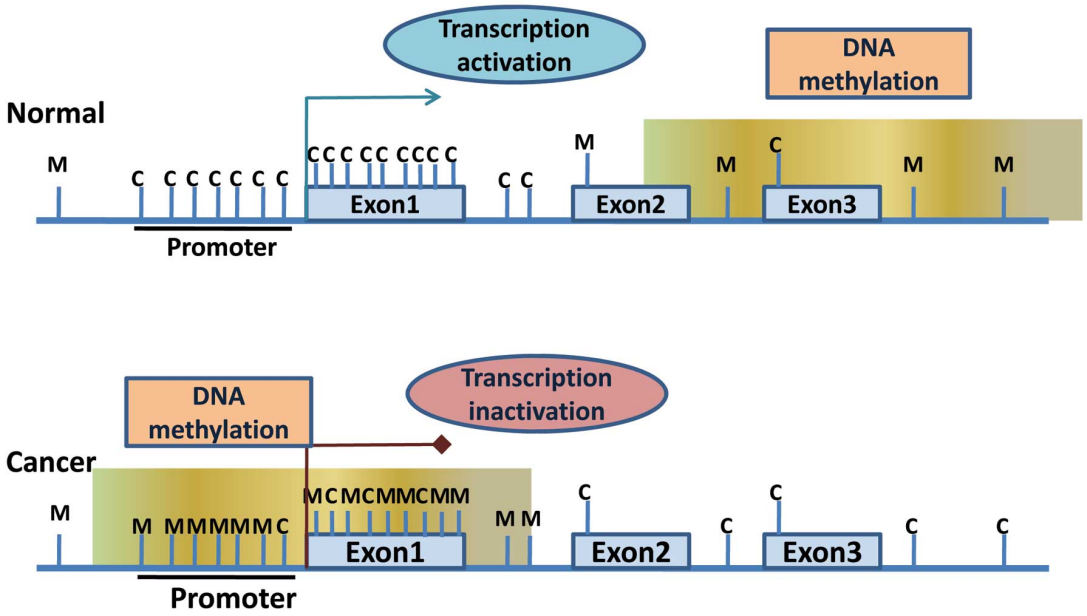

Figure 2: Structure of a gene with differently methyated regions according to normal or cancerous.

This work aims to explore the negative relationship between expression and methylation by reviewing some of the existing methodologies to detect genes with L-shaped scatterplots, and fine tune the parameters for the optimal selection of genes regulated by methylation. The final aim of this thesis is to build a package and develop an application to identify a short-list of potential candidates related to CRC from medical data.

1.2 Data used in the analysis

The data used to carry out the research has originated from 4 different datasets, 2 publicly available, one from a collaborator and another one used for testing of parameters has been artificially generated. The 2 publicly available datasets were obtained from The Cancer Genome Atlas (TCGA) and from the Gene Expression Omnibus datasets from NCBI (GEO, GSE25070 for the expression data and GSE25062 for the methylation data). These datasets are readily available from the respective websites. The collaborator's data has been then analyzed based on the L-shaped methodology tested, with the final aim to obtain biomarkers for CRC.

All datasets consist of expression microarray data and methylation data (Illumina 27k methylation array).The TCGA dataset has originated from clinical data and has gene expression values for 223 samples. The GEO dataset has 25 samples and the researchers dataset has 30 samples which are the same ones as the TCGA. The TCGA dataset has 11788 genes, the GEO has 11191 genes and the experimental dataset has 11359 genes in it. Most genes on these datasets are common.All datasets have been previously normalized

before download or pre-processed in other projects prior to this anlysis. The data has also been formatted for analysis. The formatting has ensured that both methylation and expression matrices have the same genes and samples and that missing values were removed from the dataset, since some functions implemented will not accept missing values. In addition to that, the order of the genes and the samples has to match in both matrices.
Finally, the creation of an artificial dataset has helped in the standardization of the scagnostics method. This dataset has been constructed in 2 parts, one with true L-shaped genes, by randomly creating data with a logarithmic distribution. The non-L-shaped data has been created by using a random normal distribution. Both artificial datasets have been created with R software functions.

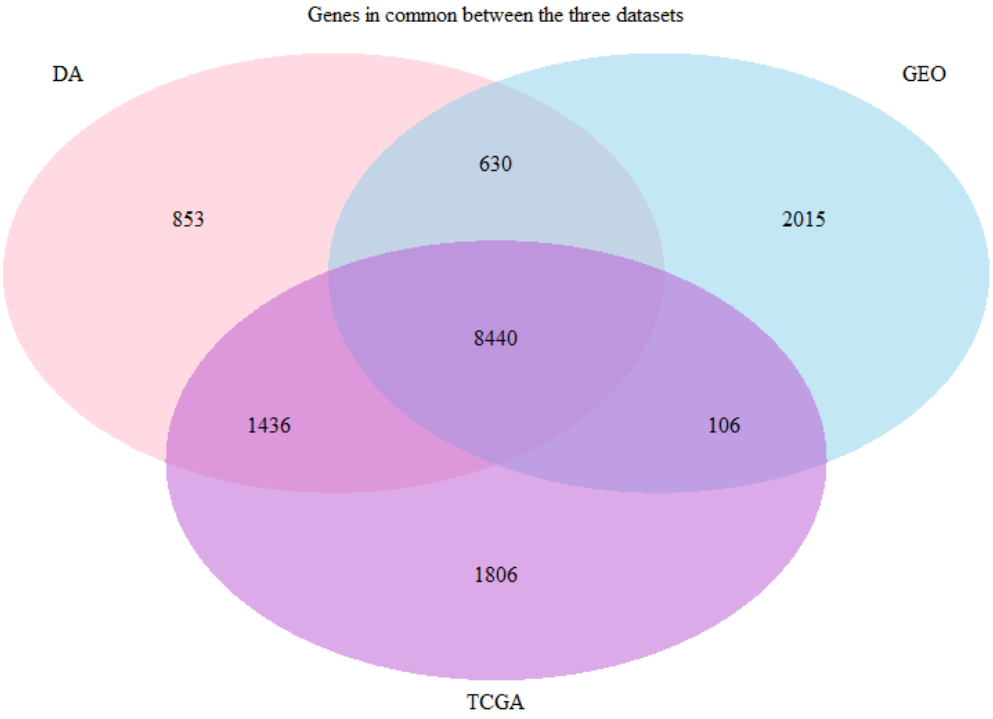

Figure 3: Venn diagram representing the overlapping of genes found between the 3 datasets used for analysis.

1.2.1 Generation of the artificial dataset

The artificial data was generated with the *rnorm* function in R. Two datasets were generated, one for the L-shaped genes and another one for the non L-shaped genes. For the L-shaped genes 2 parameters were adjusted to force the L in the scatterplot. The parameters were the distribution and the shape of the slope. The distribution was set between 0.3 and 0.7, with a step of 0.09 units. The slope was set between 0.1 and 0.3 with a step of 0.03 units. The flexibility in the parameters was introduced to have diverse L-shaped scatterplots; however they all had a "perfect' L-shaped distribution. To include more variablility int he "true" L-data, another artificial dataset was created with negative correlations, ranging between $r = (-0.5, -0.9)$. These two "true" positive datasets were merged into one.
The non L-shaped were created with the same function, but with random sd and random mean, to generate more diversity in the scatterplots.
Various dataset combinations were tested to create a realistic sample despite being artificially generated:

- Equal numbers of L and non L-shaped data
- Different dataset sizes
- Unequal combinations of L and non L-shaped scatterplots

For the final dataset used for testing and method evaluation, with the data for both L and non L artificial gene values, a subset of 30 L-shaped "genes" had been mixed with 200 and 500 random "genes" to test for diagnoastic measures to evaluate the various methods. Smaller sample numbers did not seem realistic to test. However, there are no changes in the parameters between the 2 datset sizes and therefore the 200+30 sample has been selected to evaluate the methodologies and to optimize the parameters for the heuristic and for the scagnostic methods.

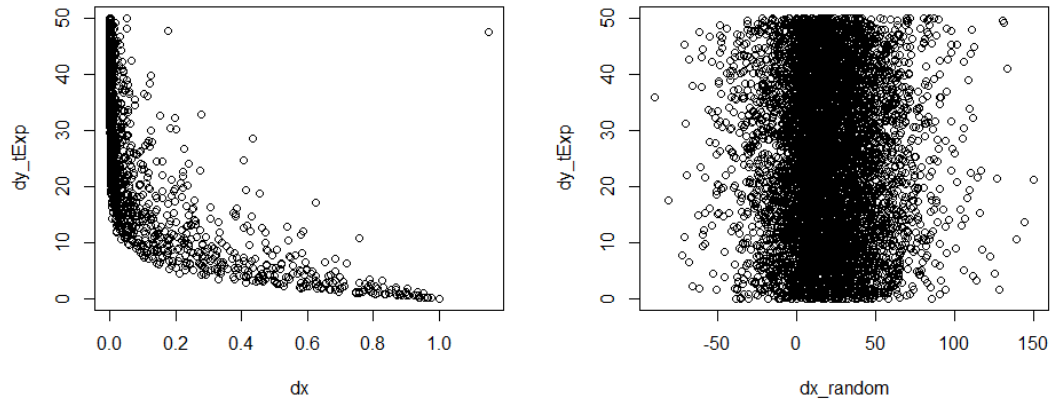

Figure 4: scatterplots representing the L-shaped and the non L-shaped artificial data.

1.3 Project Objetives

1. To evaluate and fine-tune existing methods for selection of L-shaped genes potentially regulated by methylation and apply the tool for the selection of genes related to colorectal cancer

2. To create an application for cancer driven gene selection.

1.3.1 Specific Objetives

The project objectives are detailed below:

1. To construct and/or evaluate a tool to identify L-shaped genes as potential biomarkers

   (a) To identify existing methods and test them with sample lists of genes and identify optimal parameters

   (b) To test the methods against other datasets, including L-simulation data.

   (c) To fine tune the parameters with the selected method(s).

   (d) To analyze the biological significance of the candidate gene list.

2. To create an application to select genes regulated by methylation.

   (a) To compile all the methodology developed in an R package.

   (b) To finalize an application tool (Shiny) widely available for the selection of genes potentially regulated by methylation.

To achieve the objectives, the methods will be tested against various relevant datsets, the most optimal parameters will be evaluated and a final methodology and working protocol will be developed for the final creation of an r package.

1.4 Focus and methodology

This work is the continuation of previous theses that focused on the identification and optimization of various methodologies for the classification of L-shaped expression/methylation scatterplots. These L-shaped scatterplots are associated with genes the expression of which is regulated by methylation. Previous methodologies tested include a naïve method, a CMI method, a heuristic method and a scagnostics method. These four will be explained in more detail later in the corresponding section.

The methodological design includes the reviewing and testing of the existing methods; which involves a negative correlation method, a Conditional Mutual Information (CMI) and a heuristic method. These have been evaluated, tested and fine-tuned (if required) with the above mentioned datasets. An additional method called “scagnostics” has also developed and used for analysis. The method consists of scagnostics analyses of correlation plots ([**?**]). These 4 methods have been tested on an artificial dataset and also with

manually validated L-shaped gene lists. From these analyses, the optimal parameters have been identified, and then these same methodologies have been tested with other 3 datasets. The last step has been to fine-tune the parameters of all methods. These parameters and their optimization will be further discussed in the remaining chapters.

After the first parameter optimization has been completed with the artificial dataset, various lists of genes have been obtained from each methodology. At this point, another test of the parameters with visually selected L-shaped genes has also been carried out. The lists obtained have been compared between them, including the intersections of lists from different methods. These lists then have been used to evaluate the biological significance of the L-shape selection methodology. From these results, the usage of all methods has also been evaluated. Next, a working protocol describing the approach for selection of target genes potentially regulated by methylation has been developed (**??**.

