A Statistical Framework for GeneSet Enrichement Analysis based on DNA Methylation and Gene Expression

BY

AMIRA KEFI
B.S., Higher Institue of Management, Tunisia 2005
M.S., Engineering School of Computer Sciences, Tunisia 2007

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:

Yang Dai, Chair and Advisor
Jie Liang
Chunyu Liu, Department of Psychiatry

To my parents, my family, my friends, all human beings and all creatures.

ACKNOWLEDGMENTS

I want to thank GOD to make me enough lucky to get a Fulbright scholarship to join this program and join Dr.Yang Dai's lab. Thank you Fulbright, for your financial support. Thank you Dr. Yang Dai, for your advising, patience and kindness. Thank you Dr. Jie Liang and Dr. Chunyu Liu for agreeing to be in my committee.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

TABLE **PAGE**

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

BEC	Blood Endothelial Cells
BC	Breast Cancer
CGS	Combined Gene Score
CpG	Cytocine preceeding Guanine
CSeS	CpG-Set Score
CSS	CpG Single-Score
DMR	Differentielly Methylated Regions
DNA	Deoxyribonucleic Acid
FDR	False Discovery Rate
GSEA	Gene Set Enrichement Analysis
GSS	Gene Single Score
LEC	Lymphatic Endothelial Cells
NES	Normalized Enrichment Score
PCA	Principal Component Analysis
RES	Running Enrichment Score
SNP	Single Nucleotide Polymorphism

SUMMARY

Gene expression profiling is considered as an approach to define, the phenotypes of many types of complex diseases such as tumors at the molecular level. DNA microarray technology provides biologists with the ability to measure the expression levels of thousands of genes in a single experiment and therefore enhance the transition from patterns detection of gene expression to pathways analysis.

Pathways representing molecular interactions between a set of genes are perturbed by the dis-regulation of the genes expression and can be used for potentiel targeted therapies. The dis-regulation of gene expression has been associated to many factors including single nucleotide polymorphisms and epigenetic alterations such as DNA methylation. Many studies have been conducted to integrate the gene expression and gene single nucleotide polymorphism and demonstrated better detection of pathways related to the phenotype. Recent literature has explored the association of gene expression and DNA methylation, but no study yet reported the combination of these factors using a gene set enrichment analysis to enhance the pathways analysis.

In the present work, we have developed a statistical framework to combine gene expression and DNA CpGs methylation and performed a gene set enrichment analysis to detect relevant pathways to the phenotype of interest. We adopted different scoring methods by first determining a score to a gene using only its gene expression data. Then we scored the gene according to its associated set of CpGs methylation status. Finally, we combined the two previous scores using different mathematical models to obtain a gene combined-score.

SUMMARY (Continued)

We used the proposed framework to analyse two datasets, breast cancer invasive carcinoma disease and the lymphatic and blood endothelial cells. Our approach detected abnormalities in previously identified phenotype associated pathways, such as Wnt and hedghog signaling pathways and DNA damage response. In addition, our statistical framework predicted novel pathways such as RNA degradation. These results demonstrate that our approach may help uncover biological pathways underlying human diseases and complex traits.

CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

DNA code alterations as SNPs, indels, translocations...whether inherited or not, are used for cancer diagnosis, prognosis and targeted treatment. These mutations can disrupt genes functions by enhancing the cancer genes called oncogenes or disrupting the tumor-supressor genes (1). In addition to genetic alterations, epigenetic alterations of the DNA can also be related to cancer development. The epigenetic profiles are also inherited and have been related to the silencing of cancer suppressor genes, by methylation of gene promoters where transcription of DNA to RNA begins (2).

In the last decades, gene expression dis-regulation has been highly correlated to cancer initiation and proliferation (3) (4) (5) (6). However it is still a debate whether also epigenetic modification alone or with association to gene expression trigger cancer progression and gene expression alterations (2). DNA methylation and gene expression are important processes in cell proliferation, differentiation and apoptosis and many disease traits (7) (8). Furthermore, changes in methylations can alter transcription and even be fatal in fetus development (7) (9) (10), which considers DNA methylation as a key regulator of gene expression.

The increasing evidence of high correlation between DNA methylation alterations and phenotype changes makes epigenetic as a new promising target for potential treatments. Since gene

expression has been used before for targeted therapies, could also DNA methylation be used as a complement to gene expression to strengthen the correlation between the phenotype and the molecular alterations? or could it be used independently? If so, how can we aggregate these two features? And to which extend their inherent relations can synergetically give better and more significant results?

1.2 Motivation and Significance

Gene expression regulation has been widely used in deciphering many phenotype variation traits and uncovering potential targeted gene therapy in complex diseases such as cancer. However, gene-set based analysis is more advantageous than single gene based analysis since it can uncover the interactions between related different genes within a pathway context related to the phenotype in study.

Many genomic and epigenomic features such as gene expression, gene copy number, SNPs, microRNA, DNA methylation, histone modification, can be combined together in analysis. The purpose of the combination is to increase the evidence and the power to elucidate the relationship between the phenotype in study and the alterations of these genomic features. In this work, we combine gene expression and DNA CpGs methylation data and propose a statistical framework to integrate both features to detect significant pathways related to the phenotype in question.

1.3 Thesis Organisation

We developed a statistical framework that integrates genes expression and DNA CpGs methylation to detect gene sets enriched for differential expression and/or DNA CpGs methy-

lation.

The thesis is organised as follows:

Chapter 2 reviews the relevant background knowledge. We briefly introduce concepts of gene expression, DNA methylation and their relation. We also introduce the Gene-Set Enrichment Analysis (GSEA) method and a summary of the available integrative analyses of genomic and epigenomic features.

Chapter 3 explains technical details on the proposed framework, and explains the Gene Single-Score, CpGs-Set Score, the Gene Combined-Score.

Chapter 4 presents the datasets we used to run our statistical framework and the different preprocessing steps of each dataset. We selected breast cancer and primary lymphatic and blood endothelial cells datasets.

Chapter 5 discusses results and literature validations. Our method identified significant gene-sets that are related to the selected datasets. These results indicate that the proposed method may help discover biological pathways related to human diseases and complex traits.

Chapter 6 concludes our work and presents our future perspectives

CHAPTER 2

BACKGROUND

This chapter describes the genomic features relevant to this work, which are the gene expression and DNA CpG methylation. We also give an overview of the Gene Set Enrichment Analysis (GSEA) method that was used in our statistical framework to detect relevant pathways. Finally, a summary of the integrative studies in the literature is provided.

2.1 Genomic Feature: Gene Expression

Gene expression profiling came into use as a new approach to define, at the molecular level, the phenotypes of many types of complex diseases, especially tumors. So far, many studies performed genome-wide expression patterns in several cancers including breast, lung, liver, ovarian. A common feature of these studies has been the emergence, using, for example, hierarchical clustering analysis, of different tumor subtypes based on distinct gene expression profiles patterns for each of these cancers.

The differences in gene expression patterns between cancer subtypes are likely to reflect differences in the cell biology of the tumors, at the molecular level. Based on these observations and results, one might consider these molecular subtypes as separable diseases (12). Therefore, DNA microarray technology provides biologists with the possibility to measure the expression levels of thousands of genes in a single experiment. Initially, experiments suggest

that genes of similar functions yield similar expression patterns in microarray hybridization experiments. However, the accumulation of such high-throughput data raised the need to use accurate methodologies for extracting biological significance and using the data to assign genes functions. To date, many approaches to the computational analysis of gene expression data attempt to learn functionally significant classifications of genes in a supervised or unsupervised fashion (11). Additionally, the transition from gene expression patterns detection to pathways analysis (13) make it also possible to study genes in a gene-set approach and leads to new medical insights and tumors targeted therapies (14).

2.2 Epigenomic Feature: DNA Methylation

The molecule of life, known as DNA sequence, is reproduced through the replication of four bases -adenine, guanine, cytosine, and thymine- which compose the alphabets of our primary sequence. However, there is also a "fifth" base that is a covalent modification in post replicative DNA, that is the methyl group added to cytosine (1) which happens mostly in cytosines preceding guanines, called also CpG site (2).

Mammalian genomes are highly methylated compared to simple organisms like yeast and drosophila. Most of this methylation is found in or arround CpG Islands regions densily crowded with CpG sites which occurs in almost half of genes promoter regions. Several studies observed that, in normal cells, CpG sites outside of the CpG island are mostly methylated, while CpG islands sites in genes promoter region are unmethylated. These findings have been explained as a possible suppression of unwanted transcriptions. However, in cancer cells, these normally methylated CpG sites become unmethylated and unmehtylated CpGs in promoter regions of

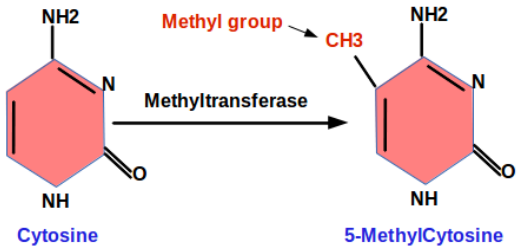


Figure 1. Methylation of Cytosine in the Mammalian Genome

other genes become methylated. These dis-regulated methylation have been related to transcriptional silencing of potentially tumor suppressor genes (2).

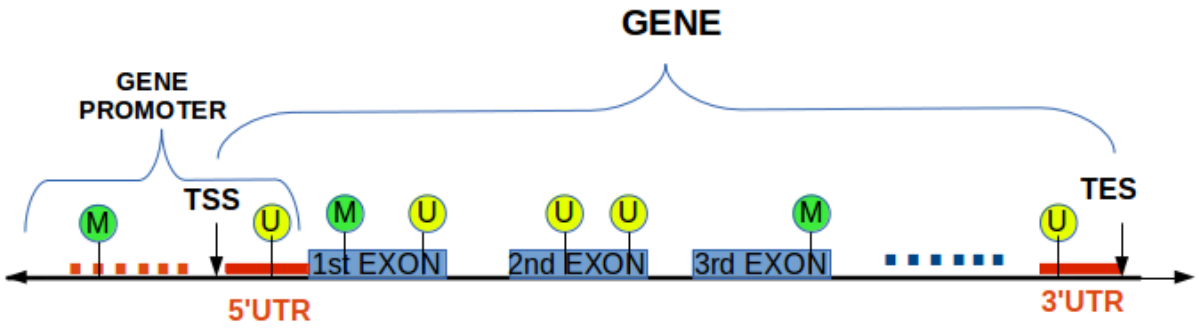


Figure 2. Gene methylation is manifested by the Mehtylation (M) or Unmethylation (U) state of every cytocine preceeding a guanine (CpG site)and can happen upstream or downstream the Transcription Starting Site (TSS) and goes until the Transcription Ending Site (TES).

2.3 Relation between Gene Expression and DNA Methylation

DNA methylation assures the silencing of genes in normal cells. Many studies have linked patterns of DNA methylation to gene expression and concluded that methylation in a gene promoter generally correlates with a silenced gene (2). However, methylation status of CpG sites in cancer cells have been correlated to losses and gains of methylation. In addition to mutations, methylation of CpG islands in gene promoter is associated with aberrant silencing of transcription and is considered as a mechanism for inactivation of tumor-suppressor genes (2). The loss of methylation was explained by a potentiel activation of normally silent regions of the genome leading to harmful expression of inserted viral genes such as imprinted genes on the inactive X chromosome. Recently, in addition to promoter region, new studies have found a high correlation between methylation of specific genomic regions, such as Exon1 (15) and gene Body (16) (17), and alterations in the gene expression.

2.4 Gene Set Enrichement Analysis: GSEA

We use the GSEA to validate our statistical framework through evaluation of the detected pathways and their numerical and biological significance. The numerical validation is well established by the GSEA and the biological one depends on the dataset selected and the literature validation.

Traditionally, different methods for gene expression analysis identify individual genes exhibiting differences between two phenotypes which does not detect biological processes that are distributed through a network of genes (31). In fact, the GSEA (31) considers all the genes in an experiment to generate a ranking list based on their associations to phenotypes. The GSEA

method aims to determine if genes belonging to a pathway, called also gene set, are ranked on the top (bottom) of the ranked list of genes (31). The GSEA uses a weighted Kolmogorov-Smirnov (K-S) test to determine which gene sets have statistical significance for association of the gene-set with the given phenotype.

2.4.1 Running Enrichment Score: RES

Genes can be ordered in a list L according to their differential expression g_i between phenotypes, which can be any test statistic suitable to assess the difference between the measurements of genes between phenotypes. Let us denote:

1. $g_i : i = 1,\ldots, N$ genes in the ranked list.
2. $s_j : j = 1, \ldots, M$ gene sets.
3. $\pi : 1,\ldots, \Pi$ permutations.

Given the rank list L and a gene-set s_j with H genes, a running enrichment score $\mathrm{RES}_{s_j}(i)$ at positions $i = 1, \ldots, N$ is computed as:

$$\mathrm{RES}_{s_j}(i) = \frac{1}{N_{sj}} \sum_{k=1}^{i} I(k \in s_j) - \frac{1}{N-H} \sum_{k=1}^{i} I(k \notin s_j), \tag{2.1}$$

where

$$N_{sj} = \sum_{k=1}^{N} I(k \in s_j), \tag{2.2}$$

The Equation 2.1 is as provided in the GSEA method, where $I(k \in s_j)$ is an indicator function which is one if the gene at the position k of the ranked list belongs to gene set s_j, otherwise 0.

2.4.2 Normalized Enrichment Score: NES

After computing the RES for each gene set at each position of the ranked list of our genes, these scores need to be normalized as they are dependent on the sizes of the gene sets. To do so, we define the best enrichment score $\mathrm{ES}(s_j)$ for gene-set s_j as follow:

$$\mathrm{ES}(s_j) = \begin{cases} \max_{i=1,\ldots,N} \mathrm{RES}_{s_j}(i) & \text{if } \left|\max_{i=1,\ldots,N} \mathrm{RES}_{s_j}(i)\right| > \left|\min_{i=1,\ldots,N} \mathrm{RES}_{s_j}(i)\right| \\ \min_{i=1,\ldots,N} \mathrm{RES}_{s_j}(i) & \text{otherwise.} \end{cases} \tag{2.3}$$

The score $\mathrm{ES}(s_j)$ is the maximum deviation of the $\mathrm{RES}_{sj}(i)$ from zero over all the positions $i = 1, \ldots, N$. The absolute magnitude of $\mathrm{ES}(s_j)$ indicates the strength of the association between the gene set and the phenotype. The sign indicates which phenotypic class the gene set is enriched with.

A normalized enrichment score (NES) for each gene set is calculated to adjust for difference in gene set size. The GSEA method uses a mean-based method and normalizes the positive and negative scores separately. Therefore, the normalized enrichment score as explained in the GSEA user guide respects this formula:

$$\mathrm{NES}(s_j) = \frac{\text{actual } \mathrm{ES}(s_j)}{\text{mean}(\mathrm{ES}(s_j)\text{against all permutations})} \tag{2.4}$$

The normalized enrichment score (NES) is the primary statistic for examining gene set enrichment results. NES is based on the gene set enrichment scores for all dataset permutations. Therefore, changing the permutation method, the number of permutations, or the size of the expression dataset affects the NES (18).

2.4.3 Permutation Test: Pvalue and FDR

The assessment of statistical significance of the gene set enrichment score and adjustment for multiple hypothesis testing are carried out on a phenotype-based permutation procedure (18). A nominal P-value is calculated relative to a null distribution that is generated by shuffling the phenotypic class labels and recalculating the gene set association scores Π times as shown in 3. The Pvalue is computed as follow when the $\mathrm{NES}(s_i, \pi_0)$ is positive :

$$P_{s_i} = \frac{\sum_{\pi=1}^{\Pi} \mathrm{I}(\mathrm{NES}(s_j, \pi) \geq \mathrm{NES}(s_i, \pi_0))}{\Pi} \tag{2.5}$$

where $\mathrm{I}(\cdot) is the indicator function and \Pi$ is the total number of permutations.

False discovery rate (FDR) control is used in multiple hypothesis testing in order to correct for multiple comparisons. FDR is generated based on the normalized gene set association scores to correct for multiple hypothesis testing and to control the proportion of false positives below a certain threshold.

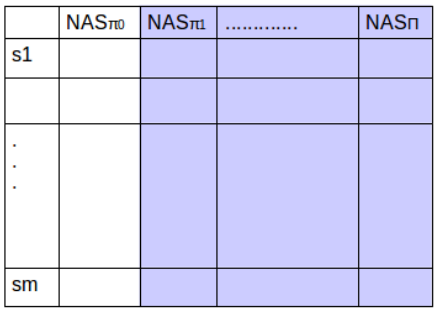


Figure 3. Permutation normalized scores. For every gene set s_j the enrichment scores RES, ES and NES are computed based on shuffled samples

Given M gene sets s_1,..., s_M and label permutations $\pi = 1,\ldots,\Pi$, and π_0, it represents the observed data, the FDR for each gene set s_j with $\text{NES}(s_j) \geq 0$ is computed according to the GSEA method as follow:

$$\text{FDR}_{s_i} = \frac{\%\,\text{NES}(s_j, \pi) \geq \text{NES}(s_i, \pi_0) \text{ for } j = 1, ..., M \text{ and } \pi = 1, ..., \Pi}{\%\,\text{NES}(s_j, \pi_0) \geq \text{NES}(s_i, \pi_0) \text{ for } j = 1, ..., M} \tag{2.6}$$

otherwise if $\text{NAS}_{si} < 0$:

$$\text{FDR}_{s_i} = \frac{\%\,\text{NES}(s_j, \pi) \leq \text{NES}(s_i, \pi_0) \text{ for } j = 1, .., M \text{ and } \pi = 1, .., \Pi}{\%\,\text{NES}(s_j, \pi_0) \leq \text{NES}(s_i, \pi_0) \text{ for } j = 1, .., M} \tag{2.7}$$

when the gene set normalised score is positive, the FDR_{s_j} is the ratio between, the proportion of the normalised gene sets scores $\text{NES}(s_j, \pi)$ that are bigger than the observed $\text{NES}(s_i, \pi_0)$ across all permutations Π and all gene sets M, and the proportion of normalised gene sets scores $\text{NES}(s_j, \pi_0)$ that are bigger than $\text{NES}(s_i, \pi_0)$ in the observed data across all gene sets M.