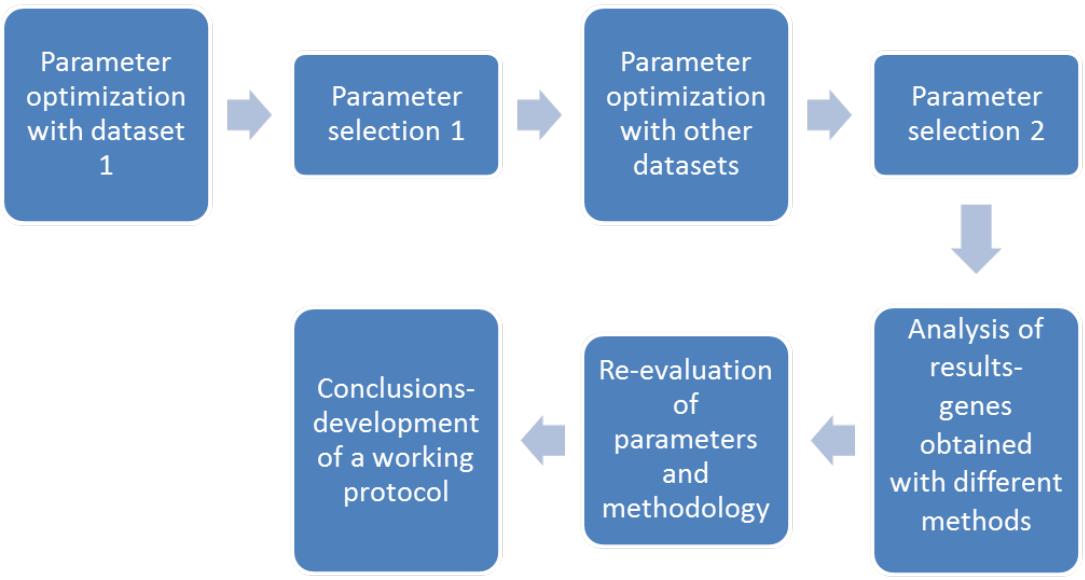

Figure 5: Flow chart sythetising the steps followed for analysis.

The final step is to convert the working protocol into an R package that will be uploaded in R Bioconductor. In addition, a Shiny (Shiny:2018) web application for online and user-friendly parameter tuning will also be developed later on.
The application will be developed in R free software environment also ([**?**]).

1.5 Work plan

The project was organized according to the following Gantt chart:

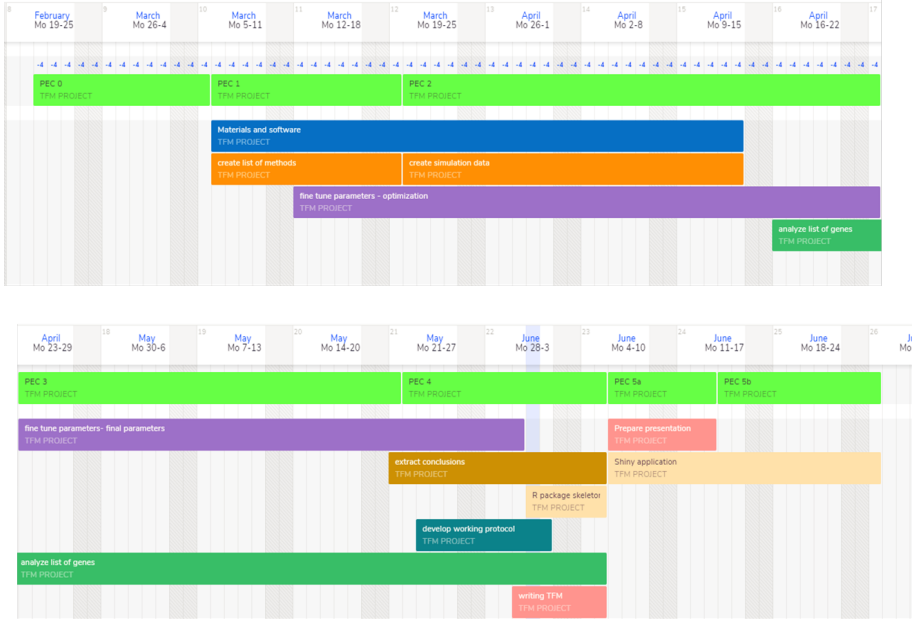

Figure 6: Gantt chart representing the final timeline for completion of the work.

The various PECs are represented above the rests of the tasks to indicate the various tasks accomplished during each deliverable.

Some adjustments were made from the original planning. In addition to that, a final task (the application Shiny) will be completed out of the evaluation period.

1.6 Description of outputs

1. **Method.** Optimized scagnostics functions for the selection of L-shaped genes based on 4 datasets
2. **Method.** Optimized scagnostics functions for the selection of L-shaped genes based on 4 datasets
3. **Product.** List of genes potentially influenced by methylation
4. **Product.** Chromosomal location of genes potentially regulated by methylation collocated to the existing CpG islands.
5. **Product.** Compilation od functions in an R package for the identification of L-shaped genes.

1.7 Brief description of the remaining chapters

The remaining chapters include the following:

1. **Introduction.** Context, focus, scope and data used

2. **Optimization of parameters.** Description and analysis of the methods tested
3. **Results.** Description of results between methods, and between datasets; brief biological significance
4. **Output.** Package development
5. **Discussion and conclusions**
6. **Future work**

2 Optimization of parameters

The relationship between gene expression and methylation has been associated with cancer; and therefore, the study of this relationship has produced fruitful results. The association between gene expression and DNA methylation in the CpG islands in particular has been long studied; and as a result, negative correlations have been found to relate to cancer driven mechanisms ([**?**]). A previously developed method was the selection of genes with an L-shape association between the expression and the methylation datasets ([**?**]). In this research, they focused on the CMI and on another method based on spline regression. They observed that the first method would detect L-shaped genes more accurately in big datasets. On the other hand, the spline clustering was not size dependent, but it would yield a smaller number of samples. Other research exists that aimed to identify genes regulated by methylation according to the expression methylation patterns; however, they only use a particular methodology like the CMI ([**?**]) with positive results. Another paper focused on the identification of genes regulated by methylation through unsupervised clustering techniques to identify CRC subtypes was able to confirm existing subtypes that clustered together as well as to define new subtypes or classifications ([**?**]).

Selection of L-shaped genes was approached based on 4 different methods, which depended on a variety of parameters. The methods tested were as follows:

1. Naive
2. CMI
3. Heuristic
4. Scagnostics

Changing the parameters affects not only the final the number of genes called "L-shaped", but also the genes themselves were different. For that there is a need to identify an optimal set of parameters for each method, such that some diagnostic measures like as accuracy, sensitivity or specificity that could be improved. For that purpose, the artificially generated data has been used in all methods except with the heuristic.

However, the artificial datset is too perfect to adjust to the reality of the gene expression and methylation data. Therefore, a set of L-shaped genes from real data has also been used for parameter calibration. However, a set of "TRUE L-shaped Positives" and "TRUE L-shaped Negatives " has not been possible to obtain. The reason for that is that none of this genes has been proven to be related to methylation.
A way around it was to visually inspect the genes' scatterplots and select a set of data that were clearly L-shaped and another set with opposite characteristics (non-L-shaped). This methodology, however, incurred in human error, since the differentiation between L and non L is not that clear. Moreover, the amount of genes with an L-shape in a particular genome is not expected to be very high.

2.1 Naive method

The naive method is based on a negative and significant correlation to identify L-shaped scatterplots of the expression and methylation data. This method uses a correlation matrix function; where given two matrices X (m,n) , Y (m,n) this function can compute Pearson and Spearman correlation coefficients and also their significance p-values. This computation is for every pair of row vectors. The function has the following parameters:

- **X** matrix with methylation data
- **Y** matrix with expression data
- **adj** is a logical variable indicating if the p-value returned should be adjusted or not. The default is set to TRUE, which will then return an adjusted p-value.
- **pValCutoff** the upper limit to be used for the p-value. The default is 0.05.
- **rCutoff** is the upper limit to be used for the correlation coefficient. Default is 0, which means thereis no cut off.
- **sortByCorrs** is a logical that if it is TRUE, results will be ordered in ascending order by p-value. In the function, the default is set to FALSE.

The parameters that can be adjusted for optimization of the performance measures are "the pValcutOff" and the "rCutOff". This method had been previously set to an optimal adjusted p-value of 0.25 and a r coefficient cut off of -0.5. These parameters produce the following results for our three datasets:

TCGA	**GEO**	**researcher**
72	190	456

Table 2: Naive results with the TCGA, GEO, and researchers' datasets

The number of genes obtained was largest with the researchers' dataset, probably including many more genes than the expected to be regulated by cancer research.

The accuracy, sensitivity and specificity calculated with the artificial dataset resulted in the following values obtained for accuracy were of 100%, for sensitivity of 100%, and for specificity of 100%.

The accuracy, sensitivity and specificity were also calculated with a visually selected list of true L-shaped positive genes and with a subset of the true non L-shaped positive genes (from the expressiion/methylation scatterplots). The values obtained were for accuracy were of 72%, for sensitivity of 44%, and for specificity of 64%.

As an example, if we wanted to improve these values we could for example increase the r coefficient to -0.7. However, this has a negative effect, since only 5 genes of 50 are detected as true L-shaped. This could be explained by the fact that most genes have a r coefficient below -0.6 (**??**).

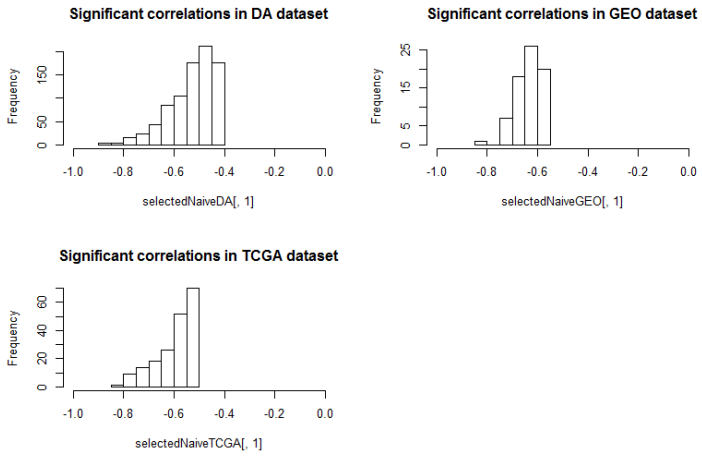

Figure 7: Distribution of the correlation coefficients of the selected genes with ajusted p-value of 0.25 and r coeff of -0.5.

2.2 Conditional Mutual Information method (CMI)

The Conditional Mutual Information method was based on the expression and methylation values computed at different points between 0 and 1 reached a minimum. This minimum should be small enough according predefined thresholds. This minimum was considered to be the cutoff point for methylation.

The cMI function computes cMI values for different t values, from 0 to 1 and a step of 0.01. The output is stored in a data frame. For each gene, the optimal threshold is the t-value that results the minimum CMI. This function has the following parameters:

- **X** is a matrix with methylation data
- **Y** is a matrix with expression data
- **h** is a number used for tuning of the kernel width, and empirically set at a default of $h = 0.2$.
- **smallR** numeric value representing the ratio mincM $I(t)/cMI(0$. Default is 0.25.
- **minCMI** Minimum value of unconditioned CMI. The default is set at 0.1.

L-shaped genes (regulated by methylation) are selected according to three conditions (parameters were chosen according to a random permutation test as in [**?**]): 1. Ratio mincM $I(t)/cMI(0)$ must be small enough, $r < 0.25$. 2. Minimum value of unconditioned CMI must be large enough, $cMI(0) > 0.1$. 3. Expression values must be higher on the left side of the plot than on the right side.
These parameter values had been previously optimized, and the values used as default in the function were found to be optimal. Therefore, this function has not been calibrated for the present work.

The number of genes identified as L-shaped with the CMI method is represented in **??**:

TCGA	**GEO**	**researcher**
301	263	795

Table 3: CMI results with the TCGA, GEO, and researchers' datasets

The accuracy, sensitivity and specificity calculated with the artificial dataset were: accuracy of 75%, sensitivity of 41%, and specificity of 69%.

The values obtained were a 84% for accuracy , for a 68% for sensitivity , and a 76% for specificity.

2.3 Heuristic method

The heuristic method is intended to select L-shaped scatterplots by overimposing a grid on it and defining certain regions which have to (or do nott have to) contain a minimum (or maximum) percentage of points if the scatterplot is to be called L-shaped. a second layer involves the method computing a scoring that favors selected cells, which then score positively and penalizes cells off the L-region that score negatively. An appropriate setting of scores and weights should yield positive scores for L-shaped scatterplots and negative scores for those that are not. One of the main interests of this approach is the possibility to tune the selection process by changing the scoring parameters.

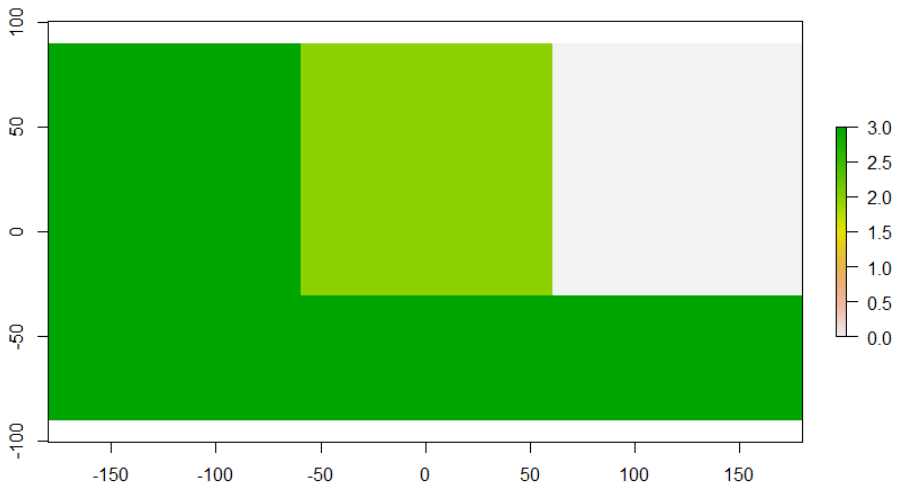

Figure 8: Graphic representation of the 3x3 grid for the heuristic method.

For this analysis a 3x3 grid is created over the scatterplot and the function calculates the frequency of point on each cell of the scatterplot with the calcFreqs function, followed by a conversion of frequencies to points to facilitate the scoring. The next step is to apply weights to each cell of the grid to select the genes based on the closest L-shaped pattern distribution. Two final weights matrices are used, one to identify L-shaped distributions and one to penalize non L-shaped scatterplots. This is represented on **??**, where the greener the color, the more favorable scored the cells are.

The parameters for this function are as follows:

- **mets** is a matrix containing methylation values.

- **expres** is a matrix containing expression values.
- **aReqPercentsMat** is a matrix of minimum maximum percentage of counts to have in a given cell.
- **aWeightMifL** is a matrix of weights to score the previous counts only if the scatterplot has been classified as L.
- **aWeightMifNonL** is a matrix of weights to score the previous counts only if the scatterplot has been classified as non-L.
- **x1, x2**, are the coordinates of the vertical points in the X axis. Since it is expected to contain methylation values that vary between 0 and 1 after normalization, the default values are 1/3 and 2/3.
- **y1, y2**, are the coordinates of vertical points in the Y axis. Leaving them as NULL assigns them the percentiles of yVec defined by 'percY1' and 'percY2'.
- **percY1, percY2** are the values used to act as default for 'y1'and 'y2' when these are set to 'NULL'.

A set of matrices with percentages, maximum and minimum counts, and weights for L and non L genes were the following:

10	20	1
5	40	20
0	5	10

Table 4: Percentatges matrix

3	6	0
2	12	6
0	2	3

Table 5: Counts matrix

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

Table 6: Weights matrix L genes

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

Table 7: Weights matrix non L genes

With the artificial dataset, accuracy was of 100%, for sensitivity of 100%, and for specificity of 100%.

With the visually selected L- positives, the accuracy with these parameters was 73%, the sensitivity was 46%, and the specificity was 69%.

Varous combinations were tested and finally a set of matrices with different percentages and weights for L and non L genes were tested again:

2	20	5
1	40	20
0	1	2

Table 8: Percentage matrix

1	6	2
0	12	6
0	0	1

Table 9: Counts matrix

2	-2	-sample size/5
1	0	1
1	2	2

Table 10: Weights matrix L genes

0	-2	-sample size/5
0	0	-2
0	0	0

Table 11: Weights matrix L genes

The accuracy with these new parameters was 95%, the sensitivity was 91%, and the specificity was 90%.

These last matrices were then found to produce an optimal number of genes from each dataset. Changing these parameters any further did not improve the results, neither quantitatively nor qualitatively.

TCGA	GEO	researcher
442	39	188

Table 12: Heuristic results with the TCGA, GEO, and researchers' datasets

2.4 Scagnostics method

The scagnostics is a method to characterize **scatterplots** according to a group of variables that are used for **diagnostics**, from here it comes the name. "Scagnostics is a Tukey neologism for the term scatterplot diagnostics. Scagnostics are characterizations of the 2D distributions of orthogonal pairwise projections of a set of points in multidimensional Euclidean space. These characterizations include such measures as density, skewness, shape, outliers, and texture ([**?**])".
The method provides 9 coefficients for each gene that describe the scatterplot. These are: outlying, skewed, clumpy, sparse, striated, convex, skinny, stringy and monotonic. In **??** there is the best graphical representation that characterized each parameter.

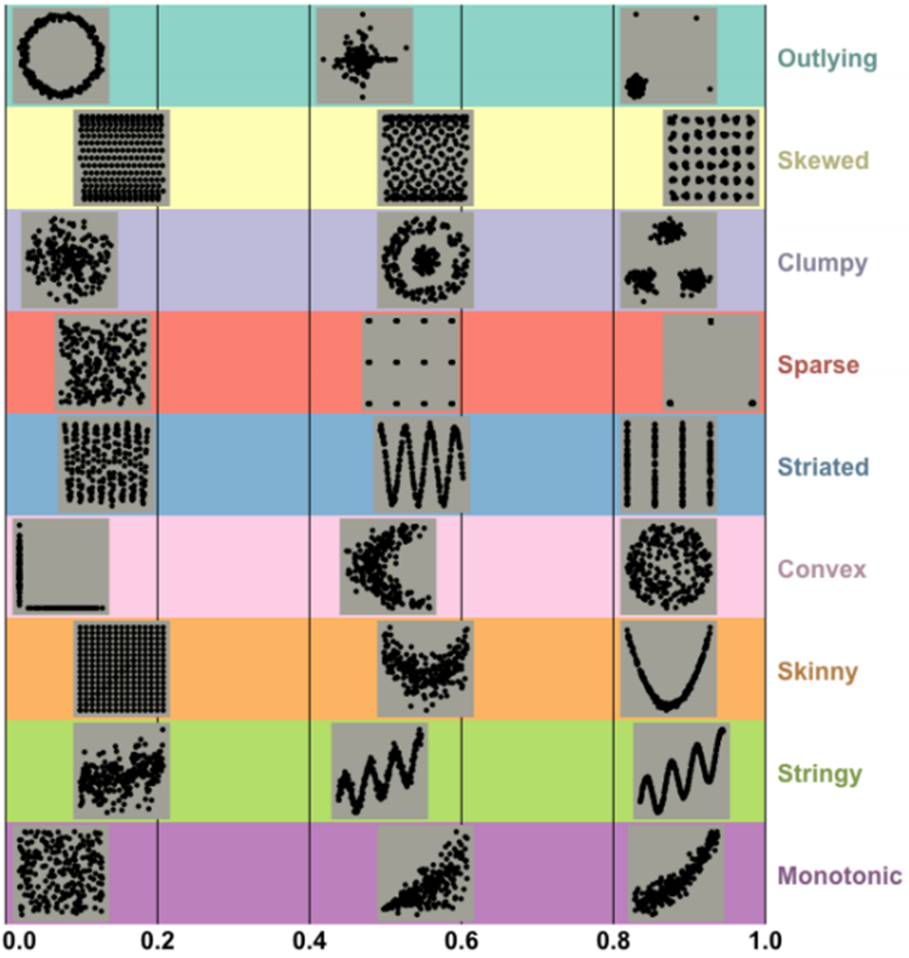

Figure 9: Examples of scatterplots and their scagnostics measures.

There are various packages in R that use the theory of scagnostics to visually characterize scatterplots: i.e. `scagnostics` and `scagExplore`. Both packages have been developed by common authors and the work is similar between both. The first one is the one used in this present thesis, since it allows the flexibility to imput different datasets for the psrameter computation. The scagExplorer has a much "fancier" visual display, but it does not easily allow for input data manipulation. Another interesting feature of 'scagExplorer' is the cluster or filtering analysis with the resulting scatterplots according to the various scagnostics measures obtained.

To fine-tune the scagnostics analysis, two artificial datasets were used, one with L-shaped data and another one with randomly distributed bivariate data. The datasets contained 200 rows and 20 columns representing the genes and the samples respectively. The scagnostics result comparison between the TRUE L-shaped genes and the non L-shaped genes showed few coefficients that could be used to discriminate between datasets (**??**).

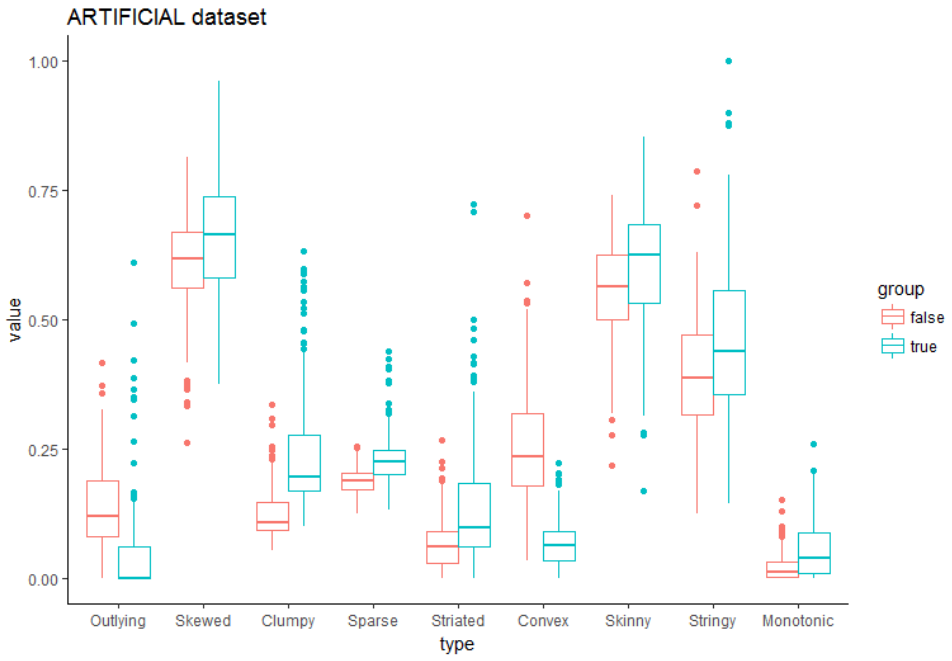

Figure 10: Boxplots representing the 9 scagnostics values for the classification of L-shaped and non L-shaped genes from artificially generated data.

The results revealed that the Outlying (< 0.06), Clumpy (> 0.17), Convex (< 0.09) and Sparse (> 0.20) could be used for classifying between the 2 groups of genes. After a fine-tuning exercise with the artificial dataset, the same exercise was carried out with a selection of TRUE/FALSE genes from the experimental, the GEO and the TCGA datasets. The first step was to select the genes selected visually as with L-shape for the DA dataset. The expression and the methylation datasets were selected based on that 'TRUE L' list of genes. These same steps were followed with the other two datasets, the GEO and the TCGA. These two datasets, instead of visual inspection, the TRUE and FALSE genes were selected using the naive, heuristic and CMI methods fort the scagnostics tuning.

To check the validity of these TRUE genes used, a visual selection was also carried out for the GEO dataset and the results were the same in terms of parameters identified. Therefore, a visual inspection of the TCGA genes was not done.
The parameters that were selected from the experimental dataset were: Monotonic, Convex, Striated and Clumpy. The parameters that were selected from the TCGA dataset were: Monotonic, Convex, Skinny and Clumpy. For the GEO dataset, none the parameters seemed to be able to discriminate between TRUE and FALSE. That is also why a visual selection of L-shaped genes was carried out, however the results did not improve.

The boxplots help visually select which are the parameters that discriminate better between the "true" L-shaped genes and the one sthat are not. Complementing that, difference Q1(true)-Q1(false); or the opposite as Q1(false)-Q3(true) to identify the chosen parameters' thresholds.

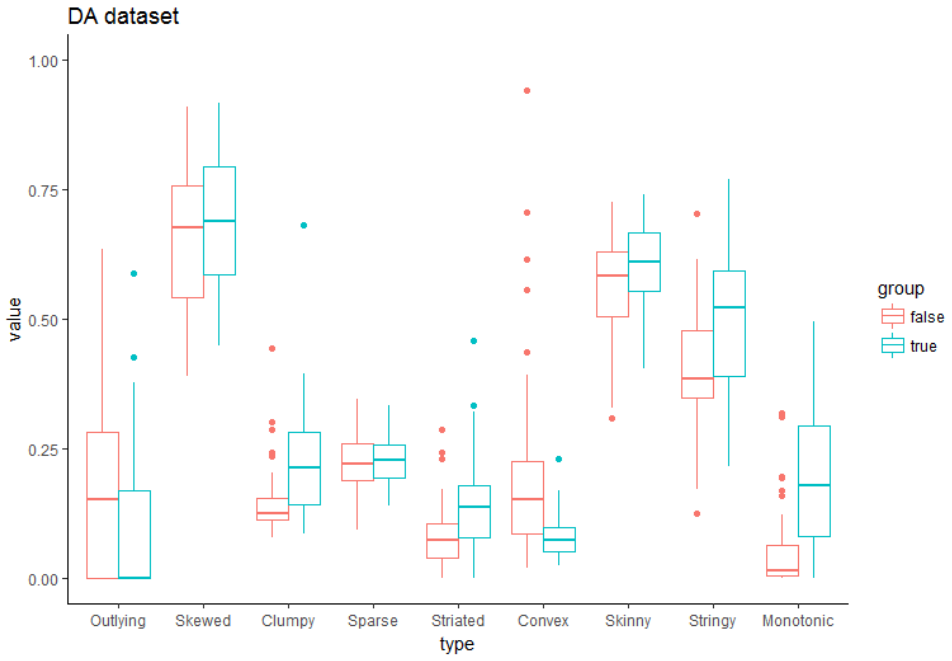

Figure 11: Boxplots representing the 9 scagnostics values for the classification of L-shaped and non L-shaped genes from the researchers' data.