2.5 Integrative Analysis of Genomic and Epigenomic Data

To classify phenotypes, single genomic feature, such as gene expression, SNPs, microRNAs or epigenomic features such as DNA methylation, have been used successfully (41) (42) (43) (44). A daunting challenge is to explore the relationship between these different genomic and epigenomic features in order to combine them and stratify different disease subtypes where the use of a single feature fails. Several integrative analysis methods and tools have been proposed to integrate different genomic and epigenomic features by using different approaches and methods such as correlation or regression.

Xiong et al. developed a statistical framework to integrate genetic and gene expression into a genome-wide association analysis of gene sets, and demonstrated that this joint analysis improved the power to detect real associations compared to the use of only one genomic feature (27). Louhimo et al. proposed an algorithm to integrate data from gene copy number, DNA methylation and gene expression. Their study revealed a synergistic effect of DNA methylation and copy number changes on gene expression for several known oncogenes as well as novel candidates (28).

Nervertheless, Li et al. conducted an integrated analysis by clustering and correlating DNA methylation and gene expression and revealed associated pathways with the phenotype in study

(29). Sun et al. used a similar approach integrating gene expression, CpG island methylation, and gene copy number in breast cancer, revealing a global pattern of differential CpG island methylation that contributes to the transcriptome (30).

Most of these studies revealed interesting results by integration of genomic and epigenomic features, however, they focus on the intersection of genes with differential expression and genes with differentially methylated CpGs. This intersection only represents a small proportion of the available data that could be explored. Moreover, a lot of genes are differentially expressed and may not have differentially methylated CpGs while other genes may not be differentially expressed but have many differentially methylated CpGs. These cases are not considered by the previous studies and they may have an interpretation in the related studied phenotypes. The existing approaches do not provide a framework that can incorporate the complete information provided from gene expression and DNA methylation data.

More importantly, the differentially methylated CpG sites or position (DMP) are less meaningful when considered individually compared to being grouped by regions called differentially methylated regions (DMR) instead of differentially methylated CpGs. Finally, so far, none of the mentioned studies have integrated gene expression and DNA methylation in a statistical framework using a gene set enrichment analysis (GSEA).

In this study, we propose a new statistical framework to associate gene expression and DNA methylated gene regions. We suggest a scoring method that may reflect the nature of relation between these genomic and epigenomic features in order to detect the most relevant and significant pathways related to the studied phenotype.

CHAPTER 3

METHODS

In this chapter we present our statistical framework that integrates gene expression and DNA CpG methylation within a GSEA procedure to detect dis-regulated pathways in the phenotype of interest. First, the Gene Single-Score (GSS) and the CpG Single-Score (CSS) are explained. Then, since every gene could be related to a set of CpGs in different regions, we explain how we elucidate this problem by assigning a CpG-Set Score (CSeS) for every gene according to different regions. Consequently, we developed two mathematical models to obtain Gene Combined-Score (GCS) by aggregating the GSS and the CSeS.

3.1 Statistical Framework

We propose a new statistical framework to combine gene expression and CpG methylation that is illustrated in 4. First, from a gene expression profile, we compute a test statistic for individual genes between the phenotypes, as the score representing the degree of the differential gene expression. We call it Gene Single-Score (GSS). The same computation is performed to determine the CpG Single-Score (CSS) from CpG methylation profiles. Second, CpG-Set Score (CSeS) for each gene is produced based on the CpGs assigned to the respective gene and a selected formula. Initially, the CpGs are grouped in a set for each gene as defined by the methylation microarray platform annotation file. CSSs are aggregated using different formulas

and will be described later in the chapter. Next, the Gene Combined-Score (GCS) is the combination of the gene GSS and its correspond CSeS. Finally, using GSEA, we performed pathway analysis test to identify significant gene sets related to the studied phenotype.

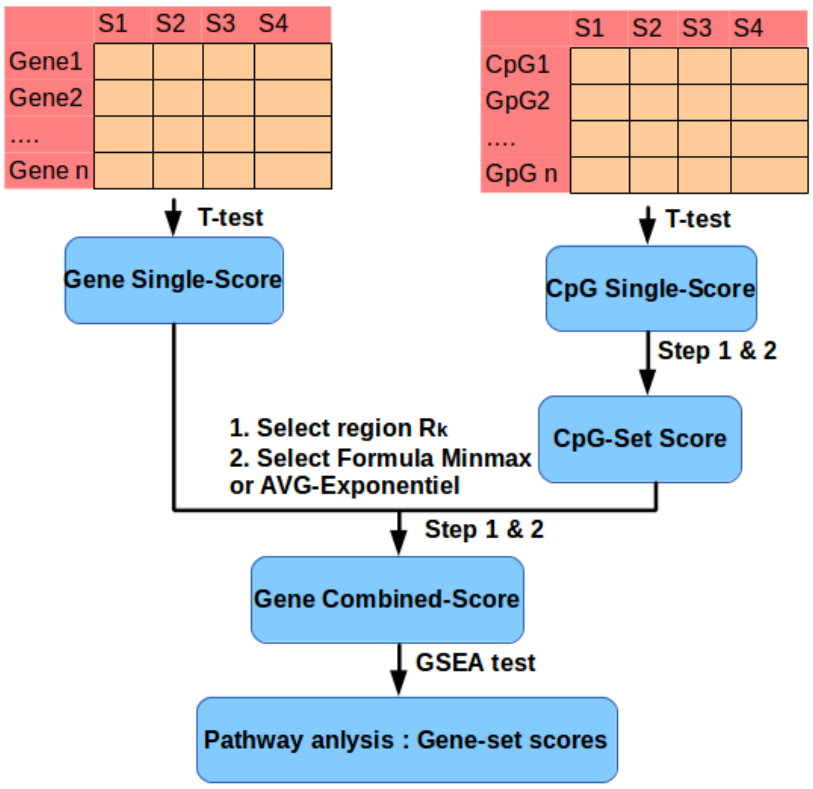


Figure 4. Proposed statistical framework

3.1.1 Gene Single-Score and CpG Single-Score: GSS and CSS

The GSS, $S_{exp}(g_i)$, is based on the gene test statistic only, and the same is true for the CSS, S_{cpg_j}. In our case we used the t-test statistic, which is computed between two phenotypes for individual genes and CpGs as follow:

$$S_{exp}(g_i) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.1}$$

where:

- $\bar{x}_k$ is the mean value of samples values from phenotype k, $k = 1, 2$.
- s_k is the standard deviation of samples values from phenotype k, $k = 1, 2$.
- n_k is number of samples in phenotype k, $k = 1, 2$.

3.1.2 CpG-Set Score: CSeS

In this step, we propose a novel approach to obtain a score for a gene according to its set of CpGs single scores. These CpGs are associated to the gene as reported by the annotation file of the microarray platform. We choose to study datasets having methylation experiments using Illumina Infinium Human Methylation450 BeadChip (illumina 450k) because it covers over 450,000 methylation sites per sample at single-nucleotide resolution (45).

Numerous studies have found that CpGs methylations in promoters around the transcription starting site (TSS) are correlated to gene expression alterations (46) (47). Nonetheless, recent studies observed that methylation of the first exon is also tightly linked to transcriptional silencing (15) and gene body methylation is positively correlated to gene expression (17) (16).

Therefore, it is more convenient to group CpGs according to regions to detect differences in methylation. For these reasons we clustered CpGs according to regions rather then using single CpG or all CpGs together. We grouped CpGs into four regions:

(a) TSS region: includes the CpGs in TSS200 and TSS1500 as defined according to the illumina 450k annotation file

(b) EXON1 region: includes the CpGs in 5'UTR and 1stEXON as defined according to the illumina 450k annotation file

(c) BODY region: includes the CpGs in body and 3'UTR as defined according to the illumina 450k annotation file

(d) Gene region: includes the TSS, EXON1 and BODY region all together and using all the CpGs in these regions to compute a CpG-set score for the gene.

Different methods can be used to score the CpG-set in a specific region for the gene association score used in the GSEA method. For example, if we select the TSS region, the max statistic, which is the maximum value of the test statistic of the CpGs, represents the most hyper-methylated CpGs associated to a gene. Likewise, the min statistic represents the most hypo-methylated CpG among the set of the CpGs related to the same gene in the TSS region. Hyper-methylation reflects the gain of methylation in the disease case relatively to the normal case and hypo-methylation is the opposite. In diseases such as cancer, the alteration of the methylome is not in one sense, some genes gain methylation and some others lose it. Surprisingly, some gene gain and lose at the same time and according to different regions. For

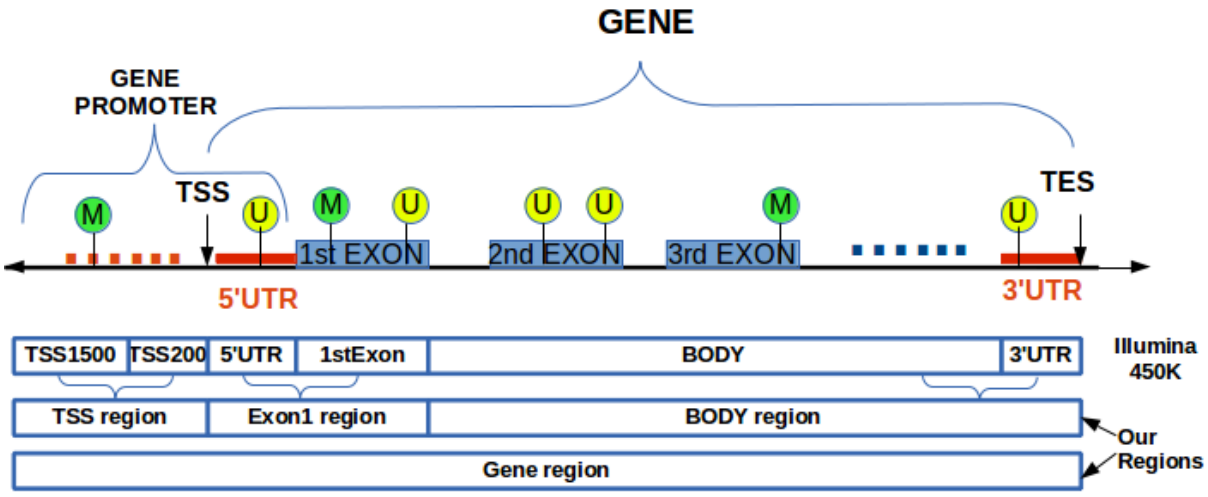


Figure 5. Gene Regions

example, over-expression of UHRF1 has been associated with epigenetic silencing of BRCA1 in sporadic breast cancer (20). In our study we found this gene observation valid, but we also found significant epigenetic activation of the same gene. We also used breast cancer dataset and found many other genes, gaining and losing methylation at the same time (see 6).

If we use the maximum statistic and select only the CpG with the highest statistic (hyper-methylated) then we ignore minimum which is also significant. Therefore, we propose the following formula to assess if the gain and loss of mehtylation together are involved in the phenotypes studied:

	A	B	C	D	E	F	G	H	I
1	**cpg**	**gene**	**chromosome**	**position**	**region**	**tstat**	**pval**	**genetstat**	**genepval**
2	cg00033371	UHRF1	19	4956786	BODY	-2,77543948	0,0105111404	11,08562689	6,33E-011
3	cg01396275	UHRF1	19	4961535	BODY	-1,5722028785	0,128995168		
4	cg01977762	UHRF1	19	4909193	TSS	-5,9354629554	3,99E-006		
5	cg04571196	UHRF1	19	4910051	EXON1	-1,2791001891	0,2130912237		
6	cg04682120	UHRF1	19	4930815	BODY	-0,0973305855	0,9232721392		
7	cg04693399	UHRF1	19	4909447	TSS	0,9044726416	0,3747349724		
8	cg06423533	UHRF1	19	4909366	TSS	0,7873660489	0,4387747031		
9	cg08422181	UHRF1	19	4909468	TSS	-0,5684921592	0,5749822067		
10	cg09317102	UHRF1	19	4909554	EXON1	-0,8703071349	0,3927544549		
11	cg09443401	UHRF1	19	4944473	BODY	**10,1231609249**	3,87E-010		
12	cg10158633	UHRF1	19	4930908	BODY	3,2697718827	0,0032422051		
13	cg10601761	UHRF1	19	4930780	BODY	3,6514547554	0,0012646724		
14	cg11888359	UHRF1	19	4936127	BODY	-1,4667814999	0,1554176444		
15	cg12159575	UHRF1	19	4910274	EXON1	-1,9270611222	0,0658909371		
16	cg17714703	UHRF1	19	4912221	BODY	-5,8842468303	4,53E-006		
17	cg18030386	UHRF1	19	4944160	BODY	8,6973567921	6,97E-009		
18	cg18322448	UHRF1	19	4910276	EXON1	-1,3367780304	0,1938297799		
19	cg18473042	UHRF1	19	4909381	TSS	0,2376166688	0,8141952206		
20	cg19230709	UHRF1	19	4933046	BODY	-2,0359856973	0,0529332613		
21	cg23290217	UHRF1	19	4909290	TSS	**-11,1983784051**	5,16E-011		
22	cg23933606	UHRF1	19	4944005	BODY	4,1137265841	0,0003951465		
23	cg25466989	UHRF1	19	4939259	BODY	-3,2059531203	0,0037857179		

Figure 6. Gene UHRF1 methylation status

$$S_{cpg}^{R_k}(g_i) = \sqrt{\left| \min_{j=1..n_{cpg}^k} (S_{cpg_j}) * \max_{j=1..n_{cpg}^k} (S_{cpg_j}) \right|}, \; k = 1, 2, 3, 4 \tag{3.2}$$

In our first approach as illustrated in the following figure, the above formula computes a score for all CpGs that are in the specific region of gene g_i, where n_{cpg}^k is the number of CpGs in a specific region R_k, and R_1, R_2 and R_3 represent TSS, EXON1 and BODY regions respectively. This CpG-set score selects the maximum score, $\max(S_{cpg_j})$, and the minimum score, $\min(S_{cpg_j})$, among all values of single CpG scores that are in the region R_k of gene g_i.

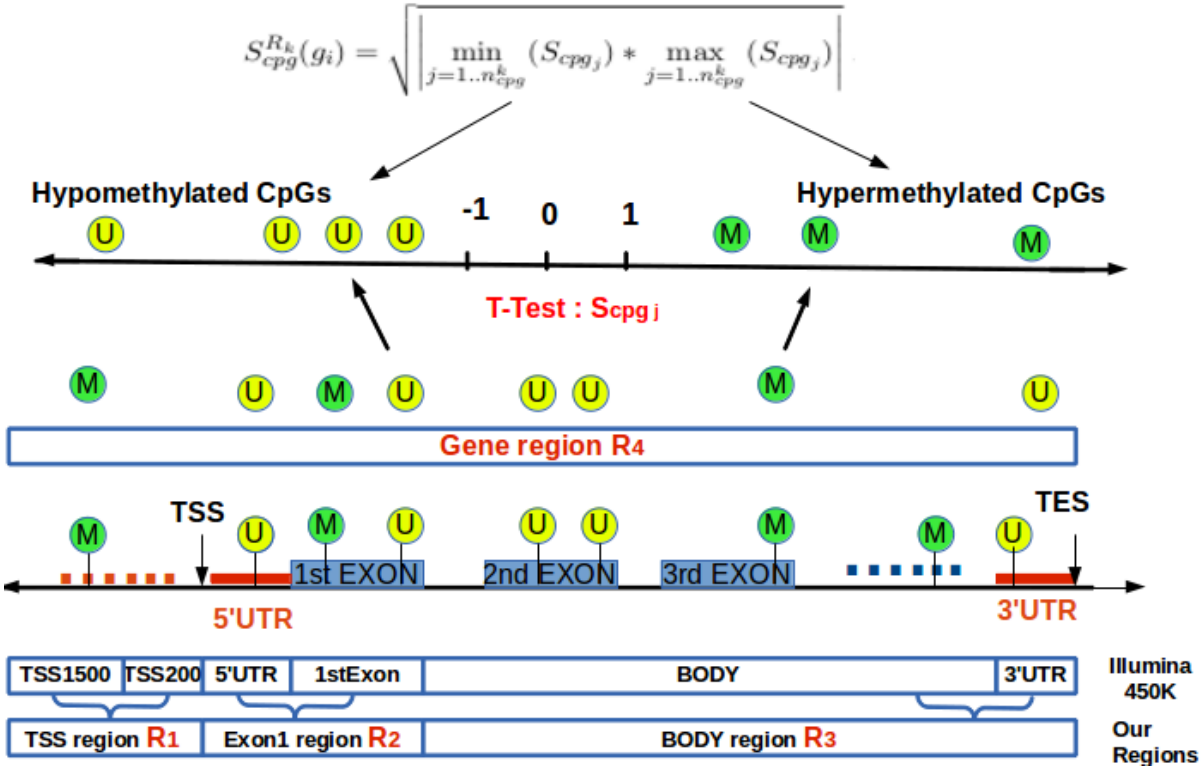


Figure 7. Scoring a CpG-Set of a Gene based on MinMax formula and different regions

Our second approach of scoring a set of CpGs targets all the CpG single scores of a particular region in order not to discard any value. To address the problem of differents signs due to CpGs having negative single scores when they are hypo-methylated and positive ones when hyper-methylated, we transformed all the CpGs single scores through an exponential function and associated them according to the following formula:

$$S_{cpg}^{R_k}(g_i) = \frac{\sum_{j=1}^{n_{cpg}^k} e^{f(S_{cpg_j})}}{n_{cpg}^k} \tag{3.3}$$

In the above formula, single scores S_{cpg_j} are rescaled to $f(S_{cpg_j})$ to fit into the interval [-2, 2] then raised to the exponential $e^{f(S_{cpg_j})}$ and finally all values averaged in one score $S_{cpg}^{R_k}(g_i)$ to represent the gene g_i (8).

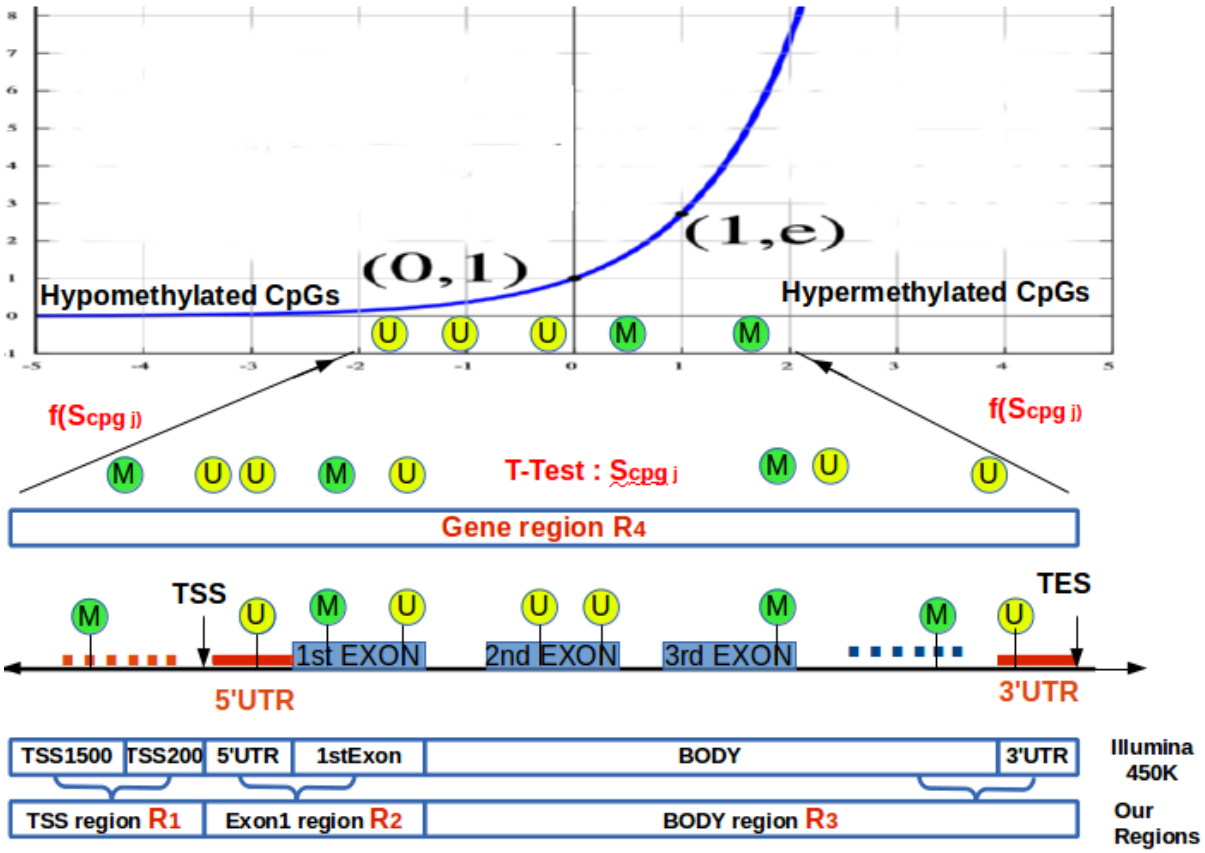


Figure 8. Gene score based on a CpG-Set using an Avg-Exponential formula and according to different regions

3.1.3 Gene Combined-Score : GCS

The gene combined score is the aggregation of the expression score of the gene, with its correspondent CpG-set score. We used either formula MinMax or Avg-Exponetial to combine the gene and CpG-set scores.

The first formula we call it formula MinMax:

$$S^{R_k}(g_i) = S_{exp}(g_i)\sqrt{\left|Min_{j=1..n^k_{cpg}}(S_{cpg_j}) * Max_{j=1..n^k_{cpg}}(S_{cpg_j})\right|} \tag{3.4}$$

The second formula we call it formula Avg-Exponential:

$$S^{R_k}(g_i) = e^{f(S_{exp}(g_i))}\frac{\sum_{j=1}^{n^k_{cpg}} e^{f(S_{cpg_j})}}{n^k_{cpg}} \tag{3.5}$$

3.1.4 Gene Set Score: GSEA

We performed a pathway analysis using the GSEA test and assessments as described in the previous chapter. However, we also used the GSEA in three different ways. First, we run the GSEA test using only the Gene Single-Score (9). Then, we run it using the CpG-Set Score (9). Finally, we used the combined score as illustrated by 4.

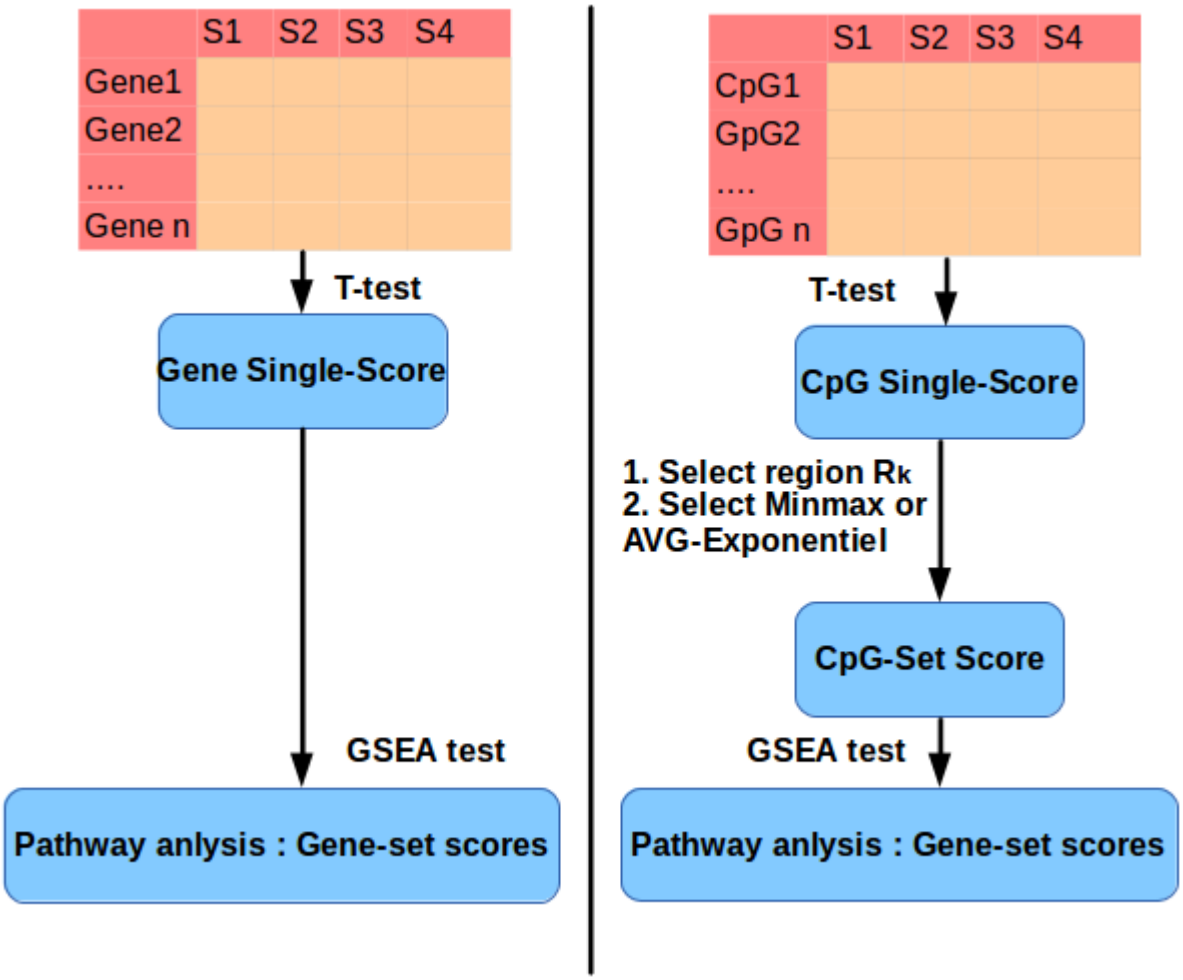


Figure 9. GSEA using first Gene Single-Score only, then CpG-Set Score only.

CHAPTER 4

DATASETS

This chapter starts by presenting the two datasets used for evaluation of our framework through the GSEA algorithm. These datasets are Breast Cancer (BC), invasive carcinoma type, and the Primary Lymphatic and Blood Endothelial Cells (LEC/BEC). Afterwards, the preprocessing steps for checking, cleaning and filtering the data are explained and illustrated through different graphs.

4.1 BC dataset

The first dataset we used is from the Cancer Genome Atlas (TCGA) which provides a data portal for researchers to search, download, and analyse data sets in cancer studies (32). The TCGA provides researchers with a research network, to consolidate different research efforts within a common infrastructure of experiments. Results are publicly available, in order to improve research findings and progress locally and globally (32).

We chose the breast cancer dataset, which is the most frequently diagnosed cancer and the second cause of cancer deaths in women (32). In 2010, according to the TCGA website, 207,090 women were estimated to have been diagnosed with invasive breast cancer in the United States and approximately 40,000 women were estimated to have died of the disease.

4.2 LEC/BEC dataset

The lymphatic system and blood vasculature share a similar molecular and developmental relationship, but they exhibit distinct features and functions (33). In fact, transitions between blood endothelial cells (BEC) and lymphatic endothelial cells (LEC) starts from the embryonic development and can occur even after terminal differentiation (33). Bronneke et al. conducted gene expression and CpGs methylation studies in LEC and BEC cells and identified a set of differentially methylated and expressed genes. Pathway analyses of the differentially methylated and upregulated genes in LEC revealed involvement in developmental and transdifferentiation processes (33). We also used this dataset to assess our satistical framework.

4.3 Data Pre-processing

4.3.1 Pre-processing BC dataset

The breast cancer dataset was suitable for our study due to its large number of case and control samples in gene expression and DNA methylation experiments. We obtained data from 27 patients, and each breast cancer patient had a sample from the solid tumor tissue (case sample) and another one from the normal tissue (control sample). The gene expression microarray experiment and the DNA CpG mehtylation microarray experiment used case and control samples from the same patients (see details in Annexe1).

The TCGA data is codified according to different criterion like project, participant, type of sample as shown in 10.

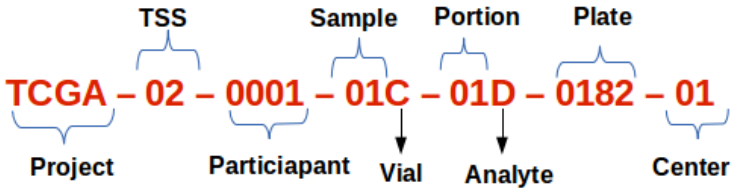
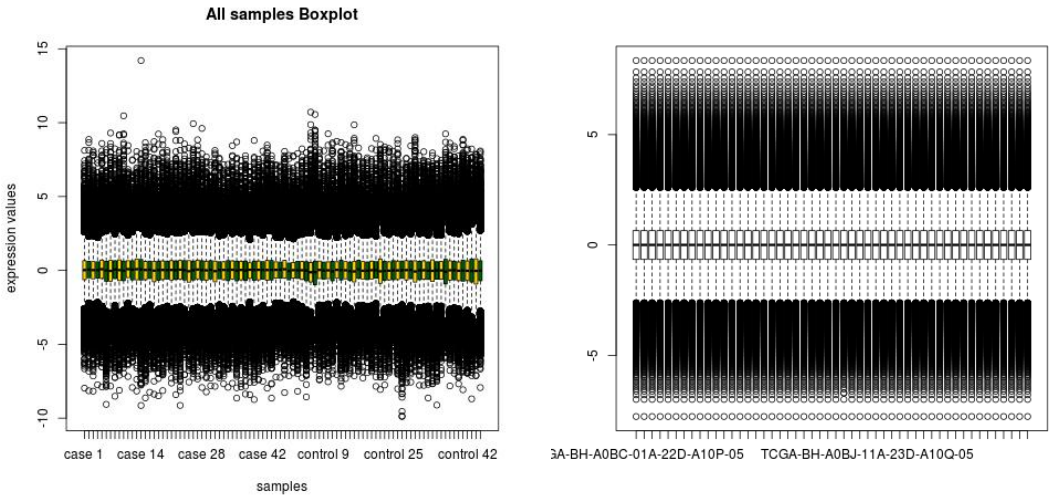


Figure 10. TCGA samples code explanation

For the complete list of our samples codes see Annexe1. The TCGA samples barcodes start with the project name which is the TCGA project for all the samples we had. Then, the tissue source site (TSS) represents the center that extracted the samples and for our samples we had "BH" which is the broad institute of MIT and Havard. The study participant code is the participant identifier. The sample type, denotes the tissue type, for example "01" stands for solid tumor, and "11" stands for normal tissue. The vial is the order of sample in a sequence of samples, for example, "A" is first sample, "B" is second, "C" is third and so forth. Then comes the portion number which is the order of portion in a sequence of 100-120 mg sample portions. The portion is associated to the analyte representing the molecular type of the sample for analysis, for example "D" for DNA or "R" for mRNA. Finally we have the plate order number and the code of the center that performed the sequencing or the microarrays experiment. For our study, the gene expression experiment was performed by University of North Carolina center "07" and the DNA methylation experiment was performed by the John hopkins/University of Southern California center "05".

The gene expression data was already pre-processed, and the level 3 folder of the downloaded BC-dataset contained the genes intensities for every sample. However, we needed to quantile normalize all the samples as showed in the 11.



(a) Before normalization. (b) After normalization.

Figure 11. Normalized gene samples: we used quantile normalization method.

The methylation data, for BC dataset and also LEC/BEC dataset, were obtained from microarray experiment using Illumina Infinium HumanMethylation450 BeadChip. For both datasets, each CpG has a beta value computed as the following:

$$\beta = \frac{\text{Methylation intensity (M)}}{\text{(Unmethylation intensity (U) + Methylation intensity (M) + 100)}} \tag{4.1}$$

The Illumina Infinium HumanMethylation450 BeadChip uses two types of probes, infinium I and Infinium II, and a bias in this platform has been reported and studied previously (34) (35) (36) (37). In fact, the CpGs probes are designed differently (34) and have different beta value distributions as presented in the distributions and densities of the BC dataset (17). Therefore, a correction is needed to ensure the effectiveness of the downstream analysis (35) (36) (37). First, we present our pre-processing pipeline in the following manner:

We used the methylumi bioconductor package to import the samples microarray files that have the extension .idat and two color channels green and red for each sample. We retrieved 54 samples, each having 485,577 CpGs. The wateRmelon bioconductor package provides a function to filter unreliable probes, which in our case indicated that zero samples having 1 % of sites with a detection p-value greater than 0.05 were removed, 4,358 sites were removed as beadcount less than 3 in 5 % of samples and 8,943 sites having 1 % of samples with a detection p-value greater than 0.05 were removed. We also used lumi bioconductor package to correct for color bias as shown in 13 and 14 and to adjust the background as indicated by 15.