We also merged results for all 3 datasets, and results for DA and TCGA datasets, to have more homogenous parameters thresholds (**??**.

The final parameters that we selected were: Outlying (< 0.06), Clumpy (> 0.17), Convex (< 0.09) and Sparse (> 0.20). However, for TCGA we have to do 2 adjsutments: 1) remove the striated (it could aslo be removed from the other selections) and 2) increase the ranges for Convex (< 0.50). The final number of genes obtained was 125 for the TCGA, 1870 for the GEO and 402 for the experimental dataset. Some of the scatterplots of the genes selected as true L-shaped were better representatives of a L-shape, but that the direction of the L is both following a positive and a negative correlation.

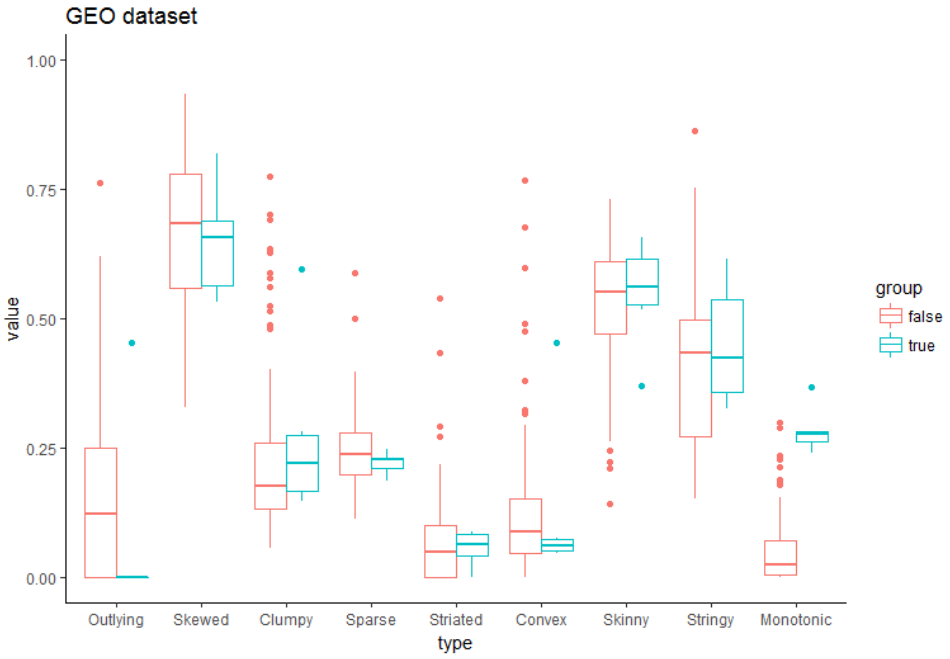

Figure 12: Boxplots representing the 9 scagnostics values for the classification of L-shaped and non L-shaped genes from the GEO data.

The combination of this method of selection with the naive in particular, and also the other 2 (CMI and heuristic) will improve the resulting list of selected L-shaped genes (when the expression is negatively correlated with methylation).

For the artificial dataset, the accuracy was of 92%, for sensitivity of 82%, and for specificity of 88%.

For the visually selected L genes data, he accuracy for these set parameters was of 59%, the sensitivity was of 18%, and the specificity was of 55%.

If we were to adjust the parameters visually to the particular dataset to improve the diagnostic measures with a more "real" dataset, it will be Monotonic > 0.09, Convex < 0.11, and Clumpy > 0.20.Then, the results would be as follows: The accuracy for these set parameters was of 67%, the sensitivity was of 34%, and the specificity was of 60%.

Still the sensitivity is very low for the visually selected L genes.

The **??** shows the number of genes identified with each dataset, with the original parameters:

TCGA	**GEO**	**researcher**
125	1870	402

Table 13: Scagnostics results with the TCGA, GEO, and researchers' datasets

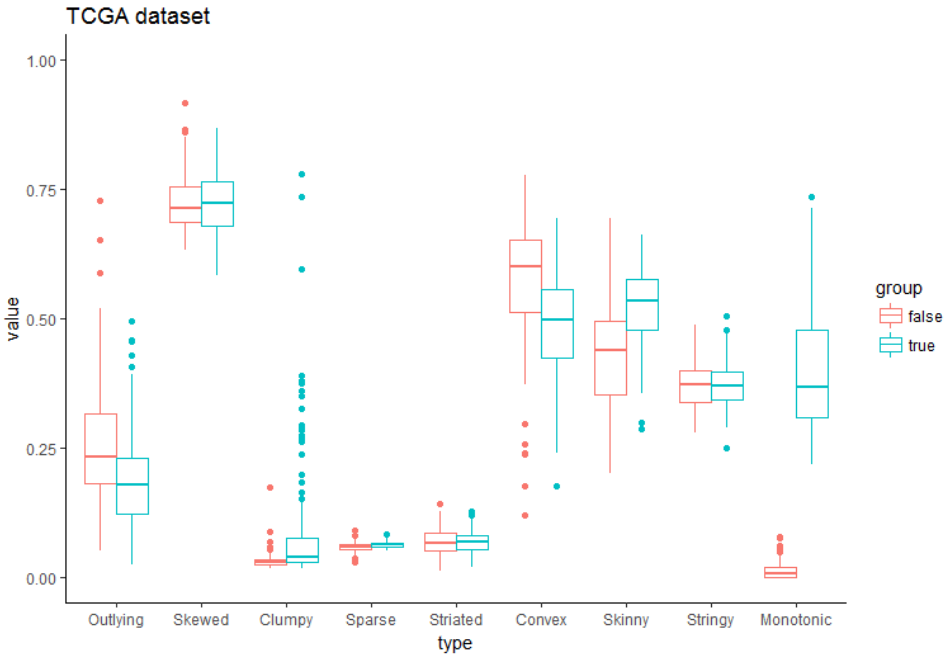

Figure 13: Boxplots representing the 9 scagnostics values for the classification of L-shaped and non L-shaped genes from the TCGA data.

The scagnostics method is set in two functions, one that runs the scagnostics function from the 'scagnostics' package for 2 large matrices; and the second one is a function that applies the selection criteia for L-shaped genes, which is basically based on a filtering or subsetting by multiple conditions. The parameters for the first function are the 2 matrices to input (methylation X, expression Y). For the second function, it is still being finalized, but the idea behind it is to have a parameter that will filter by the optimal L-shaped characteristics (Outlying (< 0.06), Clumpy (> 0.17), Convex (< 0.09) and Sparse (> 0.20)), but then to include an add-on customized option where the default values of the L-parameters can be changed or added (i.e. Convex > 0.1, Outlying < 0.23). Like that, the function could also classify for other patterns that do not adjust with the L-shape. In this way, it would be more flexible and potentially broadly used for other purposes, and not only for the L-shape classification which is of our interest.

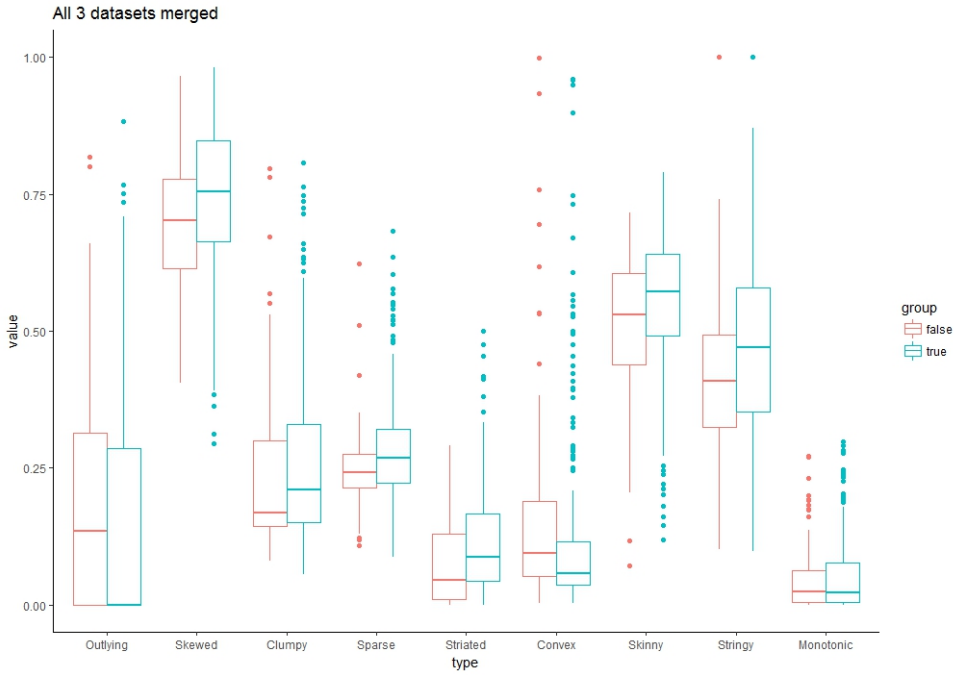

Figure 14: Boxplots representing the 9 scagnostics values for the classification of L-shaped and non L-shaped genes from all 3 experimental datasets merged.

3 Results

The 4 methods have been optimized to the best suited values, and 3 lists have been created for each method. These lists contain candidate genes which expression/methylation follows an L-shaped pattern; and therefore their expression is potentially being regulated by methylation.

3.1 Analysis of gene lists

One first exploration of these lists is by method: how many common genes have been identified with the naive method for all 3 datasets? And for the CMI, the heuristic and the scagnostics? In the Venn diagrams, the researcher's data is represented by "DA".

3.1.1 Analysis of gene lists by method

For the **naive method**, the relation of the genes identified between datasets shows that there are 40 genes in common between the TCGA and the researcher's dataset; whereas there is only one gene in common between the GEO and the researcher's dataset (**??**). It is also clear that there are no genes in common between the TCGA and GEO datasets identified by this method.

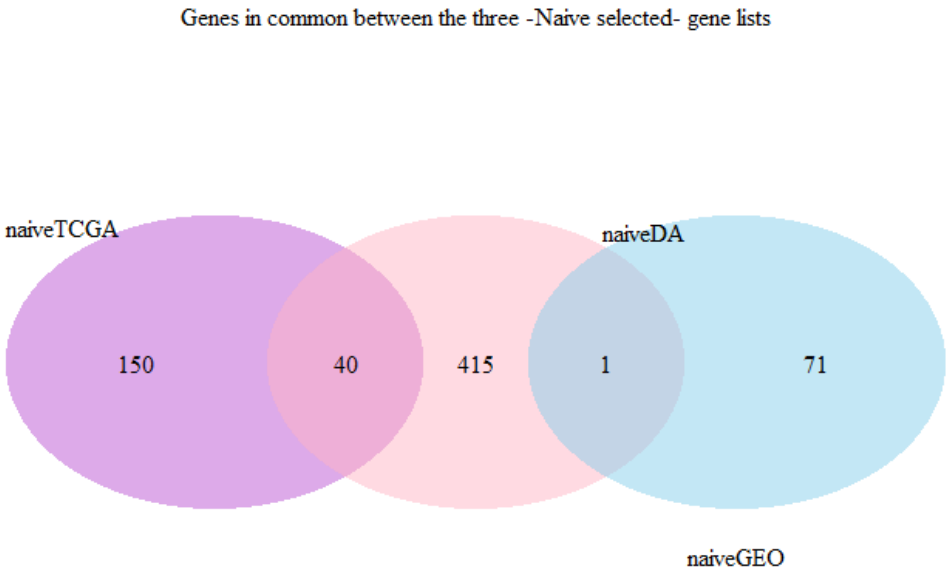

Figure 15: Venn diagram for the Naive gene selection in the 3 datasets

The number of genes selected using the **CMI method** is 301 for the TCGA, 263 for the GEO and 865 for the researcher's dataset. The common genes identified between datasets are represented in **??**. With this method, 4 genes have been commonly identified in all datasets. In addition to that, 96 more genes have been identified between the TCGA

and the researcher's datasets, 28 between the researcher's and the GEO datasets, and 7 between the TCGA and the GEO datasets.