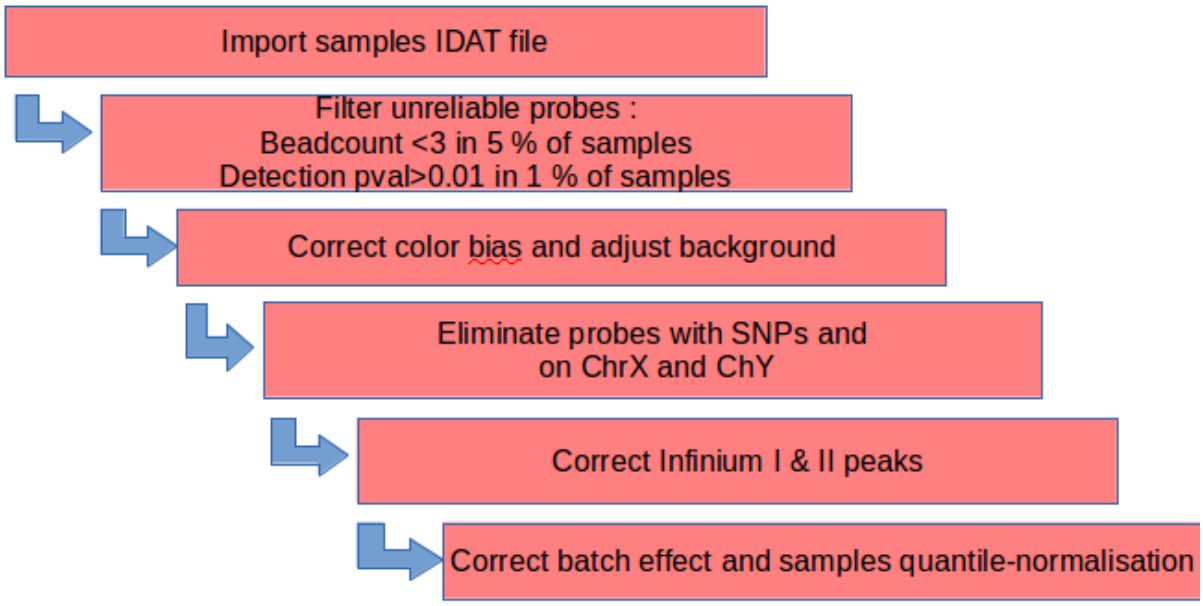


Figure 12. CpG Methylation pre-processing steps

It is worth to mention that the color bias correction is only applied to the Infinium II probes because they use one bead to catch the methylated and unmethylated probes. By contrast, the Infinium I probes have only one color channel by using two beads, one for the methylated and the second for the unmethylated. After correcting the color bias and adjusting the background we checked for batch effect and found that two patients samples data were outliers even after normalization and the batch effect was not effective, so we eliminated them.

The next step of our preprocessing pipeline was to eliminate the sex bias by filtering 11,112

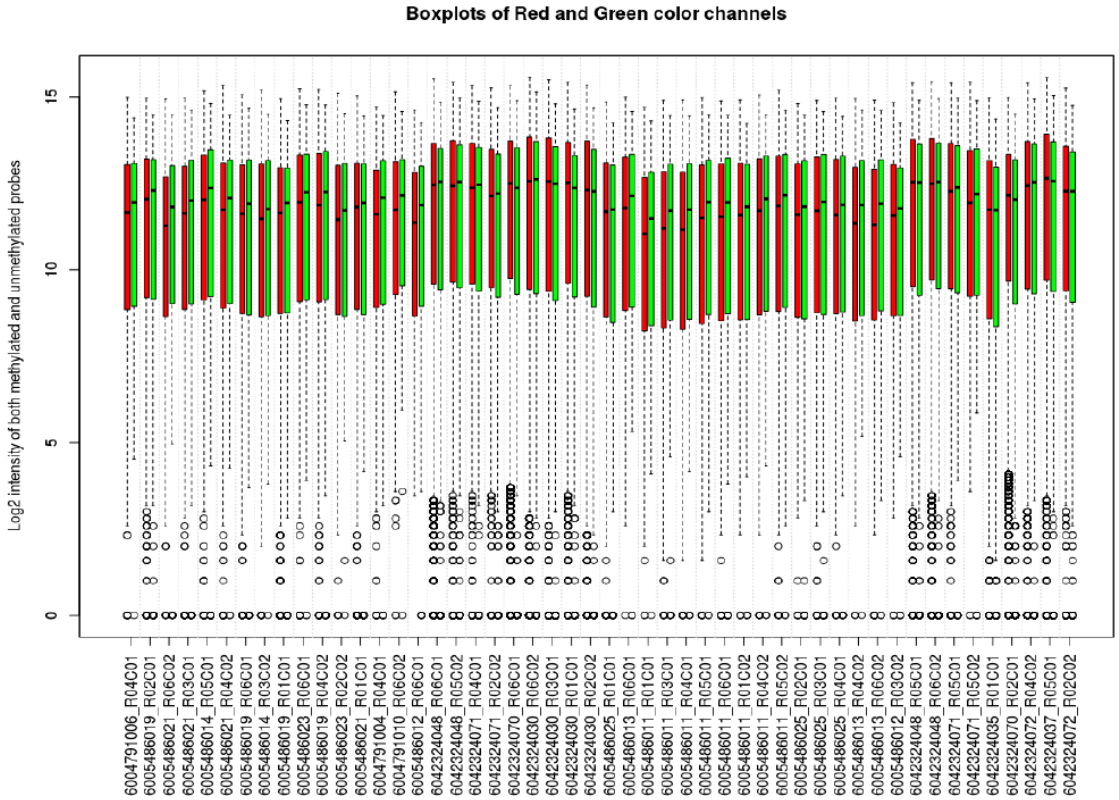


Figure 13. Samples color channels having color bias

probes that are on the X chromosome and 41 from Y chromosome and the population ethnicity bias by filtering 85,608 probes with SNPs. After eliminating all the wanted probes that could present a bias and the samples outliers, we were left with 379873 probes and 50 samples. We separated the probes in two groups, Infinium I and II and we obtained 106325 probes in type I and 273548 probes in type II.

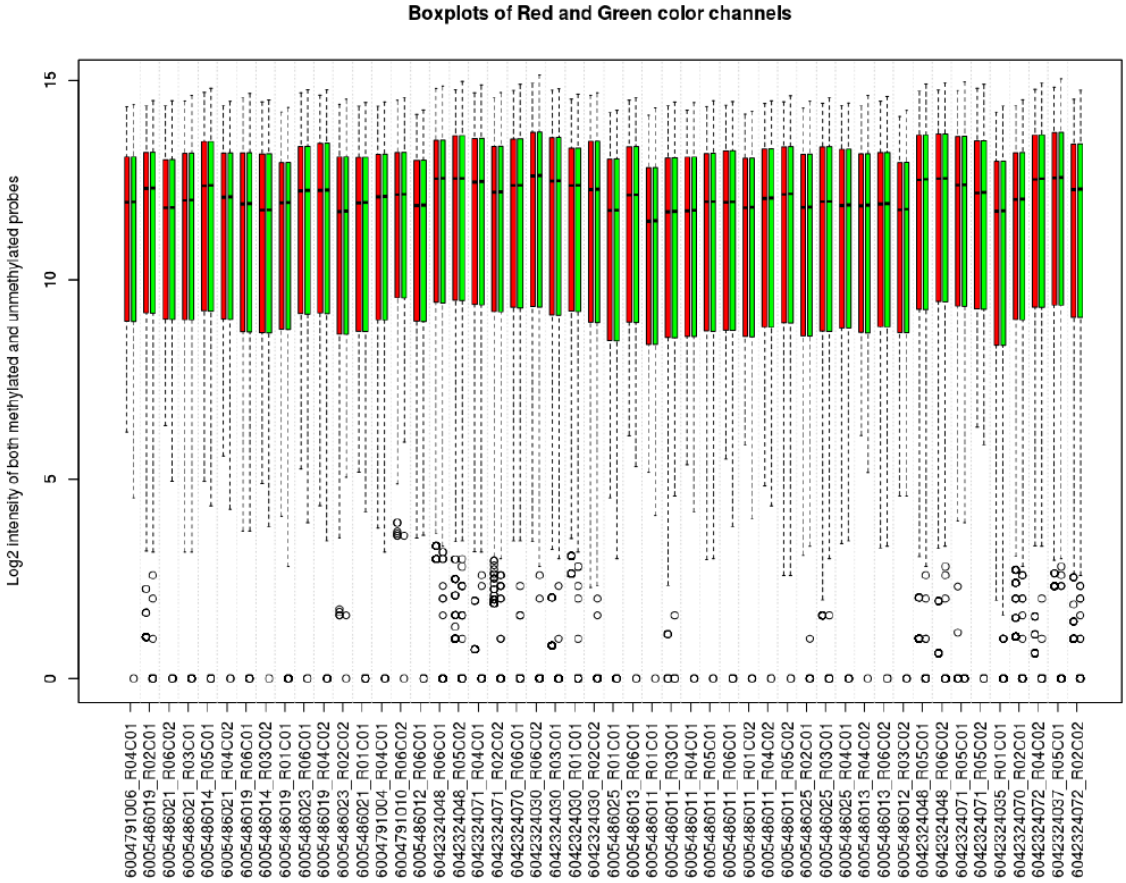


Figure 14. Samples color channels after color bias correction

To correct the two types of probes peaks bias due to their different bead design by the constructor, many methods were proposed and we used a beta mixture inter-quantile method (37) that was implemented in the bioconductor package wateRmelon. After the correction of the Infinium type I and II peaks (16), we normalized them separately (17 and 18) with quantile normalization method then we remerged them again (20).

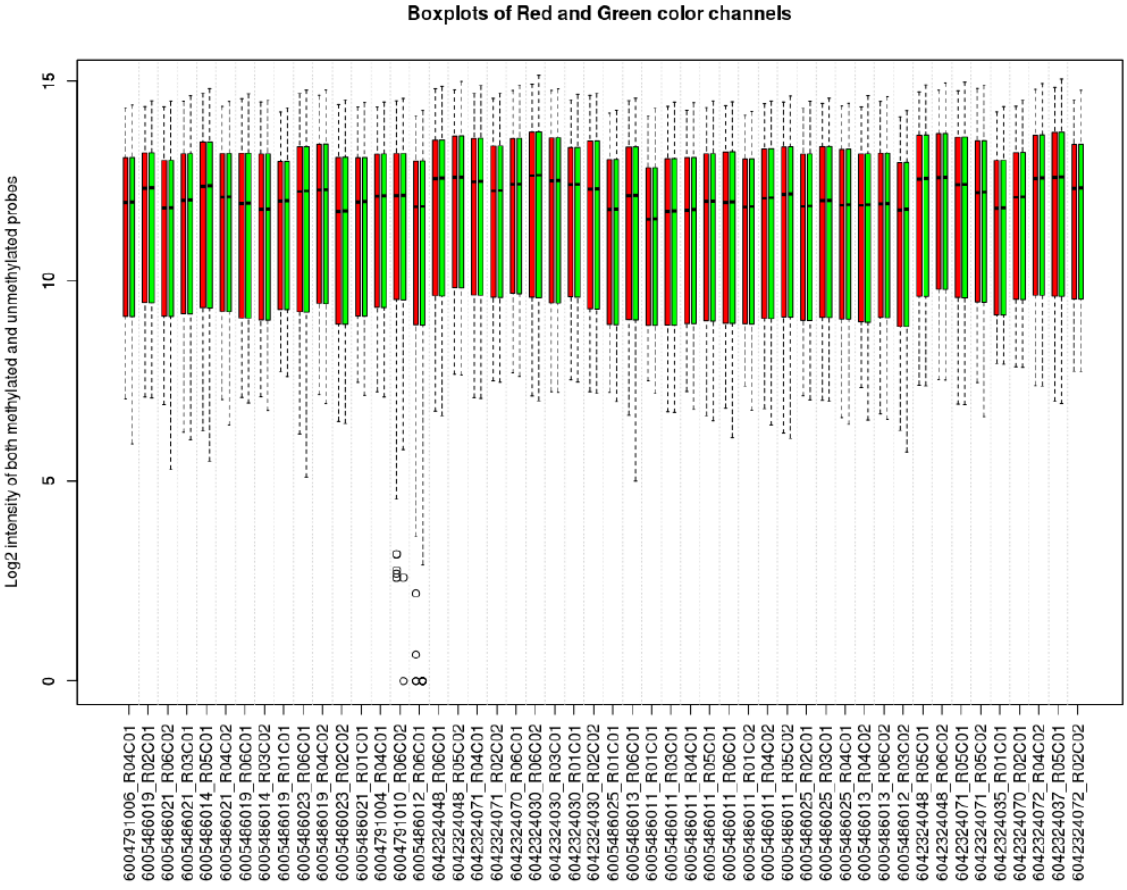
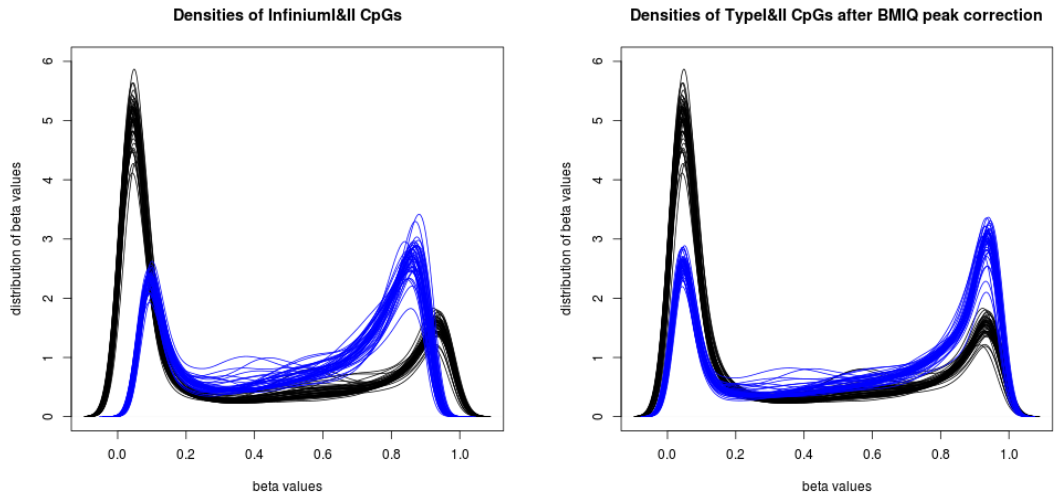


Figure 15. Samples color channels after background adjustment



(a) Before peak correction.

(b) After peak correction.

Figure 16. Boxplot of CpGs-methylation infinium I & II. on the left figure the Infinium I (black) and the infinium II (blue) have different distributions for the bumps that was corrected using the beta-mixture quantile normalization method (37) as shown on the right figure