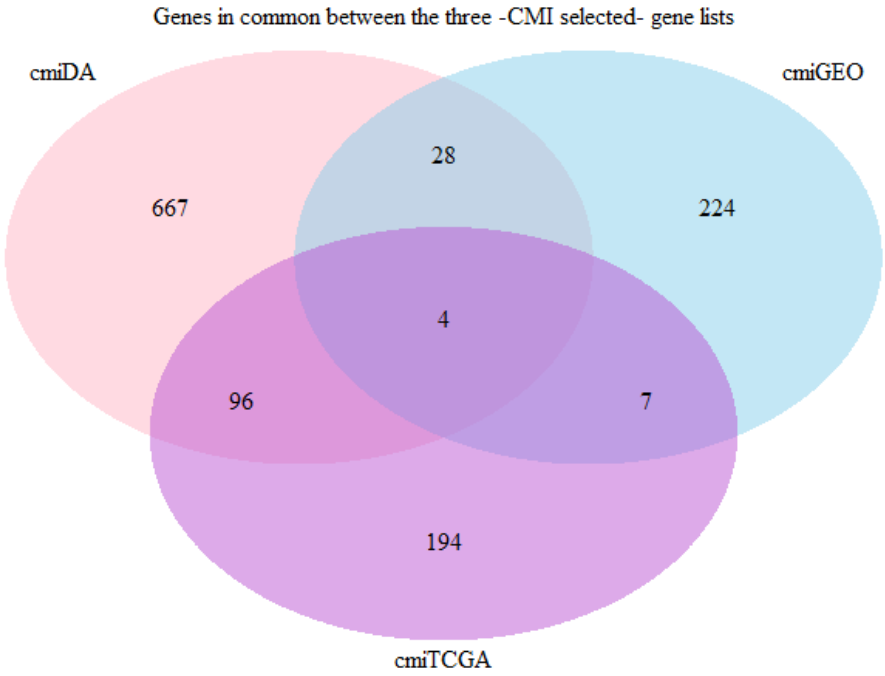

Figure 16: Venn diagram for the CMI gene selection in the 3 datasets

The number of genes selected with the **heuristic method** is 442 from the TCGA data, 39 from the GEO data and 188 from the researcher's data (**??**. There were 36 genes in common between the researcher's and the TCGA data, 1 between the researcher's and GEO data and 4 between GEO and TCGA data. No genes were selected from all datasets.

With the **scagnostics method** there were 2 genes in common between researcher' and TCGA and 1 between the researcher's and GEO datasets. With this analysis, the GEO and the TCGA had 32 genes commonly selected.

The genes common identified from the various methods is summarized in **??**

	Selection	DA	GEO	TCGA	ALL
1	All common genes	256	16	150	0
2	Interaction genes	59	0	104	

Table 14: The number of genes selected by each classifcations is represented in the following table:

One observation is that there is no common list compiled from a consensus of all methods and all datasets.

Figure 17: Venn diagram for the Heuristic gene selection in the 3 datasets

3.1.2 Analysis of gene lists by dataset

A second analysis of the list of genes obtained will focus on the gene selection of all methods in one set of data: how many common genes were found in the researcher's data by all methods? And in the GEO and TCGA data?

The analysis of the **researcher's dataset (DA)** yielded 3 genes identified by all 4 methods (**??**). There were 56 more genes that were identified by 3 out of 4 methods; and a further 232 commonly selected by 2 different methods.

The first 4 genes identified by the intersection of lists from various methods resulted in the scatterplot patterns visualized in **??** for the researcher's data.

Figure 18: Venn diagram for the Scagnostics gene selection in the 3 datasets

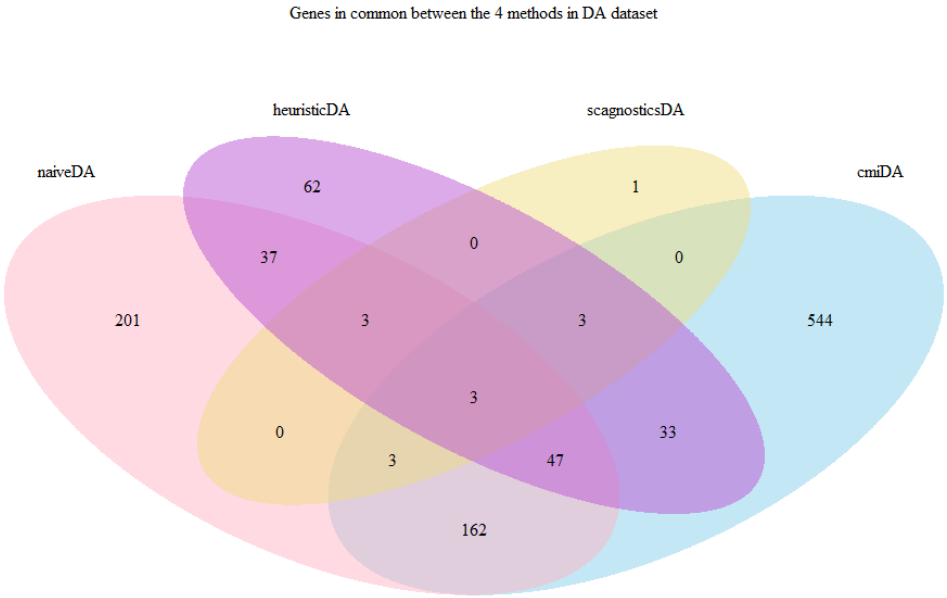

Figure 19: Venn diagram of the selected genes from the researcher's dataset with the 4 methods

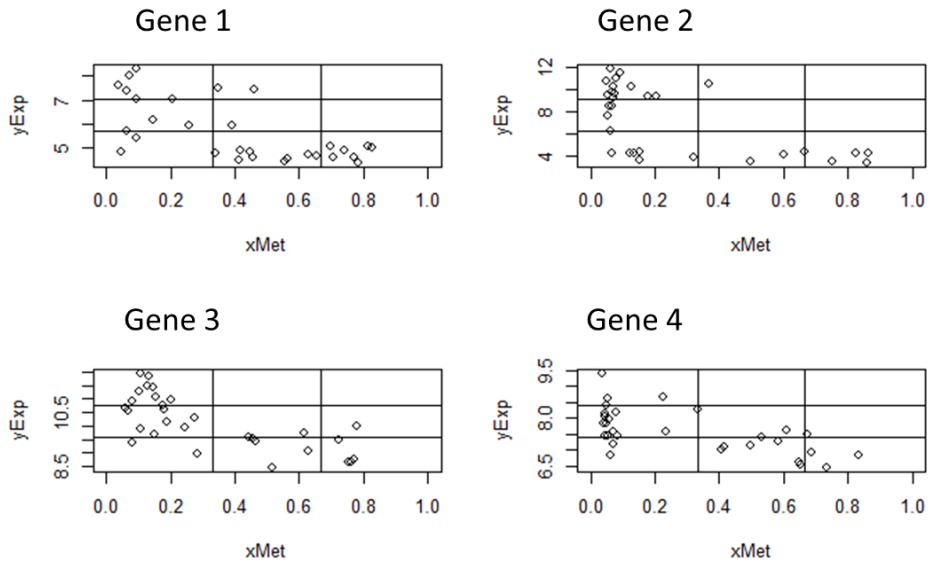

Figure 20: Sample scatterplots resulting from the genes selected from the researcher's data

The results for the **GEO dataset** are less consistant than the ones for the researcher's dataset (**??**).
There were no genes commonly identified by all methods from the GEO dataset, and only one gene was identified by 3 methods: the CMI, heuristic and scagnostics. There were 21 genes identified by a combination of 2 methods, and the rest of the genes were selected by one of the methods only.
The heuristic and naive combination did not identify any gene in common from this dataset.

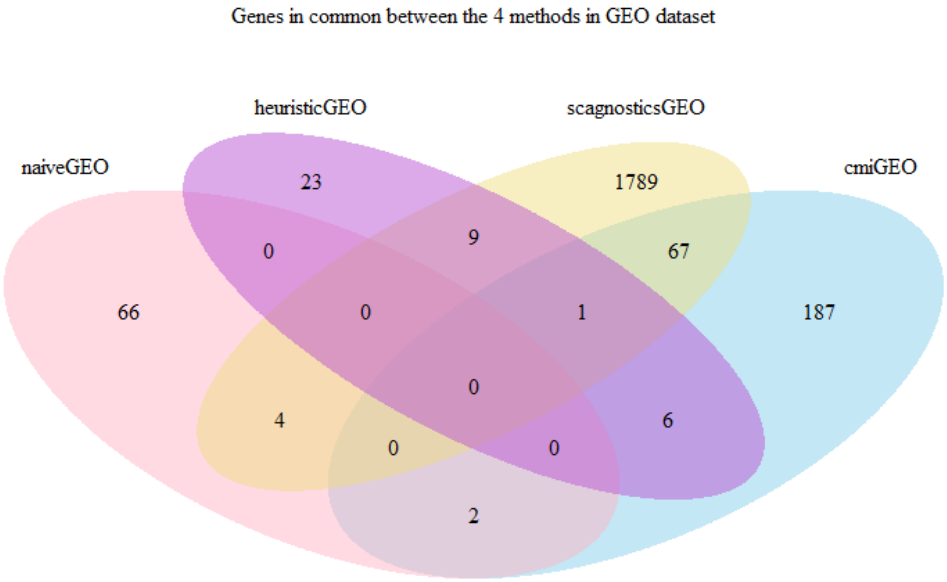

Figure 21: Venn diagram of the selected genes from the GEO dataset with the 4 methods

All four genes display a very neat L-shaped pattern.

The first 4 genes identified by the intersection of lists from various methods resulted in the scatterplot patterns visualized in **??** for the GEO data.

If we observe the scatterplots in **??**, the selection of L-shaped genes does not show as clear a pattern as with the researcher's dataset. One reason for that could be that the gene list used was the union from all methods, since the list resulting from the ntersection of all methods resulted empty for the GEO data.

There were 15 genes from the **TCGA dataset** identified by all for methods, 87 genes selected by 3 methods and 180 more genes identified by two methods.

The first 4 genes identified by the intersection of lists from various methods resulted in the scatterplot patterns visualized in **??** for the TCGA data.

The scatterplots in **??**, the selection of L-shaped genes show a very clear negative correlation between gene expression and methylation, however, the L-shape is less pronounced overall.

Four methods to identify genes potentially regulated by methylation were implemented, and the resulting lists yielded 60, 14 and 89 of selected genes for the researcher's,

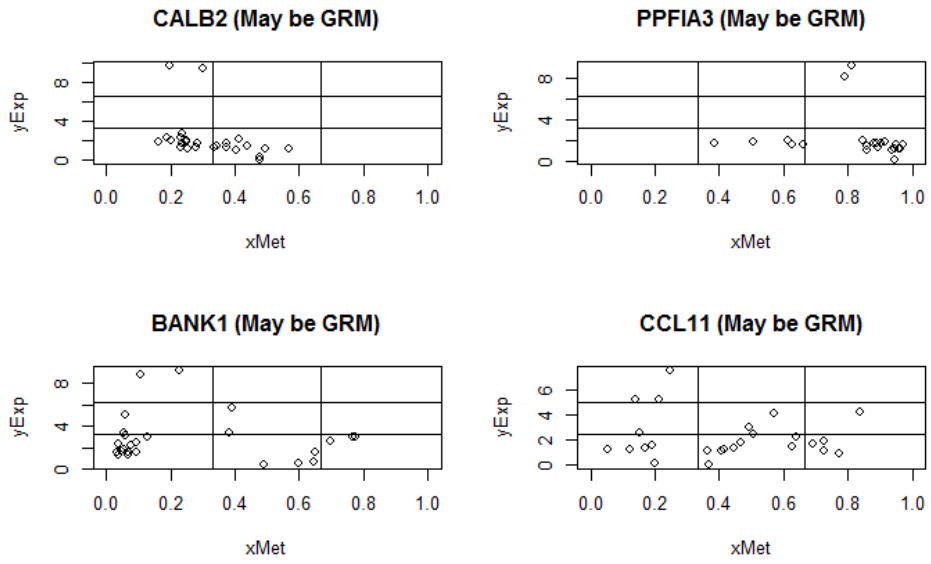

Figure 22: sample scatterplots resulting from the genes selected from the GEO data

the GEO and the TCGA datasets. The researcher's and the TCGA had 24 genes common if their selected list, the GEO dataset did not have any common genes with any of the other 2 datasets.

3.2 Brief biological significance analysis

A functional annotation analysis is a type of biological significance analysis that is based on the fact that genes that co-express ot "appear" together will participate in the same or a similar function. This is possible because of the strong association between co-expression and biological funtion. To carry out the analysis DAVID Bioinformatics resources was used ([**?**]). This is a web application that performs various funtional annotation analyses from a list of genes that the user inputs.

The list of genes obtained from each dataset were used as input into the DAVID application, and cluster analysis was performed individually. The gene-enrichment and functional annotation analysis generated a table of the clusters identified for each dataset (**??**, **??**,**??**). The scores for the significance level are a modified Fisher exact p-value.

Despite having only 24 genes in common, the functional annotation of the researcher's dataset and the TCGA dataset showed the same cluster results. In this first cluster (**??**, **??**) most genes are related to zinc finger, DNA binding, and regulation of transcription; all processes related to cander.
The GEO dataset produced different results with lower significance score, but it could be related to the low number of genes used for analysis.

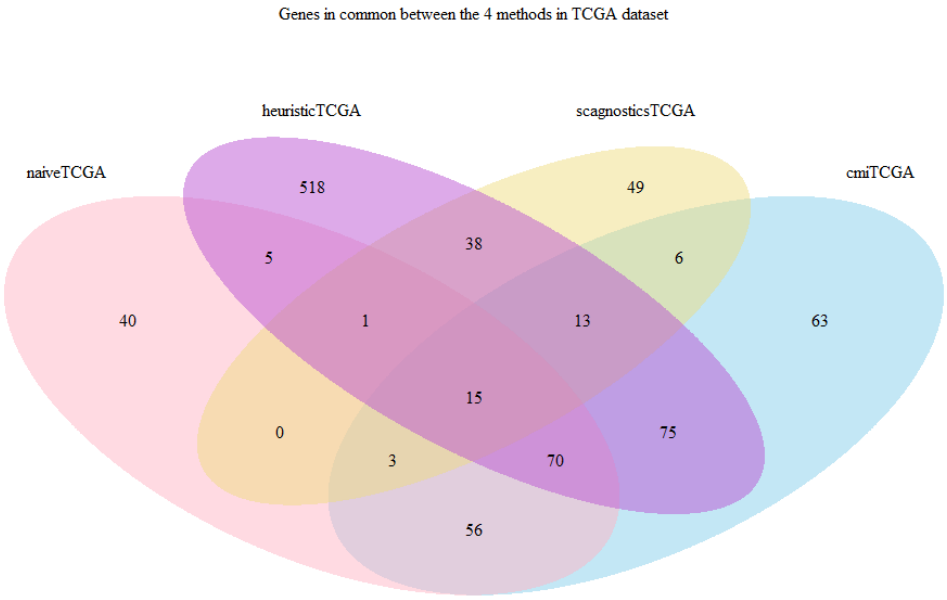

Figure 23: Venn diagram of the selected genes from the TCGA dataset with the 4 methods

3.3 Positioning of selected genes onto chromosomes and methylation pattern analysis

CpG islands are regions of the genome with higher Guanine (G) and Cytosine (C) percentage than other regions. These are usually under methylated, and are associated with the promoters of the genes, which is the region that controls expression. In the process of tumor generation, these low methylated areas (or hypomethylated), become highly methylated (or hypermethylated) ([**?**]). CpG island hypermethylation has been described in almost all cancers and many important cellular pathways are affected by the hypermethylation of these islands ([**?**]). therefore, the collocation of the candidate genes on these islands will reinforce the assumption that their expression is in fact regulated by methylation.

Another interesting feature related to methylation, expression and cancer is the DNAseI hypersensitive sites (DHSs) are chromatin regions sensitive to cleavage from DNAseI enzyme. These regions are associated with transcription, and are called regulatory regions. Methylated CpG that are found within a DHSs does not allow for the the association of the transcription factor to the DNA, by blocking the access to the chromatin. Moreover, collocation of genes with CpG islands and DNAse I hypersensitive sites provides a pattern for gene methylation and expression ([**?**]).

To test if genes that have been found to be regulated by methylation form this analysis are located at random in the genome or they collocate with known CpG islands and DNAse I hypersensitive sites, a methylation analysis workflow from [**?**] was followed.

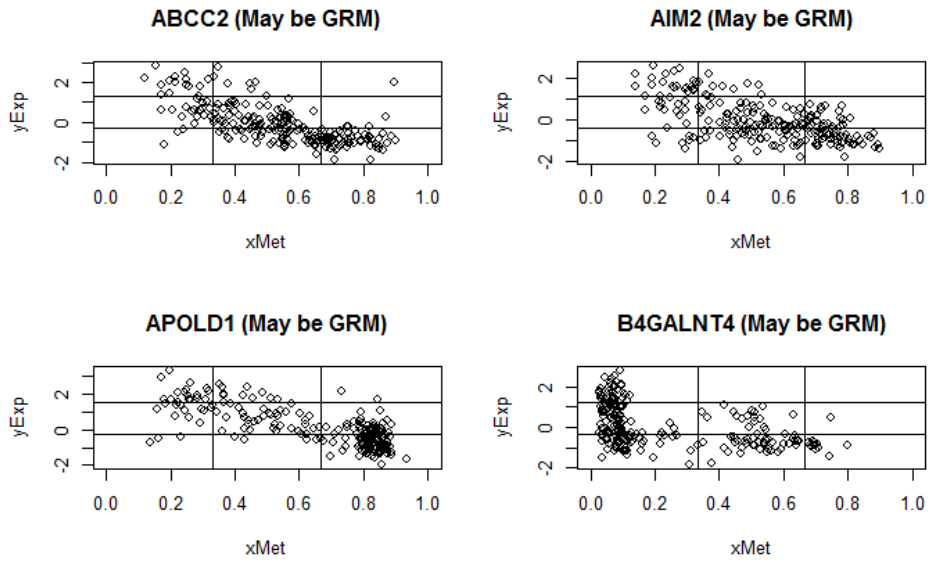

Figure 24: Sample scatterplots resulting from the genes selected from the TCGA data

3.3.1 Obtention of coordinates

The data analysis of the 3 datasets with the 4 methods (naive, cmi, heuristic and scagnostics) generated 12 lists of genes.

A dataset with the model CpG island mapped onto the hg19 genome can be obtained from the USCS (`http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/cpgIslandExt.txt.gz`). DHs can be downloaded from: `http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/`.

The first step is to obtain coordinates for the genes selected and the CpG islands and DHSs. Once the infomation is downloaded we store it in an R object of the type `GRanges`, which is a Bioconductor class to efficiently store information about sequences. Next we will repeat the precedure with the lists of selected genes. The function *getGeneLocations* is used to obtain the coordinates.

getGeneLocations is a transcript coordinate annotation tool. Given a vector with Gene Symbols, it will produce an object with ENTREZ ID, gene name, chromosome number, start position and end position for Homo sapiens. It has the following parameters:

1. **geneSymbolsSEL** is a vector containing Gene Symbols (unique abbreviation for the gene name) to be annotated.

2. **sortByChrom** is a logical indicating if the results have to be ssorted ascending by chromosome number. Default to TRUE.

3. **csvFileName** its is default is NULL; if a name is given, a csv file will be written as output.

Annotation Cluster 1	Enrichment Score: 2.07			Count	P_Value	Benjamini
UP_SEQ_FEATURE	zinc finger region:C2H2-type 8	RT		8	1.4E-4	3.3E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 11	RT		7	1.5E-4	1.8E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 7	RT		8	2.3E-4	1.8E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 10	RT		7	3.0E-4	1.8E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 6	RT		8	3.7E-4	1.8E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 9	RT		7	5.7E-4	2.3E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 5	RT		8	6.5E-4	2.2E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 14	RT		5	8.2E-4	2.5E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 4	RT		8	9.6E-4	2.6E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 3	RT		8	1.5E-3	3.6E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 13	RT		5	2.2E-3	4.7E-2
UP_KEYWORDS	Zinc	RT		15	2.4E-3	2.5E-1
INTERPRO	Zinc finger C2H2-type/integrase DNA-binding domain	RT		8	3.9E-3	4.4E-1
UP_SEQ_FEATURE	zinc finger region:C2H2-type 12	RT		5	5.0E-3	9.6E-2
UP_SEQ_FEATURE	zinc finger region:C2H2-type 15	RT		4	5.5E-3	9.8E-2
INTERPRO	Zinc finger, C2H2-like	RT		8	5.7E-3	3.4E-1
UP_KEYWORDS	DNA-binding	RT		13	6.0E-3	3.0E-1
UP_SEQ_FEATURE	zinc finger region:C2H2-type 2	RT		7	6.0E-3	9.9E-2
INTERPRO	Zinc finger, C2H2	RT		8	7.3E-3	3.0E-1
UP_KEYWORDS	Metal-binding	RT		18	9.7E-3	3.2E-1
SMART	ZnF_C2H2	RT		8	1.2E-2	3.9E-1
UP_SEQ_FEATURE	domain:KRAB	RT		5	1.4E-2	2.0E-1
UP_SEQ_FEATURE	zinc finger region:C2H2-type 1	RT		6	1.6E-2	2.2E-1
GOTERM_MF_DIRECT	nucleic acid binding	RT		8	2.5E-2	8.1E-1
INTERPRO	Krueppel-associated box	RT		5	2.8E-2	6.4E-1
GOTERM_MF_DIRECT	metal ion binding	RT		12	3.7E-2	8.0E-1

Figure 25: Cluster 1 resulting from the functional annotation analysis for the researcher's list of genes

3.3.2 Obtention of coordinates

Finally, the visualization of the genes onto each chromosome is carried out with `Gviz` with the function and *plotGenesInChroms*..

The *plotGenesInChroms* is a visualization function that to plot specific genes onto the corresponding chromosomes. Given a list of genes with their transcript coordenates, the function will plot the genes on their specific locations on each corresponding chromosme. It can also collocate the genes with CpG islands information and DNAseI hypersesitive sites.

1. **transcriptCoords** object of class containing a list of genes annotated with the getGenesLocations
2. **plotsFilename** object with the name of the file that will be used to save the .pdf with the graphs.
3. **minbase** number for the smallest basepair position of the chromosome. First position on the chromosome from the 5' side.
4. **maxbase** number for the largest basepair position of the chromosome. Last position on the chomosome from the 5' side.
5. **islandData** object of class GRanges containing CpG islands position data.
6. **dnaseData** object of class GRanges containing DNAseI hypersensitive sites position data.

Annotation Cluster 1	Enrichment Score: 0.45			Count	P_Value	Benjamini
UP_KEYWORDS	Glycoprotein	RT		7	4.7E-2	9.7E-1
UP_KEYWORDS	Signal	RT		5	2.6E-1	1.0E0
GOTERM_CC_DIRECT	extracellular region	RT		3	2.9E-1	1.0E0
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		5	2.9E-1	1.0E0
UP_KEYWORDS	Secreted	RT		3	3.6E-1	1.0E0
UP_SEQ_FEATURE	signal peptide	RT		4	3.7E-1	1.0E0
UP_SEQ_FEATURE	disulfide bond	RT		3	5.8E-1	1.0E0
UP_KEYWORDS	Disulfide bond	RT		3	6.6E-1	1.0E0
UP_KEYWORDS	Polymorphism	RT		8	7.4E-1	1.0E0
UP_SEQ_FEATURE	sequence variant	RT		7	9.3E-1	1.0E0
Annotation Cluster 2	**Enrichment Score: 0.12**			**Count**	**P_Value**	**Benjamini**
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		5	2.9E-1	1.0E0
UP_SEQ_FEATURE	transmembrane region	RT		3	8.8E-1	1.0E0
GOTERM_CC_DIRECT	integral component of membrane	RT		3	8.9E-1	1.0E0
UP_KEYWORDS	Transmembrane helix	RT		3	9.1E-1	1.0E0
UP_KEYWORDS	Transmembrane	RT		3	9.1E-1	1.0E0
UP_KEYWORDS	Membrane	RT		3	9.8E-1	1.0E0

Figure 26: Clusters resulting from the functional annotation analysis for the GEO list of genes

Figure **??** shows the genes selected using the Naive method on the DA dataset in the first chromosome only.