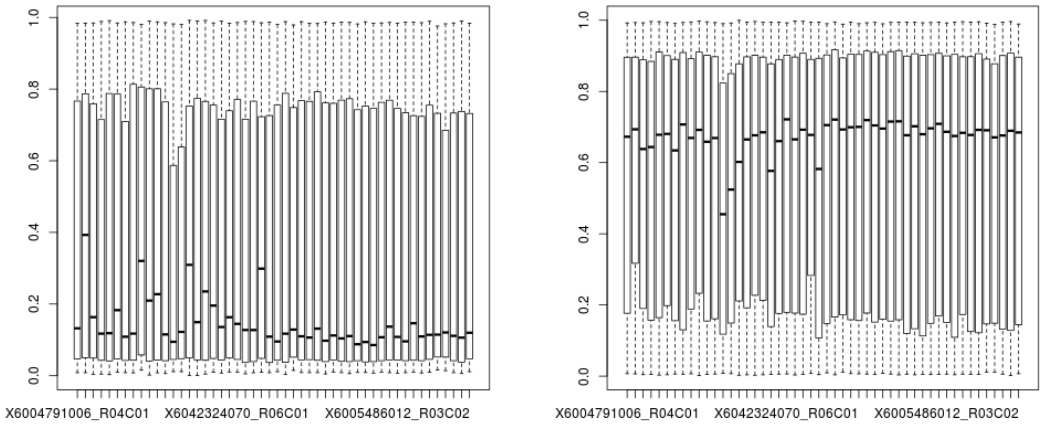


Figure 17. Infinum I&II boxplot after peak correction

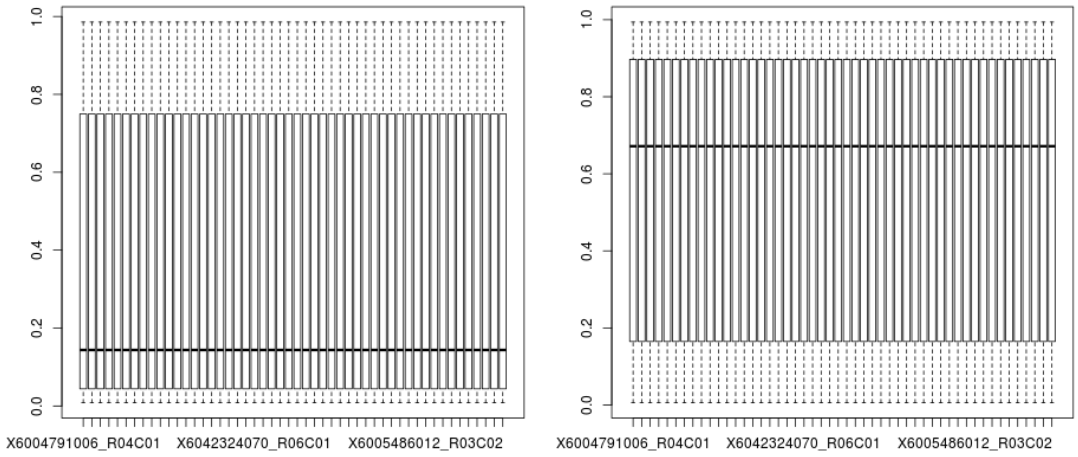


Figure 18. Infinum I&II boxplot after normalization

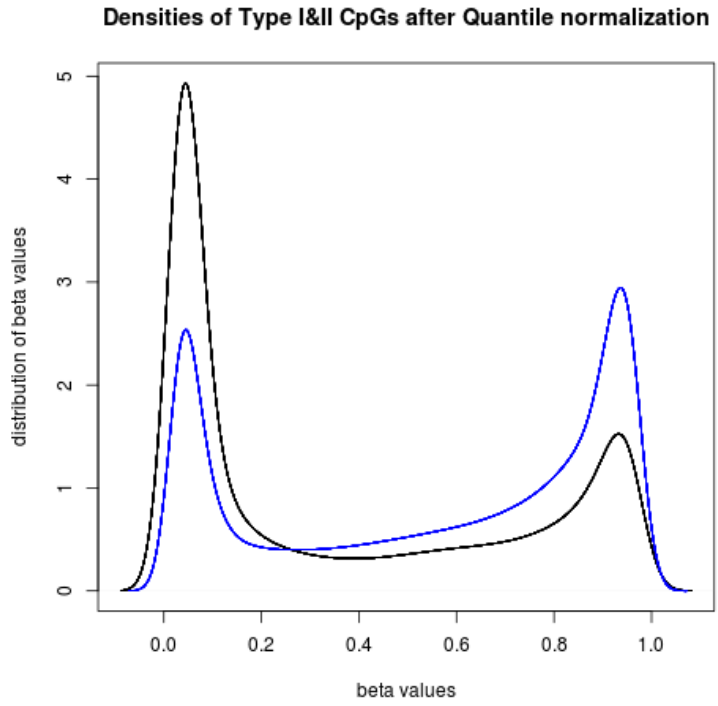


Figure 19. Density of Infinum I and II probes after normalization

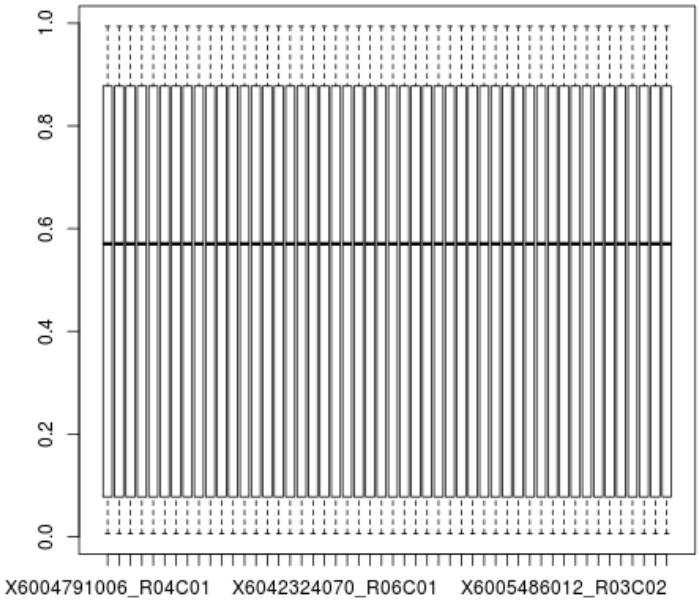


Figure 20. Boxplot after merging all probes from InfiniumI and II

After finishing with all the preprocessing steps, we proceeded to filter our data to fit the requirements of our statistical framework. Some genes did not have associated CpGs so we discarded them and some CpGs were also eliminated because they did not have associated genes (TABLE I). Grouping the CpGs into regions also reduced the number of genes and CpGs (TABLE II).

	Genes	*CpGs*
Initial number	17,811	485,577
After filtering	15,320	379,873
Matched number	14,731	246,835

TABLE I

NUMBER OF GENES HAVING CPGS AND VICE VERSA

	Genes	*CpGs*
Initial Total number	14,731	246,835
TSS	13,439	70,363
EXON1	12,151	47,890
BODY	13941	128,582

TABLE II

NUMBER OF GENES AND CPGS PER REGION

4.3.2 Pre-processing LEC/BEC dataset

The LEC/BEC dataset has less number of samples, 10 LEC and 6 BEC, in gene expression and DNA methylation experiments. The expression data were obtained from Agilent microarray experiment. The values were log2 transformed and normalized as shown by 21.

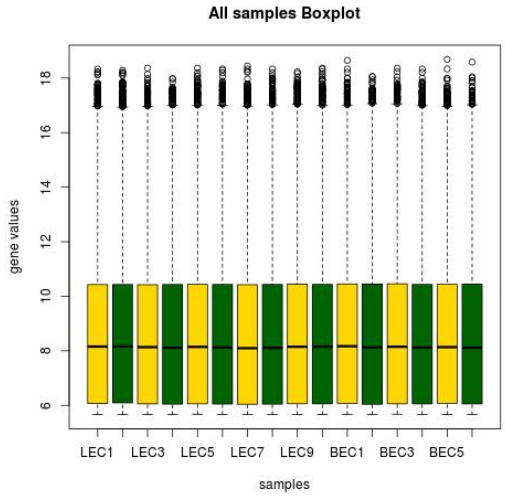


Figure 21. Normalized gene-expression samples

We also poltted the Principal Component Analysis(PCA) to verify that there were no outliers in the experiment (22).

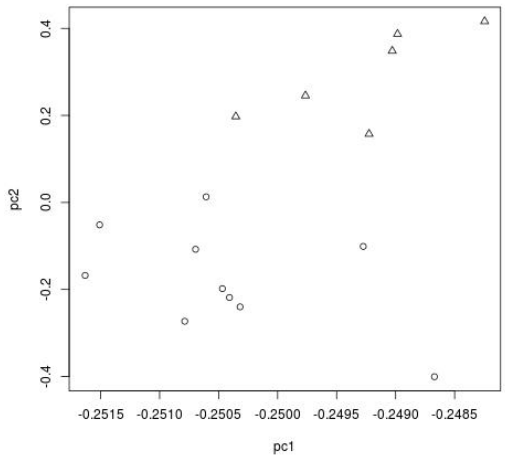


Figure 22. PCA plot of the samples

We plotted the density of LEC samples as well as the BEC samples in 23.

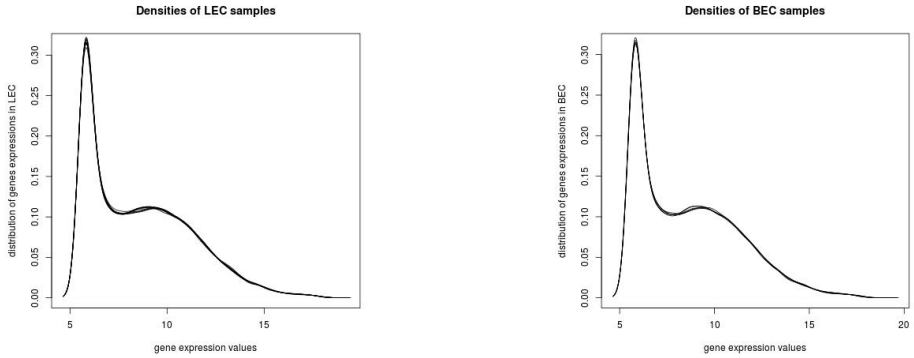


Figure 23. Density plot of gene-expression samples

Concerning the methylation data, the Beta values are normalized and have no outliers as confirmed by the boxplot and PCA plots shown in 24 and 25 and the density plot in 26.

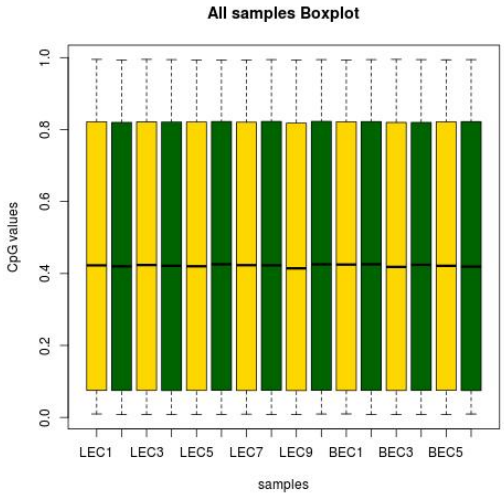


Figure 24. Boxplot of normalized samples

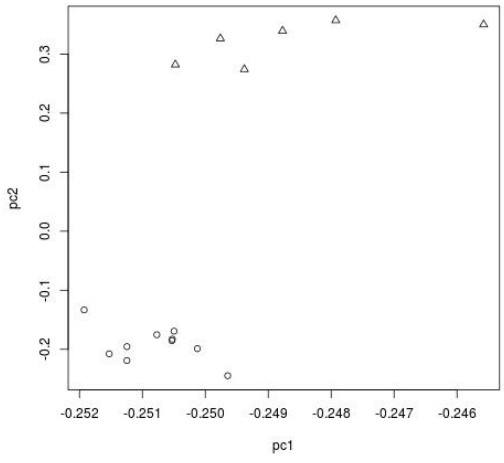


Figure 25. PCA plot of the samples

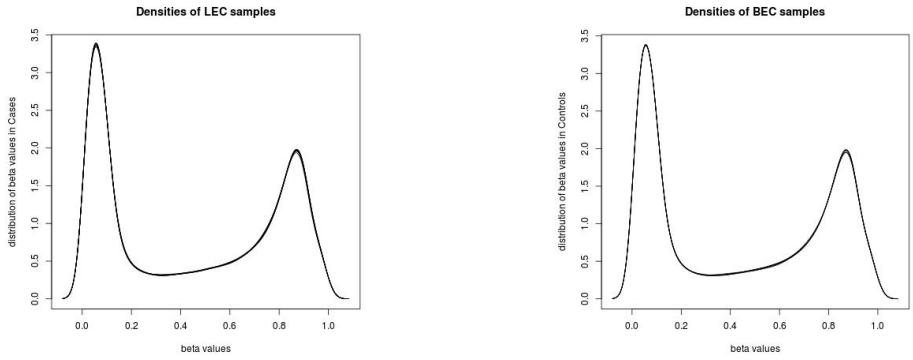


Figure 26. Density plot of all samples CpGs methylation

Since we integrate gene expression information and CpG sites methylation information at the gene level, we needed to discard the genes that don't have CpGs and the CpGs that do not belong to any gene (TABLE III).

	Genes	*CpGs*
Initial number	18,691	485,512
After eliminating those with no values	18,691	182,937
Matched number	16,887	163,467

TABLE III

NUMBER OF GENES HAVING CPGS AND VICE VERSA

Also, grouping the CpGs into regions reduced the number of genes and CpGs as follow:

	Genes	*CpGs*
Initial Total number	16,887	163,467
TSS	15,281	52,899
EXON1	11,603	30,195
BODY	14,375	80,373

TABLE IV

NUMBER OF GENES AND CPGS PER REGION

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents the experimental results of our statistical framework tested on the BC and LEC/BEC datasets. The results are highlighted in summary tables and more details about the detected pathways can be found in subsequent tables. We also validated our results through the literature.

5.1 BC Results and Literature Validation

- Based on Gene Single Score (GSS)

We start by ranking genes with at least one CpG in the ascending order of GSS. The rank list was used as input for the GSEA algorithm. We considered all the genes when we counted for the CpGs that are in the gene region. However, when we restrict the CpGs to the TSS or EXON1 or BODY regions, then the number of genes will be reduced as was described in chapter 4 (Table II).

	Formula MinMax	*Formula AvgExponentiel*
TSS Region	0 pathways	0 pathways
EXON1 Region	0 pathways	0 pathways
BODY Region	0 pathways	0 pathways
Gene Region	TABLE VI	TABLE VII

TABLE V

SUMMARY OF RESULTS USING GENE SINGLE-SCORE

The GSS did not detect any significant pathways with FDR cutoff less than 25%, except for gene region that gave two significant pathways as presented in TABLE VI and TABLE VII. Consulting the literature, K1 is associated with proliferation of human breast cancer cells due to the dis-regulation of ribosome bio-genesis and translational capacity (48). The pathway taurine and hypotaurine metabolism (K0) is also significant as taurine is abundant in the brain, heart, breast,... and has important roles in health and disease in these organs (19). Moreover, taurine is related to tumor cells (21) and is significantly involved with breast cancer (22).

TABLE VI. PATHWAYS IN BC DETECTED BY GENE SINGLE SCORE (FORMULA MINMAX, GENE REGION)

Index	Geneset Name	P	FDR
K0	KEGG_TAURINE_AND_HYPOTAURINE_METABOLISM	0.0580	0.1977
K1	KEGG_RIBOSOME	0.1090	0.1890

TABLE VII. PATHWAYS IN BC DETECTED BY GENE SINGLE SCORE (FORMULA AVG-EXPONENTIEL, GENE REGION)

Index	Geneset Name	P	FDR
K0	KEGG_TAURINE_AND_HYPOTAURINE_METABOLISM	0.0550	0.1882
K1	KEGG_RIBOSOME	0.1030	0.2360

- Based on Gene CpG-Set Score

The scoring of genes based on CpGs-set gave a more significant pathways than the GSS. A summary of these results is presented in TABLE VIII and more details about the significant detected pathways is given in the subsequent tables.

	Formula MinMax	*Formula AvgExponentiel*
TSS Region	TABLE IX	TABLE XIII
EXON1 Region	TABLE X	TABLE XVIII
BODY Region	TABLE XI	0 pathways
Gene Region	TABLE XII	TABLE XV

TABLE VIII

SUMMARY OF RESULTS USING CPG-SET SCORES

TABLE IX. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA MINMAX TSS REGION)

Index	Geneset Name	P	FDR
K1	KEGG_RIBOSOME	0.0000	0.0000
K2	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.0000	0.0000
K3	KEGG_RNA_DEGRADATION	0.0000	0.0006
K4	KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0000	0.0010
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0020	0.0040
K6	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.0000	0.0136
K7	KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	0.0010	0.0217
K8	KEGG_SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT	0.0010	0.0236
K9	KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	0.0000	0.0410
K10	KEGG_NON_HOMOLOGOUS_END_JOINING	0.0040	0.0450
K11	KEGG_CELL_CYCLE	0.0000	0.0453
K12	KEGG_SPLICEOSOME	0.0000	0.0523
K13	KEGG_PURINE_METABOLISM	0.0060	0.1303
K14	KEGG_GRAFT_VERSUS_HOST_DISEASE	0.0010	0.1365
K15	KEGG_PROTEASOME	0.0010	0.1630
K16	KEGG_OOCYTE_MEIOSIS	0.0000	0.1670
K17	KEGG_TIGHT_JUNCTION	0.0120	0.1941
K18	KEGG_ALZHEIMERS_DISEASE	0.0020	0.2004
K19	KEGG_ASTHMA	0.0020	0.2058
K20	KEGG_ONE_CARBON_POOL_BY_FOLATE	0.0150	0.2132
K21	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_BIOSYNTHESIS	0.0420	0.2185
K22	KEGG_HUNTINGTONS_DISEASE	0.0050	0.2239
K23	KEGG_NOTCH_SIGNALING_PATHWAY	0.0230	0.2279
K24	KEGG_VEGF_SIGNALING_PATHWAY	0.0150	0.2303
K25	KEGG_RNA_POLYMERASE	0.0410	0.2410

In general, the CpG-set scoring led to more results than GSS or GCS. For BC dataset, the TSS region gave a large number of pathways that were significant and at the same time biologically relevant. The neuroactive ligand receptor interaction pathway (K2) was reported as highly significant in the breast cancer cell line MCF-7 treated with 17β Estradiol (49). In addition, other significant pathways such as cytokinecytokine receptor interaction (K6), calcium

signaling (K53), cell adhesion molecules CAMs (K33) axon guidance (K50) and ErbB signaling pathway were also reported as relevant to cancer (49).

The ubiquitin mediated proteolysis pathway (K4) has a very important role in the molecular basis of carcinogenesis and specifically in breast cancer (51) (52) and a study about applying drugs affecting the ubiquitin-proteasome pathway to the therapy of breast cancer has been proposed (53). The induction of olfaction and cancer-related genes in mice fed with a high-fat diet also propose the olfactory transduction pathway (K5) as a cancer related pathway (54) and highly correlated to breast cancer (55). The snare interactions in vesicular transport pathway (K8) is regulated by Rab GTPases and it was shown that Rab25 is upregulated in certain ovarian and breast cancers due to amplification of a chromosomal region containing the Rab25 gene (58).

A study was conducted about neurotrophins and their receptors in breast cancer which relates neurotrophin signaling pathway (K9)to breast cancer (59). In fact, nerve growth factor stimulates proliferation and survival of human breast cancer cells through a specific signaling pathways (60). In addition, nerve growth factor promotes breast cancer angiogenesis by activating multiple pathways (61). We also found a multigenic study on cancer susceptibility stating that the risk to breast cancer was associated with genotypic polymorphism of the non-homologous end-joining (K10) genes (62).

Furtheremore, one study identified a gene signature in cell cycle pathway (K11) for breast cancer prognosis using gene expression profiling data (63). Moreover, in cancer cells the deregulated spliceosome (K12) core machinery can be targeted for potential therapy and the interaction

of p53 and SAP145 represents a novel role for p53 in splicing (64). This finding is likely to have direct implications for breast cancer research, considering p53 and cyclin E as prognostic markers for breast cancer, since both proteins converge their pathways at the spliceosome (65). Another point, concerns the effect of methotrexate on intracellular folate pools in human MCF-7 breast cancer cells, which was found as an evidence for direct inhibition of purine synthesis (K13) (66) (67). Molecularly targeted therapies for breast cancer, do also include the proteasome pathway (K15) especially for chemotherapy (69).