	Annotation Cluster 1	Enrichment Score: 2.77			Count	P_Value	Benjamini
	UP_KEYWORDS	DNA-binding	RT		24	1.2E-5	1.9E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 8	RT		11	1.8E-5	6.0E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 7	RT		11	3.5E-5	5.9E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 4	RT		12	5.2E-5	5.8E-3
	INTERPRO	Krueppel-associated box	RT		10	6.3E-5	1.2E-2
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 6	RT		11	6.8E-5	5.8E-3
	INTERPRO	Zinc finger, C2H2-like	RT		13	1.1E-4	1.1E-2
	SMART	KRAB	RT		10	1.1E-4	6.3E-3
	UP_SEQ_FEATURE	domain:KRAB	RT		9	1.3E-4	8.9E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 5	RT		11	1.5E-4	8.2E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 1	RT		11	1.6E-4	7.5E-3
	INTERPRO	Zinc finger, C2H2	RT		13	1.7E-4	1.1E-2
	INTERPRO	Zinc finger C2H2-type/integrase DNA-binding domain	RT		12	2.6E-4	1.3E-2
	SMART	ZnF_C2H2	RT		13	2.7E-4	7.4E-3
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 11	RT		8	3.6E-4	1.5E-2
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 10	RT		8	7.5E-4	2.8E-2
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 2	RT		10	1.5E-3	4.9E-2
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 9	RT		8	1.5E-3	4.6E-2
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 3	RT		10	1.9E-3	5.2E-2
	UP_KEYWORDS	Transcription	RT		21	2.4E-3	1.2E-1
	UP_KEYWORDS	Transcription regulation	RT		20	4.1E-3	1.5E-1
	UP_KEYWORDS	Zinc	RT		20	4.4E-3	1.3E-1
	GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	RT		12	5.0E-3	6.2E-1
	UP_SEQ_FEATURE	zinc finger region:C2H2-type 12	RT		6	5.4E-3	1.3E-1
	GOTERM_BP_DIRECT	transcription, DNA-templated	RT		18	5.7E-3	6.7E-1
	GOTERM_MF_DIRECT	nucleic acid binding	RT		12	6.0E-3	4.4E-1
	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		15	7.0E-3	6.4E-1

Figure 27: Clusters resulting from the functional annotation analysis for the TCGA list of genes

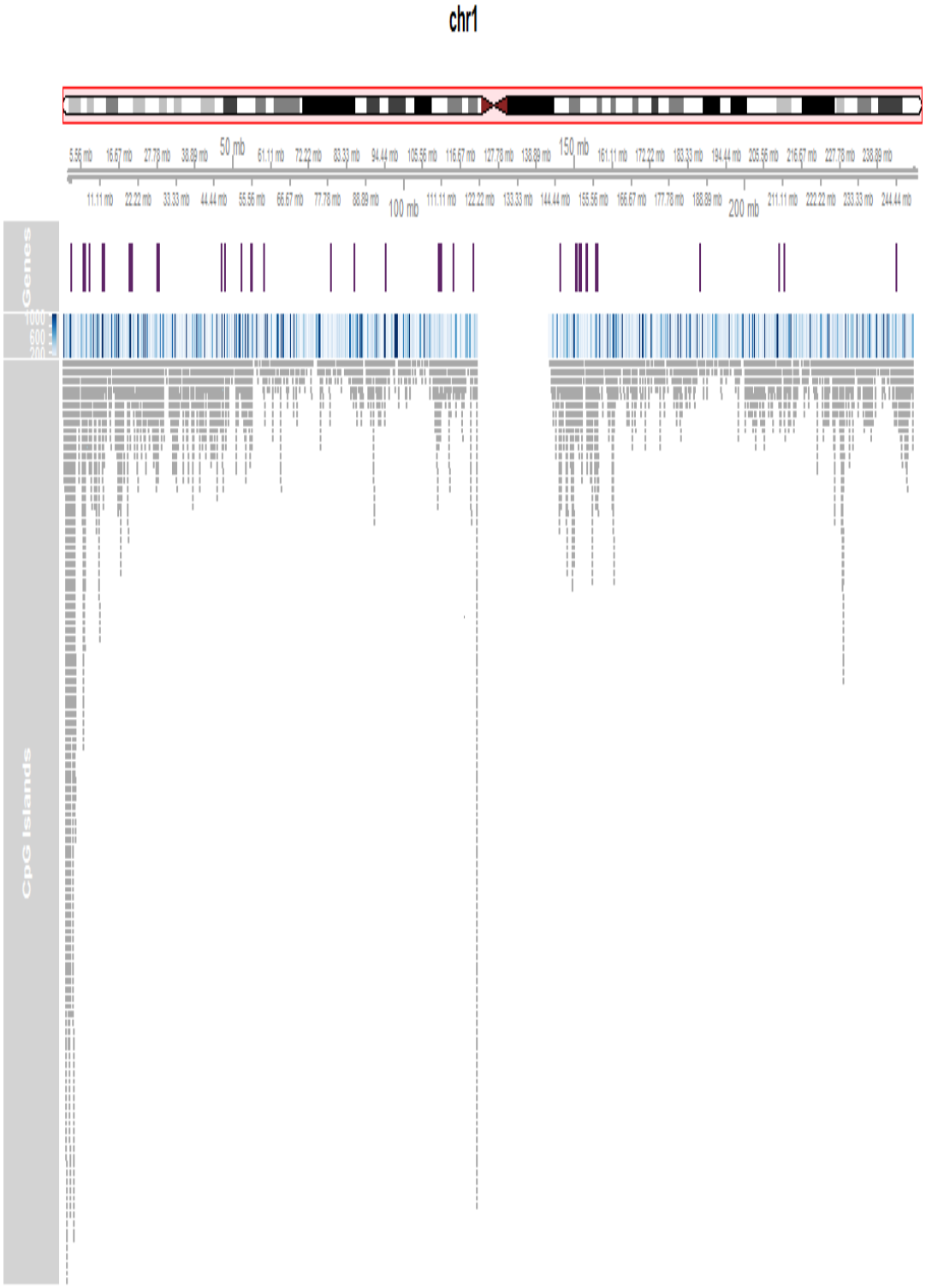

Figure 28: Plot of chromosome 1 with tracks for the genes selected from all 3 datasets, CpG islands and DHSs

3.3.3 Overlap between CpG islands and selected genes

In addition to drawing genes and CpG regions together, the overlap between gene regions and CpG islands provides a numeric value to the visual display. This can be done by calling "CpG islands" genomic windows whose GC content $> 50\%$ and observed-to-expected CG ratio > 0.6.

The overlap can be computed based on two methods: "standard" and "within" defined in the documentation of Bioconductor `GRanges` package.
For example, for the first chromosome and the gene list obtained by the Naive method on DA dataset we have:

The overlaps between the model CpG islands identified on the human genome and the genes identified from our query dataset (Naive-DA) were selected by 2 different methods with the following results: For a non-defined overlap, there were 560 overlaps between the model CpG islands and the genes selected from the DA data; and for a within feature overlap, there was 1 overlap found.

3.3.4 Evaluation of the methylation rate of the genome

Once the selected genes with an L-shaped pattern are distributed onto the chromosomes we want to test whether these genes fit a Poisson Process. The parameters for the Poisson process are λ and μ. The human genome has 22 + 1 chromosomes, and it is estimated the total number of coding genes is between 20,000 to 25,000; however, this number is diferently distributed in each chromosome. We can calculate the Poisson Process per chromosome, considering only the protein coding genes.

For this analysis, it is considered

$$\lambda = methylatedgenes/chromosome \tag{1}$$

and

$$\mu = \lambda chromosome \tag{2}$$

However, there is not data for λ, which is the rate of methylated genes in a chromosome (or the genome).

	Chromosome	Mb	ProteinCodingGenes	percentMet
1	chr1	249	2058	17
2	chr2	242	1309	11
3	chr3	198	1078	18
4	chr4	190	752	25
5	chr5	182	876	6
6	chr6	171	1048	10
7	chr7	159	989	10
8	chr8	145	677	18
9	chr9	138	786	23
10	chr10	134	733	8
11	chr11	135	1298	17
12	chr12	133	1034	22
13	chr13	114	327	33
14	chr14	107	830	5
15	chr15	102	613	11
16	chr16	90	873	25
17	chr17	83	1197	10
18	chr18	80	270	59
19	chr19	59	1472	15
20	chr20	64	544	27
21	chr21	47	234	47
22	chr22	51	488	27
23	chrX	156	842	24
24	chrY	57	71	3

Table 15: Percentage of methylated genes per chromosome

4 Development of an R package

To reuse the code and pack it in and ordered way, all functions used in the analysis of L-shaped genes have been compiled in an R package. The package is called *lpattern* and its creation had already been started before. Within this project, the fucntions parameters have been properly documented and tested.

To create a package in R we need to have installed Rtools (for Windows) in our computer or an equivalent for the appropiate OS used. LaTeX installation is also a prerequisite. There is the need to install 3 packages (if not previously done so), that will be required for the package development:

- devtools
- roxygen2
- testthat

There are 2 basic functions that are used to test if a package is working:

- *document()* created the documentation file
- *check()* checks for errors, warnings and notes of any kinds that may affect the creation or use of the package

The *lpattern* package has the following structure:

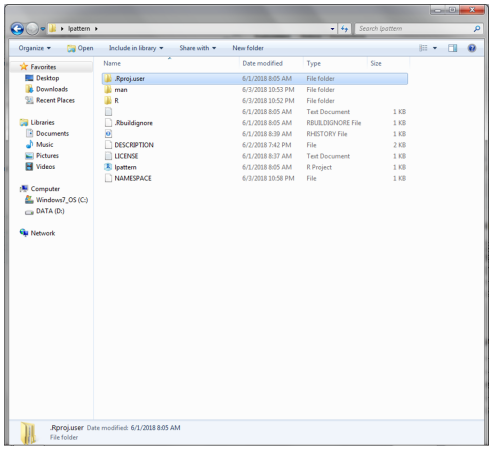

Figure 29: Structure of the R package *lpattern*

Finally, the following figure shows an example of a function created in .R with roxygen2 template for writing an .Rd file (**??**).

Next developments for the package are the inclusion of the scagnostics set of functions, that are still under preparation and the polishing of some of the functions that are currently in the package, but that generate some warnings.

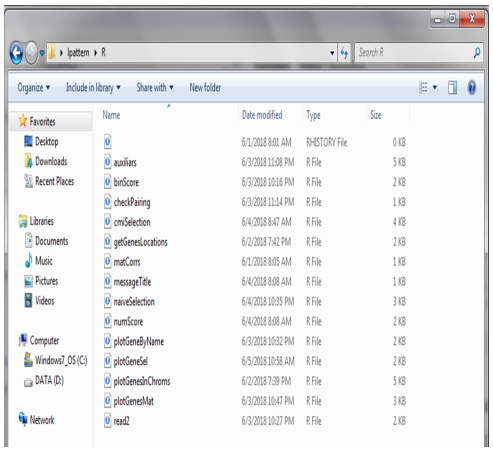

Figure 30: List of functions of the R package *lpattern*

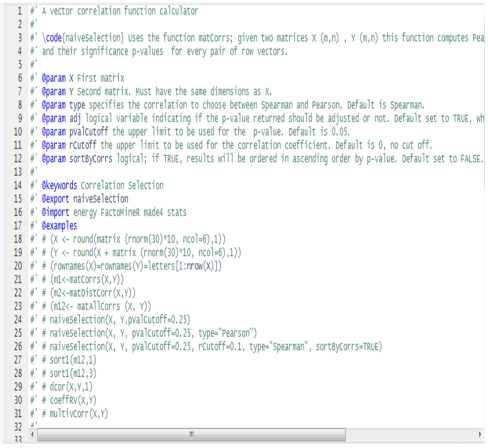

Figure 31: Example of a documented function of the R package *lpattern*

5 Discussion, conclusions and future work

This work has evaluated 4 different methodologies for selecting candidate genes for colorectal cancer (CRC) which expression and methylation plot follows an L-shaped pattern. To acomplish that, 3 different datasets from experimental datahave been used, and one artificial dataset has been generated for parameter tuning.
To evaluate the various methods, diagnostic measures have been calculated for different parameters and tested ewith different datasets. In addition to that, the methods have also been evaluated for their biological significance.

5.1 Discussion

The method evaluation base don the accuracy, sensitivity and specificity produced a series of results with the 2 datasets used (**??**)

Method	Accuracy	Sensitivity	Specificity
Naive	100	100	100
CMI	75	41	69
Heuristic	100	100	100
Scagnostics	92	82	88

Table 16: Diagnostic measures obtained with the artificial dataset

Another different set of diagnostic measures were obtained with the visually selected L-datset (**??**).

Method	Accuracy	Sensitivity	Specificity
Naive	72	44	64
CMI	84	68	76
Heuristic	95	91	90
Scagnostics	59	18	55

Table 17: Diagnostics measures obtained with the TCGA, GEO, and researchers' datasets

The measures of diagnostic show that the ideal method would be the **Heuristic method** and the least recommended the **scagnostics method**. Looking at these measures, the heuristic method could be used by itself, without the conjunction of the other methodologies. A test that would need to be used in conjunction with other methods is the scagnostics. This methodology also has the setback that, despite selecting for a particular scatterplot shape, it cannot select for positive or negative correlations and therefore, some of the selected henes have an inverse L-shape (**??**).

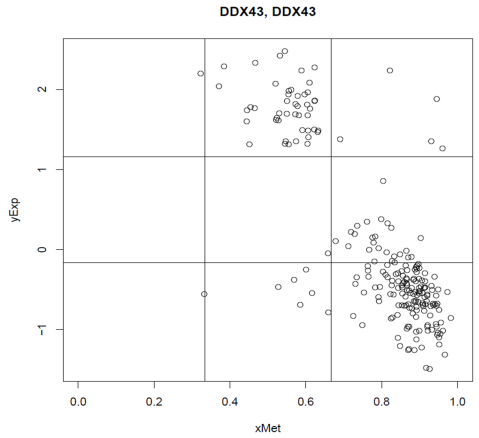

Figure 32: Example of a scatterplot of a gene selected with the scagnostics package that shows an inverse L-shape.

Most methods failed to detect true positives (as the sensitivity is the lowest measure). On of the reasons for that could be the lack of real negative data. This refers to the fact that the analysis is based on L-shaped, which is well defined, and non L-shaped, which actually includes all the ones that do not belong to the first category.

The methods have also been compared by looking at the final lists of selected genes. From these lists, the common number of selected genes identified have been contrasted both by method and by type of dataset.
The final lists of genes obtained from the intersection of all datasets have been compared between all 4 methods. This comparison shows that there are no genes commonly identified by all methods from all datasets, and that only 4 genes have been identified by 3 out of for methods. One gene between scagnostics, heuristic and naive and 3 between CMI, heuristic and naive (venn9). These results could be related to the different outputs obtained from each dataset. When comparing the results between methods for a particular gene list (as in Venn diagrams venn6, venn7, venn8), the researcher's dataset has 3 genes in common selected from all methods, and the TCGA dataset had 15 genes in common between all methods. If we remember that the initial overlap between lists was of about 8-9000 genes, these final numbers may seem too low
These results reflect the fact that each method selects a different set of genes, which also means that each method has different strengths and witnessess.

The gene lists have also been compared after the functional annotation analysis (dav1, dav2, dav3). Both the TCGA and the researcher' list have shown the same functional

Figure 33: Venn diagram representing common genes between all datasets and all methods.

annotation clustering. This implies that despite selecting for different genes from each list (meaning) that the genes as such are not the same, the gene functions of the principal clusters are the same. Biologically, this could be explained by the fact that the methylation pattern of a particular gene may be unstable or variable from sample to sample and change the expression and methylation pattern. However, in global terms, there may be a group of genes with the same function that are more or less expressed/methylated and they can exchange between them.

That brings us to the question:

What is the best method or combination of methods to select genes that have an L-pattern (that are regulated by methylation)?

Probably the answer to that question is that it will depend on our particular dataset for analysis. However, as a general pipeline for analysis, the pooling of various candidate genes selected by at least any 2 methods is a good starting point. This argument may seem a bit arbitrary, however, if a gene has been selected by at least 2 methods with different strategies, it is a validation of one selection by a second independent method. In addition to that, the resulting number of genes selected by at least 2 methods is of good length for biological significance tests.

This work has some limitations:

There is no validated data on what are called "TRUE genes" or "FALSE genes" for adequate calculations of accuracy, specificity and sensitivity. In addition to that, the visual inspection and selection of L-shaped genes may not be incorrect, bu then it is also missing some confirmation to support the information on L-shaped vs non L-shaped genes.

Another limitation encountered is the base for this work. It is widely believed and

understood that gene expression and methylation levels follow a negative correlation, however this probably is not always the case.

A limitation related to the above mentioned is the fact that the gene expression and methylation relationship is not stable, that means it can vary from sample to sample, and not necessarily due to a change in conditions.

5.2 Conclusions

This work is a description of 4 different methods used for the selection of genes regulated by methylation in relation to CRC. It also performs a functional annotation of these genes and their colocations with other experssion and methylation markers such as CpG island and DHSs.

The functions have been compiled in an R package emphlpattern. This package has the flexibility to select for L-shaped genes as well as for scatterplots with different patterns by changing the various function parameters.
Finally, a researcher's dataset has been analyzed and a list of candidate genes has been presented both as a proof of concept and for use as candidates for further analysis in the search of biomarkers for CRC.
Aparently, some test datsets like the TCGA are more suited for that analysis than the GEO datasets tested. In addition to that, the positive selection of L-shaped genes would improve with experimentally proven data for genes regulated by methylation and related to cancer (not any specific one) to fine tune the parameters better,

5.3 Future work

Next steps will be the finalization of the R package emphlpattern, with the scagnostics functions which are still under construction, proof reading of the documentation and the creation of vignettes for the package and examples.

In addition to that, a Shiny application will be updated with the latest optimizations and methodologies.

6 Glosary

- **CRC**: colorectal cancer.Also called bowel cancer, is the development of abnormal growth cells or tumors in the large intestine (colon or rectum).
- **Gene expression**: is the process by which information of a gene is used in the synthesis of a functional gene product.
- **Gene**: is a subunit of DNA that has the information to realize a particular function.
- **DNA**:is a complex molecule that contains all of the information necessary to build an maintain an organism. It is composed of 4 bases: Cytosine, Guanine, Adenine and Timine.
- **Methylation**: is a process by which methyl groups are addd to the DNA molecule. DNA methylation typically acts to repress gene expression.
- **CpG island**: is a sequence of the DNA bases Cytosine (C) followed by a Guanine (G) highly repeated for 500 to 1500 bases.
- **DHSs**: DNAse I hypersensitive sites, are regions of DNA chromatin that are sensitive to cleaveage of the DNAse I enzyme. This regions are usually egulating gene expression.
- **Scagnostics**: scatterplot diagnostics that are used to characterize 2D plots.
- **R**: a free programming software environment for statistical computing and graphics.
- **TCGA**: The Cancer Genome Atlas, is a project to catalogue genetic mutations responsible for cancer and stores related highthroughput datasets.
- **GEO**: Gene Expression Obnibus, is an NCBI repository that contains various omics datsets.
- **Shiny**: R package to build interactive web apps based on R.

7 Bibliography

References

[1] Arango D, Bazzocco SD, Carton-Garcia H, Macaya F, Andretta I, Chionh E, Rodrigues F, Garrido P, Alazzouzi M, Nieto H, Sanchez R, Schwartz Jr A, Bilic S, Mariadason J, *Highly expressed genes in rapidly proliferating tumor cells as new targets for colorectal cancer treatment*, Clinical Cancer Research, 21: 3695-3704, 2015.

[2] Ashktorab H, Brim H, *DNA methylation and colorectal cancer*, Curr Colorectal Cancer Rep 10: 425-430, 2014.

[3] Ashktorab H, Daremipouran M, Goel A, Varma S, Leavitt R, Sun X, Brim H, *DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia* Epigenetics 9(4):503-12, 2014.

[4] Ashktorab H, Rahi H, Wansley D, Varma S, Shokrani B, Lee E, Daremipouran M, Laiyemo A, Goel A, Carethers JM, Brim H, *Toward a comprehensive and systematic methylome signature in colorectal cancers*, Epigenetics 8(8):807-15, 2013.

[5] Barat A, Ruskin HJ, Byrne AT, and Phren JHM, *Integrating colon cancer microarray data: associating locus-specific methylation groups to gene expression-based classifications*, Microarrays, 4: 630-646, 2015.

[6] DAVID functional annotation bioinformatics microarray analysis, url-https://david.ncifcrf.gov/, 2018.

[7] Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, et al., *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts*, Nat Genet 41:1350–1353, 2009.

[8] Sandoval J, Esteller M, *Cancer epigenomics: beyond genomics*, Curr Opin Genet Dev 22:50–55, 2012.

[9] GEO datasets, `https://www.ncbi.nlm.nih.gov/gds`, 2018.

[10] Gonzalo V, Lozano JJ, Muñoz J, Balaguer F, Pellise M, Rodríguez de Miguel C, Andreu M, Jover R, Llor X, Giraldez MD, Ocaña T, Serradesanferm A, Alonso-Espinaco V, Jimeno M, Cuatrecasas M, Sendino O, Castellvi-Bel S, Castells A, Aberrant gene promoter methylation associated with sporadic multiple colorectal cancer. *Gastrointestinal Oncology Group of the Spanish Gastroenterological Association*, PLoS One 5(1):e8777, 2010.

[11] Gopalakrishnan S, Emburgh BOV, Robertson KD, *DNA methylation in development and human disease*, Mutation research 647(1-2):30-38, 2008.

[12] Illinworth RS, Bird AP, *CpG islands - 'a rough guide'*, FEBS letters 583, 1713-1720, 2009.

[13] Jin B, Tao Q, Peng J, et al, *DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function*, Hum Mol Genet 17(5):690-709, 2008.

[14] Jin B, Yao B, Li JL, et al, textitDNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer, Cancer Res 69(18):7412-21, 2009.

[15] Jin B, Li Y, Robertson KD, *DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy?*, Esteller M, ed Genes and Cancer 2(6):607-617, 2011.

[16] Kishore K, de Pretis S, Lister R, Morelli MJ, Bianchi V, Amati B, Ecker JR, Pelizzola M, *methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data*, BMC Bioinformatics 16:313, 2015.

[17] Koch A, De Meyer T, Jeschke J,Van Criekinge W, *MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data*, BMC Genomics 16:636, 2015.

[18] Lao VV, Grady WM, *Epigenetics and colorectal cancer* Nat Rev Gastroenterol Hepatol 8:686–700, 2011.

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

[20] Lou S, Lee H M, Qin H, Li JW, Gao Z, Liu X, et al., *Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation*, Genome Biol 15:408, 2014.

[21] Maksimovic A, Phipson B, Oshlack A, *A cross-package Bioconductor workflow for analysing methylation array data*, `http://www.bioconductor.org/packages/devel/workflows/vignettes/methylationArrayAnalysis/inst/doc/methylationArrayAnalysis.html`, 2018.

[22] Moghadam BT, Zamani N, Komorowski J, Grabherr M, *PiiL: visualization of DNA methylation and gene expression data in gene pathways*, BMC Genomics 18:571, 2017.

[23] Mosquera Orgueira A, *Hidden among the crowd: differential DNA methylation-expression correlations in cancer occur at important oncogenic pathways*, Front Genet `https://doi.org/10.3389/fgene.2015.00163`, 2015.

[24] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`, 2014.

[25] Robertson K, *DNA methylation and human disease*, Nature Reviews Genetics 6, 597–610, 2015.

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

[27] Sanchez-Pla A, Ruiz de Villa MC, Carmona F, Bazzocco S, Arango D, *Integrative Analysis to Select genes regulated by methylation in a cancer colon study*, in "Extended Abstracts Fall 2015", eds: Ainsbury Elizabeth A, Calle MLuz, Elisabeth, Einbeck Jochen Gomez Guadalupe and Puig Pere, Springer International Publishing, pg: 53-57, 2015.

[28] Shiny, `https://www.rstudio.com/products/shiny/`, 2018.

[29] Stimson KM, Vertino PM, *Methylation-mediated silencing of TMS1/ASC is accompanied by histone hypoacetylation and CpG island-localized changes in chromatin architecture*, JBC Manuscript M109809200, 2001.

[30] Tukey JW, Tukey PA, *Computer graphics and exploratory data analysis: An introduction*, In: National Computer Graphics Association (ed), Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85 III, Fairfax, VA, 1985.

[31] VanderKraats ND, Hiken JF, Decker KF, and Edwards JR, *Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes*, Nucleic Acids Research 41, 1-12, url-http://digitalcommons.wustl.edu/open_access_pubs/1651, 2013.

[32] Colorectal cancer statistics, `https://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/colorectal-cancer-statistics`, 2018.

[33] Wilkinson L, and Wills G, *Scagnostic distributions*, Journal of computational and graphical statistics, 17: 473-491, 2012.

[34] Zhang Y, Li Q, Chen H, *DNA methylation and histone modifications of Wnt genes by genistein during colon cancer development*, Carcinogenesis 34(8):1756-63, 2013.

8 Annexes

This work has been developed with R software. Within R the package *knitr* has been used to create the documentation, which is then written into TeX language.

The package *lpattern* is hosted in /lpattern.

Within the package there are the functions used, the package documentation and a *vignettes* folder with examples of application of the various functions.