So far, the evidences reported in literature support the effectiveness of our proposed framework. In addition, our method also suggests new significant pathways involved with the methylation of the TSS region of certain genes that can be strongly related to breast cancer and can be further investigated to become a new target treatment. The RNA degradation pathway (K3) has been found to regulate GAS5 function which is a non-coding RNA in mammalian cells and far less is known about the mechanisms and biological importance of ncRNA (50). Furthermore, the aminoacyl-tRNA biosynthesis pathway (K7) can be important in breast cancer since the aminoacyl-tRNA synthetase is deeply involved with cancer disease such as in glioblastoma tumor (56) and potentielly in breast cancer (57). Additionally, the induction of graft versus host disease pathway (K14) has been proposed as an immuno-therapy for relapsed chronic myleoid leukemia (68) and may also be breast cancer related.

TABLE X. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA MINMAX EXON1 REGION)

Index	Geneset Name	P	FDR
K2	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.0000	0.0000
K4	KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0010	0.0220
K11	KEGG_CELL_CYCLE	0.0000	0.0235
K26	KEGG_BASE_EXCISION_REPAIR	0.0000	0.0306
K27	KEGG_PHENYLALANINE_METABOLISM	0.0010	0.0706
K28	KEGG_DILATED_CARDIOMYOPATHY	0.0000	0.0707
K29	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	0.0010	0.0772
K30	KEGG_DNA_REPLICATION	0.0030	0.0772
K31	KEGG_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI_ANCHOR_BIOSYNTHESIS	0.0030	0.0798
K6	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.0050	0.0898
K32	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.0030	0.0920
K33	KEGG_CELL_ADHESION_MOLECULES_CAMS	0.0040	0.0947
K17	KEGG_TIGHT_JUNCTION	0.0080	0.0995
K34	KEGG_PROPANOATE_METABOLISM	0.0040	0.1052
K12	KEGG_SPLICEOSOME	0.0050	0.1165
K35	KEGG_N_GLYCAN_BIOSYNTHESIS	0.0030	0.1272
K1	KEGG_RIBOSOME	0.0140	0.1648
K36	KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	0.0100	0.1888
K37	KEGG_VIRAL_MYOCARDITIS	0.0120	0.2367
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0010	5e-04

TABLE XI. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA MINMAX BODY REGION)

Index	Geneset Name	P	FDR
K15	KEGG_PROTEASOME	0.002	0.134
K20	KEGG_ONE_CARBON_POOL_BY_FOLATE	0.003	0.1956
K1	KEGG_RIBOSOME	0.008	0.24125

CpG-Set Scoring using the MinMax formula in EXON1 region detected many interesting significant pathways, and some of them are overlapping with the ones detected in the TSS region and they are highlighted in TABLE X. Also, when CSeS scoring in BODY region was

used, only 3 significant pathways were detected and they all overlap with the ones detected by using the same scoring in TSS and/or EXON1 region. The pathways detected by using CSeS scores in EXON1 region include also breast cancer related pathways and others are unexplored but could be a new topic or hypothesis of research and interest.

The base-excision repair pathway (K26) was significant and has also genetic polymorphisms in its genes presenting a risk for breast cancer (70). As a therapeutic strategy, it was also suggested to target the DNA repair defect in BRCA mutant cells (71). Phenylalanine metabolism pathway(K27) is very important in cancer since an analysis of metabolic correlation network was established and found it as a potential biomarkers for breast cancer (72), in addition, it is already used in tumor cell metabolism imaging (73). Another study related dilated cardiomyopathy (K28) and HER2-Positive breast cancer (74) through the interaction between specialties. The amplification of the gene encoding the ErbB2 (Her2/neu) receptor tyrosine kinase is also critical for the progression of many forms of breast cancer and it is showed that ErbB2 is essential in the prevention of dilated cardiomyopathy (75). DNA damage response pathwya was suggested as a candidate for anti-cancer barrier in early human tumorigenesis (76). Therefore, the targeting of DNA replication (K30) before it starts has been elaborated by proposing Cdc7 as a therapeutic target in p53-mutant breast cancers (77).

TABLE XII. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA MINMAX GENE REGION)

Index	Geneset Name	P	FDR
K38	KEGG_ECM_RECEPTOR_INTERACTION	0.000	0.0120
K39	KEGG_PARKINSONS_DISEASE	0.0010	0.0150
K15	KEGG_PROTEASOME	0.0000	0.0177
K22	KEGG_HUNTINGTONS_DISEASE	0.0010	0.0180
K20	KEGG_ONE_CARBON_POOL_BY_FOLATE	0.0010	0.0230
K40	KEGG_NON_HOMOLOGOUS_END_JOINING	0.0000	0.0306
K1	KEGG_RIBOSOME	0.0030	0.0405
K41	KEGG_OXIDATIVE_PHOSPHORYLATION	0.0140	0.0490
K42	KEGG_FATTY_ACID_METABOLISM	0.0240	0.1164
K31	KEGG _GLYCOSYLPHOSPHATIDYLINOSITOL_GPI_ANCHOR_BIOSYNTHESIS	0.0150	0.1223
K43	KEGG_PEROXISOME	0.0200	0.1243
K18	KEGG_ALZHEIMERS_DISEASE	0.0040	0.1317
K3	KEGG_RNA_DEGRADATION	0.0130	0.1444
K44	KEGG_OTHER_GLYCAN_DEGRADATION	0.0040	0.1540
K45	KEGG_BUTANOATE_METABOLISM	0.0180	0.1557
K29	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	0.0210	0.2027
K46	KEGG_REGULATION_OF_AUTOPHAGY	0.0310	0.2243

TABLE XIII. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA AVG-EXPONENTIEL TSS REGION)

Index	Geneset Name	P	FDR
K14	KEGG_GRAFT_VERSUS_HOST_DISEASE	0.0060	0.1030
K47	KEGG_ALLOGRAFT_REJECTION	0.0080	0.1060
K48	KEGG_AUTOIMMUNE_THYROID_DISEASE	0.0050	0.1080
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0040	0.1100
K49	KEGG_TYPE_I_DIABETES_MELLITUS	0.0030	0.1100
K19	KEGG_ASTHMA	0.0070	0.1315
K2	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.0090	0.1980
K25	KEGG_RNA_POLYMERASE	0.0210	0.2398
K6	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.0260	0.2463
K1	KEGG_RIBOSOME	0.0130	0.2471

TABLE XIV. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA AVG-EXPONENTIEL EXON1 REGION)

Index	Geneset Name	P	FDR
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0010	0.0730
K50	KEGG_AXON_GUIDANCE	0.0020	0.0731
K51	KEGG_MELANOGENESIS	0.0040	0.0731
K52	KEGG_BASAL_CELL_CARCINOMA	0.0040	0.0760
K2	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.0090	0.0780
K28	KEGG_DILATED_CARDIOMYOPATHY	0.0020	0.0788
K53	KEGG_CALCIUM_SIGNALING_PATHWAY	0.0010	0.092
K41	KEGG_ECM_RECEPTOR_INTERACTION	0.0030	0.0966
K4	KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0050	0.0986
K12	KEGG_SPLICEOSOME	0.0030	0.1180
K54	KEGG_PATHWAYS_IN_CANCER	0.0070	0.1353
K55	KEGG_WNT_SIGNALING_PATHWAY	0.0120	0.1914
K56	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	0.0050	0.1974
K57	KEGG_FOCAL_ADHESION	0.0160	0.1986
K0	KEGG_TAURINE_AND_HYPOTAURINE_METABOLISM	0.0020	0.2030
K58	KEGG_HEDGEHOG_SIGNALING_PATHWAY	0.0180	0.2101
K59	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_HEPARAN_SULFATE	0.0060	0.2122

For the BC dataset, we found that the best results were generated by the formula AvgExponentiel (Equation 3.3) when applied to the EXON1 region using GSeS (TABLE XVIII) since it detected multiple cancer pathways such K54 and especially breast cancer pathway, basal cell carcinoma (K52). In addition, this scoring combined with this region detected only relevant pathways like the calcium signaling pathway (K53) which is highly correlated to cancer and the melanogenesis pathway (K51) that has also been related to breast cancer cell lines in different studies (78). Moreover, there is an evidence that transgenes encoding components of the Wnt signaling pathway (K55) preferentially induce mammary cancers from progenitor cells (79). Another point, concerning the gene expression programs of human smooth muscle cells

pathways (K56), it showed a tissue-specific differentiation and prognostic significance in breast cancers (80). A study of the focal adhesion kinase (K57) and p53 signal transduction pathways in cancer (81) has been related to migration and survival of breast cancer cells (82). Finally, the hedgehog signalling pathway (K58) is well studied in breast development, carcinogenesis and cancer therapy (83) and the heparan-sulfate and glycosaminoglycans pathway(K59) have an important role in cancer in general (84) and especially in breast cancer (85).

TABLE XV. PATHWAYS IN BC DETECTED BY CPG-SET SCORE (FORMULA AVG-EXPONENTIEL, GENE REGION)

Index	Geneset Name	P	FDR
K28	KEGG_DILATED_CARDIOMYOPATHY	0.0010	0.1410
K30	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.0040	0.2226
K24	KEGG_ENDOCYTOSIS	0.0010	0.2302
K15	KEGG_ADHERENS_JUNCTION	0.0050	0.2408

- Based on Gene Combined-Score

The GCS is the aggregation of the GSS and the CSeS. This scoring procedure gave also better results than a gene single score results with more significant FDRs. Using formula MinMax the EXON1 region detected the pathway K60 glutathione metabolism with Pvalue= 0.0000 and FDR=0.0560. Also, using the whole gene as region, only one pathway was detected drug metabolism cytochrome p450 (K61) with Pvalue=0.0010 and FDR=0.2000.

	Formula MinMax	Formula AvgExponentiel
TSS Region	0 pathways	0 pathways
EXON1 Region	1 pathway	TABLE XVIII
BODY Region	TABLE XVII	0 pathways
Gene Region	1pathway	0 pathways

TABLE XVI

SUMMARY OF RESULTS USING GENE COMBINED SCORE

TABLE XVII. PATHWAYS IN BC DETECTED BY GENE COMBINED SCORE

(FORMULA MINMAX, BODY REGION)

Index	**Geneset Name**	**P**	**FDR**
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0000	0.0000
K61	KEGG_DRUG_METABOLISM_CYTOCHROME_P450	0.0000	0.0575
K62	KEGG_LYSINE_DEGRADATION	0.0000	0.0970
K63	KEGG_PRIMARY_BILE_ACID_BIOSYNTHESIS	0.001	0.1743

TABLE XVIII. PATHWAYS IN BC DETECTED BY GENE COMBINED SCORE

(FORMULA AVG-EXPONENTIEL, EXON1 REGION)

Index	**Geneset Name**	**P**	**FDR**
K5	KEGG_OLFACTORY_TRANSDUCTION	0.0070	0.0660
K64	KEGG_ENDOCYTOSIS	0.0000	0.2130

5.2 LEC/BEC Results and Literature Validation

- Based on Gene Single Score

Based on this type of scoring the first formula did not lead to any significant pathways. A summary of the results is presented in the following table:

	Formula MinMax	*Formula AvgExponentiel*
TSS Region	0 pathways	0 pathways
EXON1 Region	0 pathways	0 pathways
BODY Region	0 pathways	0 pathways
Gene Region	0 pathways	0 pathways

TABLE XIX

SUMMARY OF RESULTS USING GENE SINGLE SCORE

- Based on Gene CpG-Set Score

CSeS in TSS region gave two significant pathways using minMax formula and only one by using formula AvgExponentiel. Also, one significant pathway in EXON1 region was detected. All the significant pathways are listed below (TABLE XXI)with their correspondant Pvalues and FDRs.

	Formula MinMax	Formula AvgExponentiel
TSS Region	2pathways	1pathway
EXON1 Region	0 pathways	1pathway
BODY Region	0 pathways	0 pathways
Gene Region	TABLE XXII	0 pathways

TABLE XX

SUMMARY OF RESULTS USING CPG-SET SCORE

TABLE XXI. PATHWAYS IN LEC/BEC DETECTED BY CPG-SET SCORE

Index	**Geneset Name**	**P**	**FDR**
	Formula MinMAx TSS region		
K1	KEGG_CELL_ADHESION_MOLECULES_CAMS	0.008	0.169
K2	KEGG_HEMATOPOIETIC_CELL_LINEAGE	0.001	0.114
	Formula MinMax EXON1 region		
K4	KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	0.001	0.035
	Formula AvgExponentiel TSS region		

Continued on next page

Continued from previous page

Index	Geneset Name	P	FDR
K3	KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	0.002	0.072

TABLE XXII. PATHWAYS IN LEC/BEC DETECTED BY CPG-SET SCORE (FORMULA MINMAX, GENE REGION)

Index	Geneset Name	P	FDR
K4	KEGG_HEMATOPOIETIC_CELL_LINEAGE	0.0010	0.0160
K5	KEGG_GRAFT_VERSUS_HOST_DISEASE	0.0020	0.0200
K6	KEGG_SPLICEOSOME	0.0030	0.0210
K7	KEGG_AUTOIMMUNE_THYROID_DISEASE	0.0020	0.0226
K8	KEGG_OXIDATIVE_PHOSPHORYLATION	0.0010	0.0320
K9	KEGG_ALLOGRAFT_REJECTION	0.0040	0.0562
K10	KEGG_PARKINSONS_DISEASE	0.0040	0.0646
K11	KEGG_NUCLEOTIDE_EXCISION_REPAIR	0.0110	0.0738
K12	KEGG_HUNTINGTONS_DISEASE	0.0020	0.0752
K13	KEGG_RNA_DEGRADATION	0.0110	0.0776
K14	KEGG_VIBRIO_CHOLERAE_INFECTION	0.0000	0.0867
K15	KEGG_ALZHEIMERS_DISEASE	0.0010	0.0876
K16	KEGG_BASAL_TRANSCRIPTION_FACTORS	0.0160	0.0932
K17	KEGG_RIBOSOME	0.0100	0.0932
K18	KEGG_PROTEASOME	0.0050	0.0960
K19	KEGG_GLUTATHIONE_METABOLISM	0.0230	0.0998
K20	KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM	0.0070	0.0998
K21	KEGG_HOMOLOGOUS_RECOMBINATION	0.0110	0.1000
K22	KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO_SERIES	0.0060	0.1130
K23	KEGG_PROTEIN_EXPORT	0.0490	0.1204
K24	KEGG_ABC_TRANSPORTERS	0.0190	0.1222
K25	KEGG_TYPE_I_DIABETES_MELLITUS	0.0020	0.1293
K26	KEGG_CELL_ADHESION_MOLECULES_CAMS	0.0270	0.1365
K27	KEGG_ECM_RECEPTOR_INTERACTION	0.0010	0.1514
K28	KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS	0.0020	0.1775
K29	KEGG_FOLATE_BIOSYNTHESIS	0.0160	0.1782
K30	KEGG_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI_ANCHOR_BIOSYNTHESIS	0.0530	0.1798
K31	KEGG_TYROSINE_METABOLISM	0.0420	0.1807
K32	KEGG_SNARE_INTERACTIONS_IN_VESICULAR_TRANSPORT	0.0280	0.1953
K33	KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	0.0050	0.1978
K34	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	0.0660	0.2098
K35	KEGG_VASOPRESSIN_REGULATED_WATER_REABSORPTION	0.0160	0.2118
K36	KEGG_VIRAL_MYOCARDITIS	0.0110	0.2133

Continued on next page

Continued from previous page

Index	Geneset Name	P	FDR
K37	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_HEPARAN_SULFATE	0.0410	0.2441

- Based on Gene Combined Score

Based on the CpG-set scoring, formula MinMax in the TSS region gave one significant pathway and the EXON1 region also gave one significant pathway too. For formula AvG-Exponentiel, only BODY region gave one significant pathway (TABLE XXIV).

	Formula MinMax	Formula AvgExponentiel
TSS Region	1pathway	0 pathways
EXON1 Region	1pathway	0 pathways
BODY Region	TABLE XXV	1pathway
Gene Region	TABLE XXVI	0 pathways

TABLE XXIII

SUMMARY OF RESULTS USING GENE COMBINED SCORE

TABLE XXIV. PATHWAYS IN LEC/BEC DETECTED BY GENE COMBINED-SCORE

Index	Geneset Name	P	FDR
	Formula MinMax TSS region		
K5	KEGG_PPAR_SIGNALING_PATHWAY	0.002	0.249
	Formula MinMAx EXON1 region		
K38	KEGG_RENAL_CELL_CARCINOMA	0.003	0.207
	Formula AvgExponentiel BODY region		
K17	KEGG_RIBOSOME	0.004	0.249

TABLE XXV. PATHWAYS IN LEC/BEC DETECTED BY GENE COMBINED-SCORE

(FORMULA MINMAX BODY REGION)

Index	Geneset Name	P	FDR
K39	KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	0.0010	0.0090
K5	**KEGG_PPAR_SIGNALING_PATHWAY**	0.0030	0.1075
K40	KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	0.0020	0.1076
K41	KEGG_CHRONIC_MYELOID_LEUKEMIA	0.0010	0.1280
K42	KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	0.0030	0.1366
K43	KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	0.0120	0.1470
K44	KEGG_FOCAL_ADHESION	0.0190	0.1499
K45	KEGG_INOSITOL_PHOSPHATE_METABOLISM	0.0090	0.1506
K46	KEGG_ADHERENS_JUNCTION	0.0130	0.1542
K47	KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GLOBO_SERIES	0.0150	0.1544
K48	KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY	0.0100	0.1573
K49	KEGG_NOTCH_SIGNALING_PATHWAY	0.0210	0.1591
K50	KEGG_BLADDER_CANCER	0.0180	0.1592
K51	KEGG_SMALL_CELL_LUNG_CANCER	0.0270	0.1598
K52	KEGG_COLORECTAL_CANCER	0.0110	0.1635
K53	KEGG_APOPTOSIS	0.0140	0.1672
K54	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_HEPARAN_SULFATE	0.0190	0.1723
K55	KEGG_ENDOCYTOSIS	0.0220	0.1751
K56	KEGG_AXON_GUIDANCE	0.0290	0.1785
K57	KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	0.0100	0.1796
K58	KEGG_DILATED_CARDIOMYOPATHY	0.0280	0.1802
K59	KEGG_MAPK_SIGNALING_PATHWAY	0.0170	0.1813
K60	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.0140	0.1853

Continued on next page

Continued from previous page

Index	Geneset Name	P	FDR
K61	KEGG_INSULIN_SIGNALING_PATHWAY	0.0320	0.1854
K62	KEGG_PANCREATIC_CANCER	0.0240	0.1861
K63	KEGG_PATHWAYS_IN_CANCER	0.0190	0.1913
K64	KEGG_PROSTATE_CANCER	0.0390	0.1948
K65	KEGG_DRUG_METABOLISM_OTHER_ENZYMES	0.0010	0.2170
K66	KEGG_GLYOXYLATE_AND_DICARBOXYLATE_METABOLISM	0.0340	0.2432
K67	KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR_CARDIOMYOPATHY_ARVC	0.0440	0.2481

TABLE XXVI. PATHWAYS IN LEC/BEC DETECTED BY GENE COMBINED SCORE (FORMULA MINMAX GENE REGION)

Index	Geneset Name	P	FDR
K68	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_HEPARAN_SULFATE	0.0020	0.0390
K69	KEGG_ENDOCYTOSIS	0.0010	0.0750
K70	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.0040	0.2216
K71	KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_GANGLIO_SERIES	0.0210	0.2366
K44	KEGG_FOCAL_ADHESION	0.0090	0.2431
K5	**KEGG_PPAR_SIGNALING_PATHWAY**	0.0100	0.2444
K72	KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	0.0420	0.2475
K49	KEGG_NOTCH_SIGNALING_PATHWAY	0.0350	0.2479

The involvement of lymphatic and blood vessels in the course of tissue genesis and regeneration in several physiological mechanisms has been reported(33) . In addition, these lymphatic and endothelial cells are tightly related to the progression of many pathological states such as tumor metastasis. Nonetheless, cancer proliferation needs the development of new anomalous blood vessels, which can also cause other types of diseases. Therefore, in some conditions, blood endothelial cells (BEC) can look similar to the lymphatic endothelial cells (LEC) (33).

In the present work, through the statistical framework scoring procedures, we detected several cancer pathways such as renal cell carcinoma (K38), chronic myeloid leukemia (K41), bladder cancer (K50), small cell lung cancer (K51) and colorectal cancer(K52), in addition to

cancer related signaling pathways such as notch pathway (K49) and neurotrorphin pathway (K40). Our results confirm with the literature, indicating that epigenetic processes play an important role in specifying the vascular lineage of endothelial cells and may further be involved in cancer initiation, formation and proliferation.

5.3 Discussion

In this work, we have built a statistical framework in which, we proposed three types of scoring methods (GSS/CSS, CSeS, GCS), to be used in gene set enrichment analysis (GSEA), in order to detect phenotype related pathways that can be investigated for potential therapies and prognosis. The first score is based on gene expression as a single score and same for CpG single-score. The second score, is based on an aggregation of a set of CpGs methyaltion in a specific region. The third one, is based on the combination of the first and the second scores using two mathematical models.

The evaluation of this approach on two datasets generated significant results which demonstrates that the proposed statistical framework could detect relevant pathways involved with the disease. However, the results are very variable and dependent on the scoring type, the methylation region selected, the combination formula used and the dataset specificities.

We selected breast cancer and lymphatic/blood endothelial cells datasets, and we observed a group of gene sets (pathways) that are commonly detected by different scoring procedures through different regions and some are presented in TABLE XXVII and TABLE XXVIII. These pathways have general and specific functions according to KEGG pathways descriptions. Eventually, they belong to functional categories associated with cellular processes, especially with the processing of genetic information and environmental information, in addition to the metabolism of cofactors, vitamins, nucleotides and amino acid. These findings suggest that epigenetic mechanisms are involved in regulating metabolic, developmental and environmental processes of the cancer disease (BC) and vascular system (LEC/BEC).

TABLE XXVII. BREAST CANCER PATHWAYS DESCRIPTION

Index	Geneset function
K1,K7	Genetic Information Processing:Translation
K4,K3,K8,K15	Genetic Information Processing:Folding,sorting and degradation
K12	Genetic Information Processing:Transcription(Spliceosome)
K6,K2	Environmental Information Processing:Signaling molecules and interaction
K11,K16	Cellular Processes:Cell growth and death
K17	Cellular Processes:Cell communication
K5	Organismal Systems:Sensory system
K9	Organismal Systems:Nervous system
K10	Genetic Information Processin:Replication and repair
K25	Genetic Information Processing:Transcription
K22,K18	Human Diseases:Neurodegenerative diseases
K14,K19	Human Diseases:Immune diseases
K20	Metabolism:Metabolism of cofactors and vitamins
K13	Metabolism:Nucleotide metabolism
K21	Metabolism:Amino acid metabolism
K23,K24	Environmental Information Processing:Signal transduction

TABLE XXVIII. LEC/BEC PATHWAYS DESCRIPTION

Index	Geneset Function
K1,K3	Environmental Information Processing:Signaling molecules and interaction
K2	Organismal Systems:Immune system
K4	Metabolism:Amino acid metabolism
K5	Organismal Systems:Endocrine system
K6	Human Diseases:Cancers Specific types
K7	Genetic Information Processing:Translation

The breast cancer dataset had more significant pathways due to many factors. For example, it has a larger number of samples compared to that of the second dataset (50 versus 16 samples). Both datasets agreed on the fact that the scoring based on the gene expression only is less effective and almost gave no significant pathways. However, combining the expression with the CpG methylation showed a significant improvement and detected a large set of pathways depending on the combination of formula and region. This combination was more efficient in

the LEC/BEC dataset than it was in the breast cancer dataset, suggesting that LEC/BEC phenotype is synergetically altered by the gene expression and DNA methylation at the same time. Meanwhile, the breast cancer alteration is more connected to DNA methylation than gene expression.

In the LEC/BEC case, the first combination using the MinMax product of CpG methylation detected more significant pathways and the body region outperformed the other regions, which suggests that the genes bodies methylation is potentially more connected to the gene expression and the phenotype. The observation about the capabilty of the first formula to detect many significant pathways can be considered true also for the breast cancer case, however, when using the CpG-set scoring which is based only on methylation of the genes, the second formula which uses the exponential function to account for all the CpGs methylation values outperfomed the other one, not in the number of the pathways but in their biological meaning and reduced noise.

The effectiveness of formula Avg-Exponential is clear especially in EXON1 region by detecting the breast cancer immediate pathways such as the basal cell carcinoma, which highly suggests the importance of the methylation alteration in the EXON1 region of the breast cancer genes. Finally, the majority of the detected pathways in both datasets are specific metabolic pathways and directly related to the phenotype. However the less specific ones can be considered as noise like Asma, parkinson, huntington disease but they showed only in a couple of tables and mostly less significant which proves the great potential of our proposed framework.

CHAPTER 6

CONCLUSIONS AND PERSPECTIVES

Gene expression dis-regulation and DNA nucleotide mutation such as Single Nucleotide Polymorphisms (SNP), have been extensively studied and used for complex diseases such in cancer diagnosis, prognosis and targeted treatment. Recently, numerous studies considered epigenetic alterations, such as DNA methylation, as another factor that can distinguish between phenotypes and contribute to targeted treatment.

However, it is still unclear how methylation is affecting the gene expression, by only up-regulation or down-regulation or both, and to which extend. Several studies conducted successful integrated analysis of genomic features such as gene expression and SNPs to decipher the inherent relation between both of them, but very few studies were as successful as those ones, when it comes to gene expression and DNA mehtylation integration.

In this work, we proposed a statistical framework to assess the integration of gene expression and DNA CpGs methylation in a gene-set enrichment analysis. Through this approach, we rank the genes according to their differential expression test-statistic, such as t-test, called also gene single-score and for every ranked position we determine an enrichment score for every gene set in KEGG pathways. We assess the significance of detected pathways by computing Pvalue and FDR from the null distribution generated through permutations between phenotypes. The same procedure is repeated by ranking the genes according to their CpG-set score in different regions then again according to a combination of the gene single-score and CpGs-set score to-

gether. Consequently, we developed two mathematical models to combine the scores.

We tested our framework on breast cancer and lymphatic/blood endothelial cells datasets. Our results, showed that depending on the region the gene combined score can perform better then using gene expression only to detect pathways significantly correlated to phenotype in study. However, in most regions CpG-set score showed more pathways that are numerically and biologically more significant which shows the promising power of DNA methylation alterations in complex diseases.

In a future work, we intend to use larger datasets and pathways databases. Also, we urgently need to elaborate a simulation study to assess the sensitivity and precision of our statistical framework. Furthermore, different test statistics can be used for gene and CpGs combined scores, for example, by using Stouffer or Fisher combined Pvalues. Finally, more investigation is needed to improve our mathematical models to increase the significance and reduce and the noise.

APPENDICES

TABLE XXIX. TCGA BREAST CANCER CASE SAMPLES USED FOR DNA CPG-METHYLATION EXPERIMENT

Index	**Case Sample Barcodes**	**Chip_position**
1	TCGA_BH_A0BC_01A_22D_A10P_05	6004791006_R04C01
2	TCGA_BH_A0DK_01A_21D_A10P_05	6005486019_R02C01
3	TCGA_BH_A0H7_01A_13D_A10P_05	6005486021_R05C01
4	TCGA_BH_A0BA_01A_11D_A10P_05	6005486021_R06C02
5	TCGA_BH_A0B3_01A_11D_A10P_05	6005486021_R03C01
6	TCGA_BH_A0BJ_01A_11D_A10P_05	6005486014_R05C01
7	TCGA_BH_A0DP_01A_21D_A10P_05	6005486021_R04C02
8	TCGA_BH_A0E0_01A_11D_A10P_05	6005486019_R06C01
9	TCGA_BH_A0E1_01A_11D_A10P_05	6005486014_R03C02
10	TCGA_BH_A0HK_01A_11D_A10P_05	6005486019_R01C01
11	TCGA_BH_A0C0_01A_21D_A10P_05	6005486023_R06C01
12	TCGA_BH_A0B8_01A_21D_A10P_05	6005486019_R04C02
13	TCGA_BH_A0BM_01A_11D_A10P_05	6005486023_R02C02
14	TCGA_BH_A0H9_01A_11D_A10P_05	6005486021_R01C01
15	TCGA_BH_A0DQ_01A_11D_A10P_05	6004791004_R04C01
16	TCGA_BH_A0DH_01A_11D_A10P_05	6004791010_R06C02
17	TCGA_BH_A0B2_01A_11D_A10N_05	6005486012_R06C01
18	TCGA_BH_A1EO_01A_11D_A138_05	6042324048_R06C01
19	TCGA_BH_A1F0_01A_11D_A138_05	6042324071_R06C02
20	TCGA_BH_A1EW_01A_11D_A138_05	6042324048_R05C02
21	TCGA_BH_A1ET_01A_11D_A138_05	6042324071_R04C01
22	TCGA_BH_A1EU_01A_11D_A138_05	6042324071_R02C02
23	TCGA_BH_A0HA_01A_11D_A12R_05	6042324070_R06C01
24	TCGA_BH_A0C3_01A_21D_A12R_05	6042324030_R06C02
25	TCGA_BH_A0AU_01A_11D_A12R_05	6042324030_R03C01
26	TCGA_BH_A0BZ_01A_31D_A12R_05	6042324030_R01C01
27	TCGA_BH_A0BS_01A_11D_A12R_05	6042324030_R02C02

TABLE XXX. TCGA BREAST CANCER CONTROL SAMPLES USED FOR DNA CPG-METHYLATION EXPERIMENT

Index	**Control Sample Barcodes**	**Chip_position**
1	TCGA_BH_A0BC_11A_22D_A093_05	6005486025_R01C01
2	TCGA_BH_A0DK_11A_13D_A10Q_05	6005486013_R06C01
3	TCGA_BH_A0H7_11A_13D_A10Q_05	6005486013_R05C02
4	TCGA_BH_A0BA_11A_22D_A10Q_05	6005486011_R01C01
5	TCGA_BH_A0B3_11B_21D_A10Q_05	6005486011_R03C01
6	TCGA_BH_A0BJ_11A_23D_A10Q_05	6005486011_R04C01
7	TCGA_BH_A0DP_11A_12D_A10Q_05	6005486011_R05C01
8	TCGA_BH_A0E0_11A_13D_A10Q_05	6005486011_R06C01
9	TCGA_BH_A0E1_11A_13D_A10Q_05	6005486011_R01C02
10	TCGA_BH_A0HK_11A_11D_A10Q_05	6005486011_R04C02
11	TCGA_BH_A0C0_11A_21D_A10Q_05	6005486011_R05C02
12	TCGA_BH_A0B8_11A_41D_A093_05	6005486025_R02C01
13	TCGA_BH_A0BM_11A_12D_A093_05	6005486025_R03C01
14	TCGA_BH_A0H9_11A_22D_A093_05	6005486025_R04C01
15	TCGA_BH_A0DQ_11A_12D_A10Q_05	6005486013_R04C02
16	TCGA_BH_A0DH_11A_31D_A10Q_05	6005486013_R06C02
17	TCGA_BH_A0B2_11A_11D_A10N_05	6005486012_R03C02
18	TCGA_BH_A1EO_11A_31D_A138_05	6042324048_R05C01
19	TCGA_BH_A1F0_11B_23D_A138_05	6042324048_R03C02
20	TCGA_BH_A1EW_11B_33D_A138_05	6042324048_R06C02
21	TCGA_BH_A1ET_11B_23D_A138_05	6042324071_R05C01
22	TCGA_BH_A1EU_11A_23D_A138_05	6042324071_R05C02
23	TCGA_BH_A0HA_11A_31D_A12R_05	6042324035_R01C01
24	TCGA_BH_A0C3_11A_23D_A12R_05	6042324070_R02C01
25	TCGA_BH_A0AU_11A_11D_A12R_05	6042324072_R04C02
26	TCGA_BH_A0BZ_11A_61D_A12R_05	6042324037_R05C01
27	TCGA_BH_A0BS_11A_11D_A12R_05	6042324072_R02C02

TABLE XXXI. TCGA BREAST CANCER CASE AND CONTROL SAMPLES USED FOR GENE EXPRESSION EXPERIMENT

Index	**Case Sample Barcodes**	**Control Sample Barcodes**
1	TCGA_BH_A0BC_01A_22R_A084_07	TCGA_BH_A0BC_11A_22R_A089_07
2	TCGA_BH_A0DK_01A_21R_A056_07	TCGA_BH_A0DK_11A_13R_A089_07
3	TCGA_BH_A0H7_01A_13R_A056_07	TCGA_BH_A0H7_11A_13R_A089_07
4	TCGA_BH_A0BA_01A_11R_A056_07	TCGA_BH_A0BA_11A_22R_A089_07
5	TCGA_BH_A0B3_01A_11R_A056_07	TCGA_BH_A0B3_11B_21R_A089_07
6	TCGA_BH_A0BJ_01A_11R_A056_07	TCGA_BH_A0BJ_11A_23R_A089_07
7	TCGA_BH_A0DP_01A_21R_A056_07	TCGA_BH_A0DP_11A_12R_A089_07
8	TCGA_BH_A0E0_01A_11R_A056_07	TCGA_BH_A0E0_11A_13R_A089_07
9	TCGA_BH_A0E1_01A_11R_A056_07	TCGA_BH_A0E1_11A_13R_A089_07
10	TCGA_BH_A0HK_01A_11R_A056_07	TCGA_BH_A0HK_11A_11R_A089_07
11	TCGA_BH_A0C0_01A_21R_A056_07	TCGA_BH_A0C0_11A_21R_A089_07
12	TCGA_BH_A0B8_01A_21R_A056_07	TCGA_BH_A0B8_11A_41R_A089_07
13	TCGA_BH_A0BM_01A_11R_A056_07	TCGA_BH_A0BM_11A_12R_A089_07
14	TCGA_BH_A0H9_01A_11R_A056_07	TCGA_BH_A0H9_11A_22R_A089_07
15	TCGA_BH_A0DQ_01A_11R_A084_07	TCGA_BH_A0DQ_11A_12R_A089_07
16	TCGA_BH_A0DH_01A_11R_A084_07	TCGA_BH_A0DH_11A_31R_A089_07
17	TCGA_BH_A0B2_01A_11R_A10J_07	TCGA_BH_A0B2_11A_11R_A10J_07
18	TCGA_BH_A1EO_01A_11R_A137_07	TCGA_BH_A1EO_11A_31R_A137_07
19	TCGA_BH_A1F0_01A_11R_A137_07	TCGA_BH_A1F0_11B_23R_A137_07
20	TCGA_BH_A1EW_01A_11R_A137_07	TCGA_BH_A1EW_11B_33R_A137_07
21	TCGA_BH_A1ET_01A_11R_A137_07	TCGA_BH_A1ET_11B_23R_A137_07
22	TCGA_BH_A1EU_01A_11R_A137_07	TCGA_BH_A1EU_11A_23R_A137_07
23	TCGA_BH_A0HA_01A_11R_A12P_07	TCGA_BH_A0HA_11A_31R_A12P_07
24	TCGA_BH_A0C3_01A_21R_A12P_07	TCGA_BH_A0C3_11A_23R_A12P_07
25	TCGA_BH_A0AU_01A_11R_A12P_07	TCGA_BH_A0AU_11A_11R_A12P_07
26	TCGA_BH_A0BZ_01A_31R_A12P_07	TCGA_BH_A0BZ_11A_61R_A12P_07
27	TCGA_BH_A0BS_01A_11R_A12P_07	TCGA_BH_A0BS_11A_11R_A12P_07

CITED LITERATURE

1. Hanahan D, and Weinberg RA: the hallmarks of cancer. *Cell 2000*, pp. ;100:57-70,2000.

2. James G, Herman, and Stephan B.Baylin: Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N ENGL J MED*, pp 349;21, nov20,2003.

3. Rainusso, N., Man, T.-K., Lau, C. C., Hicks, J., Shen, J. J., Yu, A., Rosen, J. M: Identification and gene expression profiling of tumor-initiating cells isolated from human osteosarcoma cell lines in an orthotopic mouse model. *Cancer Biology and Therapy*, 278287, 2011.

4. Liu, G., Yuan, X., Zeng, Z., Tunici, P., Ng, H., Abdulkadir, I. R., Yu, J. S.: Analysis of gene expression and chemoresistance of CD133+ cancer stem cells in glioblastoma. *Molecular cancer*, 5, 67,2006.

5. Wang, Z., Liu, Y., Mori, M., and Kulesz-Martin, M.: Gene expression profiling of initiated epidermal cells with benign or malignant tumor fates. *Carcinogenesis*, 23(4), 63543,2002.

6. Desai, K. V, Xiao, N., Wang, W., Gangi, L., Greene, J., Powell, J. I., Green, J. E.: Initiating oncogenic event determines gene-expression patterns of human breast

cancer models. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 696772, 2002.

7. Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J., and Ramsahoye, B.: Severe Global DNA Hypomethylation Blocks Differentiation and Induces Histone Hyperacetylation in Embryonic Stem Cells. 24(20), 88628871, 2002.

8. Manuscript, A.: methylation in the developing hippocampus of mouse fetal brains. 20(1), 4349, 2006.

9. Carlone, D. L., Lee, J., Young, S. R. L., Dobrota, E., Butler, J. S., Ruiz, J., and Skalnik, D. G.: Reduced Genomic Cytosine Methylation and Defective Cellular Differentiation in Embryonic Stem Cells Lacking CpG Binding Protein, 25(12), 48814891,2005.

10. Takahashi, K., and Yamanaka, S.: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4), 66376, 2006.

11. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1), 2627,2000.

12. Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Botstein, D.: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14), 841823, 2003.

13. Slonim, D. K.: From patterns to pathways: gene expression data analysis comes of age. Nature Genetics, 5028, 2002.

14. Hu, Y., Swerdlow, S., Duffy, T. M., Weinmann, R., Lee, F. Y., Li, S.: Targeting multiple kinase pathways in leukemic progenitors and stem cells is essential for improved treatment of Ph leukemia in mice. *Proceedings of the National Academy of Sciences of the United States of America*, 103(45), 168705, 2002.

15. , F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., and Scandura, J. M.: DNA methylation of the first exon is tightly linked to transcriptional silencing. *PloS one*, 6(1), 2002.

16. Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Church, G. M.: Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology*, 27(4), 3618, 2009.

17. Aran, D., Toperoff, G., Rosenberg, M., and Hellman, A.: Replication timing-related and gene body-specific methylation of active human genes. *Human molecular genetics*,

20(4), 67080, 2011.

18. Gene Set EnrichmentAnalysis, *GSEA user guide*, The broad Institue, 2009-2010.

19. http://www.hmdb.ca/metabolites/HMDB00251.

20. Jin W1, Chen L, Chen Y, Xu SG, Di GH, Yin WJ, Wu J, Shao ZM.:UHRF1 is associated with epigenetic silencing of BRCA1 in sporadic breast cancer. *Breast Cancer Res Treat.* Sep;123(2):359-73, 2010.

21. Lin TSAI et al, Metabolomic Dynamic Analysis of Hypoxia in MDA-MB-231 and the Comparison with Inferred Metabolites from Transcriptomics Data, *Cancers* , 5, 491-510, 2013.

22. Mathilde Bayet-Robert et al.: Biochemical disorders induced by cytotoxic marine natural products in breast cancer cells as revealed by proton NMR spectroscopy-based metabolomics *Biochemical Pharmacology*, Volume 80, Issue 8, Pages 11701179,2010.

23. http://www.yalemedicalgroup.org/circadiangene

24. Tanja Kunej et al.: Epigenetic regulation of microRNAs in cancer:An integrated review of literature, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 717, 77-84, 2011.

25. http://breast-cancer-research.com/1755-8794/4/74/table/T5

26. http://www.itb.cnr.it/breastcancer/php/KOTree2.php?idKO=01110

27. Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S.: Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Research* 386397, 2012.

28. Louhimo, R., and Hautaniemi, S.: CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics (Oxford, England)*, 27(6), 8878, 2011.

29. Li, M., Balch, C., Montgomery, J. S., Jeong, M., Chung, J. H., Yan, P., Nephew, K. P.: Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC medical genomics*, 2, 34, 2009.

30. Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R., Thompson, E. A.: Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PloS one*, 2011.

31. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., and Ebert, B. L.: Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. 2005.

32. http://cancergenome.nih.gov/

33. Brnneke, S., Bruckner, B., Peters, N., Bosch, T. C. G., Stab, F., Wenck, H., Winnefeld, M.: DNA methylation regulates lineage-specifying genes in primary lymphatic and blood endothelial cells. *Angiogenesis*, 15(2), 31729, 2012.

34. Bibikova, M., Le, J., Barnes, B., Saedinia-melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L.: Genome-wide DNA methylation profiling using, Technology Report, 177200.

35. Dedeurwaerder, S., Defrance, M., & Calonne, E.: Evaluation of the Infinium Methylation 450K technology, Technology Report, 771784.

36. Touleimat, N. (2012). T echnology R eport Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation T echnology R eport, 4, 325341.

37. Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S.: A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* (Oxford, England), 29(2), 18996, 2013.

38. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:6671.

39. Venkitaraman AR. Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell 2002;108:17182.

40. Keniry M, Parsons R. The role of PTEN signaling perturbations in cancer and in targeted therapy. Oncogene 2008;27:547785.

41. Zhu, J., and Yao, X. (2007). Use of DNA methylation for cancer detection and molecular classification. Journal of biochemistry and molecular biology, 40(2), 13541.

42. Golub, T. R. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439), 531537.

43. Wang, D. G. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science, 280(5366), 10771082.

44. Lu, J., Getz, G., Miska, E. a, Alvarez-Saavedra, E., Lamb, J., Peck, D., Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. Nature, 435(7043), 8348.

45. http://res.illumina.com

46. Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics, 39(4), 45766.

47. Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America, 103(5), 14127.

48. Belin, S., Beghin, A., Solano-Gonzlez, E., Bezin, L., Brunet-Manquat, S., Textoris, J., Diaz, J.-J. (2009). Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. PloS One, 4(9), e7147.

49. Huan, J., Wang, L., Xing, L., Qin, X., Feng, L., Pan, X., and Zhu, L. (2014). Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17beta -estradiol (E2). Gene, 533(1), 34655.

50. Tani, H., Torimura, M., and Akimitsu, N. (2013). The RNA degradation pathway regulates the function of GAS5 a non-coding RNA in mammalian cells. PloS One, 8(1), e55684.

51. Rossi, S., and Loda, M. (2014). The Role of the Ubiquitination Proteasome Pathway in Breast Cancer : Use of Mouse Models for Analyzing Ubiquitination Processes

52. Ohta, T., and Fukuda, M. (2004). Ubiquitin and breast cancer. Oncogene, 23(11), 207988.

53. Orlowski, R. Z., and Dees, E. C. (2003). The role of the ubiquitination-proteasome pathway in breast cancer Applying drugs that affect the ubiquitin-proteasome pathway to the therapy of breast cancer, 17.

54. Choi, Y., Hur, C.-G., and Park, T. (2013). Induction of olfaction and cancer-related genes in mice fed a high-fat diet as assessed through the mode-of-action by network identification analysis. PloS One, 8(3), e56610.

55. Muranen, T. a, Greco, D., Fagerholm, R., Kilpivaara, O., Kmpjrvi, K., Aittomaki, K., Nevanlinna, H. (2011). Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. Breast Cancer Research : BCR, 13(5), R90.

56. Kim, Y.-W., Kwon, C., Liu, J.-L., Kim, S. H., et Kim, S. (2012). Cancer association study of aminoacyl-tRNA synthetase signaling network in glioblastoma. PloS One, 7(8), e40960.

57. http://www.itb.cnr.it

58. Manuscript, A., et Physiology, C. (2013). NIH Public Access, 91(1), 119149.

59. http://www.cgfr.co.uk/

60. Descamps, S., Toillon, R. a, Adriaenssens, E., Pawlowski, V., Cool, S. M., Nurcombe, V., Hondermarck, H. (2001). Nerve growth factor stimulates proliferation and survival of human breast cancer cells through two distinct signaling pathways. The Journal of Biological Chemistry, 276(21), 1786470.

61. Romon, R., Adriaenssens, E., Lagadec, C., Germain, E., Hondermarck, H., et Le Bourhis, X. (2010). Nerve growth factor promotes breast cancer angiogenesis by activating multiple pathways. Molecular Cancer, 9, 157.

62. Fu, Y., Yu, J., et Cheng, T. (2003). Breast Cancer Risk Associated with Genotypic Polymorphism of the Nonhomologous End-Joining Genes : A Multigenic Study on Cancer Susceptibility Breast Cancer Risk Associated with Genotypic Polymorphism of the Nonhomologous End-Joining Genes : A Multigenic Study on, 24402446.

63. Liu, J., Campen, A., Huang, S., Peng, S.-B., Ye, X., Palakal, M., Li, S. (2008). Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. BMC Medical Genomics, 1, 39.

64. Quidville, V., Alsafadi, S., Goubar, A., Commo, F., Scott, V., Pioche-Durieu, C., Andr, F. (2013). Targeting the deregulated spliceosome core machinery in cancer cells triggers mTOR blockade and autophagy. Cancer Research, 73(7), 224758.

65. http://cbcrp.org.127.seekdotnet.com/research/PageGrant.asp

66. Allegras, C. J., Fine, R. L., Drake, J. C., et Chabner, A. (1986). The Effect of Methotrexate on Intracellular Folate Pools in Human MCF-7 Breast Cancer Cells, 261(14), 64786485.

67. Schramm, G., Surmann, E.-M., Wiesberg, S., Oswald, M., Reinelt, G., Eils, R., et Konig, R. (2010). Analyzing the regulation of metabolic pathways in human breast cancer. BMC Medical Genomics, 3, 39. ubiquitin mediated proteolysis

68. Davis L. Porter et Al.(1994). induction of graft versus host disease as immunotherapy for relapsed chronic myleoid leukemia. The new England Journal of medicine.

69. Build, H., et Road, L. (2002). REVIEW The proteasome: a novel target for cancer chemotherapy, 433443.

70. Zhang, Y., Newcomb, P. a, Egan, K. M., Titus-Ernstoff, L., Chanock, S., Welch, R., Garcia-Closas, M. (2006). Genetic polymorphisms in base-excision repair pathway genes and risk of breast cancer. Cancer Epidemiology, Biomarkers and Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 15(2), 3538.

71. Farmer, H., McCabe, N., Lord, C. J., Tutt, A. N. J., Johnson, D. a, Richardson, T. B., Ashworth, A. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature, 434(7035), 91721.

72. http://pubs.rsc.org/en/content/articlehtml/2009/an/b907243h

73. Plathow, C., and Weber, W. a. (2008). Tumor cell metabolism imaging. Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine, 49 Suppl 2, 43S63S.

74. Moraes, S., Montemor, J., and Silva, A. (2008). Clinical Update Interaction between Specialties : Dilated Cardiomyopathy and HER2- Positive Breast Cancer, 1115.

75. Crone, S. a, Zhao, Y.-Y., Fan, L., Gu, Y., Minamisawa, S., Liu, Y., Lee, K.-F. (2002). ErbB2 is essential in the prevention of dilated cardiomyopathy. Nature Medicine, 8(5), 45965.

76. Tort, F., Zieger, K., Guldberg, P., Sehested, M., Bartkova, J., Hor, Z. (2005). DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis, 434(April).

77. Rodriguez-acebes, S., Proctor, I., Loddo, M., Wollenschlaeger, A., Rashid, M., Falzon, M., Williams, G. H. (2010). Targeting DNA Replication before it Starts Cdc7 as a Therapeutic Target in p53-Mutant Breast Cancers. The American Journal of Pathology, 177(4), 20342045.

78. Koo, H., Vanbrocklin, M., Mcwilliams, M. J., Leppla, S. H., Duesbery, N. S., & Woude, G. F. Vande. (2001). Apoptosis and melanogenesis in human melanoma cells induced by anthrax lethal factor inactivation of mitogen-activated protein kinase kinase.

79. Li, Y., Welm, B., Podsypanina, K., Huang, S., Chamorro, M., Zhang, X., Varmus, H. E. (2003). Evidence that transgenes encoding components of the Wnt signaling pathway preferentially induce mammary cancers from progenitor cells, 100(26), 1585315858.

80. Chi, J., Rodriguez, E. H., Wang, Z., Nuyten, D. S. A., Mukherjee, S., Rijn, M. Van De, Brown, P. O. (2007). Gene Expression Programs of Human Smooth Muscle Cells : Tissue-Specific Differentiation and Prognostic Significance in Breast Cancers, 3(9).

81. Manuscript, A., and Kinase, F. A. (2011). NIH Public Access, 53(3), 901912.

82. Gordon, J. A. R., Sodek, J., Hunter, G. K., and Goldberg, H. A. (2009). Cellular Biochemistry, 1128(June), 11181128.

83. Hui, M., Cazet, A., Nair, R., Watkins, D. N., Toole, S. A. O., and Swarbrick, A. (2013). The Hedgehog signalling pathway in breast development , carcinogenesis and cancer therapy.

84. Sasisekharan, R., Shriver, Z., Venkataraman, G., & Narayanasami, U. (2002). ROLES OF HEPARAN-SULPHATE GLYCOSAMINOGLYCANS IN CANCER, 2(July), 18.

85. Mitropoulou, T. N., Theocharis, A. D., and Nikitovic, D. (2004). IGF-I affects glycosaminoglycan / proteoglycan synthesis in breast cancer cells through tyrosine kinase-dependent and independent pathways, 86, 251259.

VITA

Amira Kefi

EDUCATION

Master of Bioinfomatics: Fulbright scholar fall2011-Fall2014
University of Illinois at Chicago, Bioengineering Department
Concentration: Functional Genomics, Microarrays and NGS Data Analysis

SSW08 Certification
Summer school on WEB semantics, Madrid, Spain, July 2008

Master of Computer Science
National school of computer sciences, la Manouba, Tunisia, December 2007
Concentration: Model driven Engineering of Object-oriented systems
Minor: Artificial Intelligence

Bachelor of Computer Science applied to Management
Higher Institute of Management, le Bardo, Tunisia, June 2005
Concentration: Business Analysis and Design

Music Diploma
Ministry of culture, conservatoire H-lif, Tunisia, June 2000
Concentration: Oriental Music

COMPUTER SKILLS

Programming Languages: R, MATLAB, Java, C, Pascal.
Web designing: HTML, Javascript, EasyPHP/PHP.
Databases: MSAccess, MySQL, Oracle.
Operating Systems: Linux, Windows.
Development IDEs: Eclipse.

BIOinformatic SKILLS

Microarrays data Analysis
Gene expression, MicroRNA : Affymetrix and Agilent platforms
DNA CpG-methylation : Illumina Infinium HumanMethylation27k and 450k

Next Generation sequencing Analysis
RNAseq tools: Bowtie, Tophat.

Pathway Analysis: GSEA, GO term analysis.

PROFESSIONAL EXPERIENCE

Research Assitant internship February 2014-August 2014
Reaserch center of Ann and Robert Lurie's childrens hospital of Chicago

- Statistical analysis of Genomic and Epigenomic data (Hapmap, Encode, TCGA and GEO DATA)

Thesis Research work Fall 2012-Fall 2013
BioEngineering Department, University of Illinois at chicago
- Statistical framework and pipeline to integrate Gene expression and DNA methylation in pathways analysis of Breast Cancer using the Cancer Genome Atlas Data.

Business Analyst internship Spring 2005
Cars insurance-system Re-engineering, STAR company, Tunis, Tunisia
- Analysis, Design and Development of a prototype for migrating the legacy system which is in Cobol files to relational database,web-oriented and object-oriented system using Java/Oracle/Linux.

Programmer internship Summer 2003
STEG electricity company, Tunis, Tunisia
- Development of Application to automatically generate payment coupons using VB/SQL-Server

TEACHING EXPERIENCE

Teaching Assistant Fall 2009 - Fall 2010
Computer science Department, ESTI University, la Charguia, Tunsia
- Teached system programming C/Linux.

High school Teacher Fall 2006 - Spring 2011
Ministry of Education and Training, Tunis, Tunisia
- Teached Algorithm (PASCAL), DataBases (Microsoft Access, mySQL), Opertating Systems and networks