Master of Science in Omics Data Analysis

Master Thesis

Epigenomic data integration for characterization of promoter regions

by

Sandra Garcia Mulero

Supervisor: Emanuele Raineri, Statistical Genomics Group, CNAG-CRG

Co-supervisor: Simon Heath, Statistical Genomics Group, CNAG-CRG

Department of Systems Biology

University of Vic - Central University of Catalonia

September 19, 2016

Acknowledgements

Foremost, I would like to express my sincere gratitude to my master thesis supervisor, Dr Emanuele Raineri, for having supported me and guided my work during these months, for his consistent help and his time. Also for giving me the opportunity of opening the project to more complex analysis and teaching me new and interesting approaches.

I would also like to thank all my colleagues at the Statistical Genomics group at CNAG-CRG for the interesting meetings and open discussions I have been invited to participate and contribute, and specially Dra Angelika Merkel, for her helpful tips and opinions.

To my colleague Iago Maceda for his help and advices with the programming, and to Ana Karina, Beatriz Miguel and Manu Ferrando for their support when difficulties and the good moments afterwork.

And last but not least, to all the master's teachers and collaborators, specially the coordinator Dra Malu Calle, for gently having given us the knowledge and tools to start a career in the exciting world of bioinformatics.

Abstract

Understanding the regulatory machinery is one of the current challenges for research in cancer epigenomics. Regulation of gene expression is a complex process with many components involved in it such as histone modifications, DNA methylation and chromatin accessibility, among others. In this project we are going to explore the Blueprint database of human hematopoietic cells in order to characterize regulatory regions on samples from this dataset. For this purpose we are going to carry out a number of steps to analyse the DNA methylation, epigenomic features and RNA-seq.

The characterization of methylation across 92 samples of 12 different cell types have determined that methylation state at promoters and CpG islands are cell type-specific. Exploration and integration of epigenomics in one sample of cell type monocyte was carried out. We have seen the implications of two marks studied (histone modification H3K4me3 and open chromatin state) in gene expression. This characterization of promoters functionality enabled us to create a pipeline for the identification of novel regulatory regions based in the integration of epigenomic features and RNA-seq data, which we later reproduced in four samples. This project could be a start point of a bigger project for the cell type-specific promoters discovery within the Blueprint projects.

Contents

List of Tables

List of Figures

1 Introduction

As far as we know, DNA is static and does not change among the different cells of an organism, whereas chromatin is highly dynamic and changes in response to signalling systems and environmental stimulus. Chromatin dynamics govern the diverse response to such stimulations through the addition of chemical tags by specialized enzymes [2]. The study of the epigenome consists in the study of those chemical tags on the DNA and histone proteins in the nucleus, and its changes in different conditions, tissues and cell types.

The epigenome is involved in embryonic development, tissue differentiation, and regulation of the gene expression in tissues [3]. It has also been shown its strong relation with neurological and autoimmune diseases, cancer development and metastasis, as reported in different studies [4,5].

Promoters are the regions that initiate the transcription of a gene and are located at the 5′ end of the gene, surrounding the Transcription Start Site (TSS) of the gene. Promoters are formed of two parts, the core promoter, which is found in the upstream aside the gene and is approximately 250 base pairs long and where transcription factors (TFs) are bounded, and the distal promoter, which can go up to 1000 to 5000 base pairs upstream from the gene. Most of the genes have more than one TSS that control the transcription of its gene. The regulatory machinery carries out the initiation of transcription at promoters by recruitment of the RNA polymerase II [6].

Although there is a lot of knowledge about promoters, in recent papers it has been reported in mouse and some human tissues that there could be unannotated regulatory regions which would be cell type-specific [1, 7–9]. This means that there could be up-regulated transcripts on some cell types which transcription initiation is carried out by specific regulatory elements that will not have this functionality on other cell types.

The exploration of those cell type-specific regulatory regions would be interesting in order to understand the possible functionalities of these regions; which genes are they regulating, how far are they from the transcripts, which cell types show a higher density of specific promoters, etc. Those regions could be intragenic or intergenic, and could intersect CpG islands, repetitive domains or other elements.

This information could be of high interest for the epigenomics community because it may lead to insights in how cell type-specific regulatory elements lead to differences in the phenotypes, and further investigation on how do these regulatory regions behave in blood diseases. In this project we start a first exploration of this problem, by integrating datasets from the Blueprint consortium.

1.1 The Human Epigenome

Each cell in our body has the same information, although each cell type have their own regulation that alters the gene expression. Chromatin dynamics determines the mechanisms of cell type-specific transcriptomics after stimulus [10]. For many years, regulatory elements have been deeply studied and annotated.

One of the biggest projects created for the understanding and annotation of regulatory elements is the ENCODE Project [4] (ENCyclopedia Of DNA Elements), which has created a catalogue of DNA regulatory elements, including promoters, enhancers, silencers, and transcription factors binding sites, among others.

Gene promoters, as well as gene-distal regulatory elements (enhancers), control the gene transcription by the interplay of regulatory elements (**Figure1.1**). It is well known and described that promoters

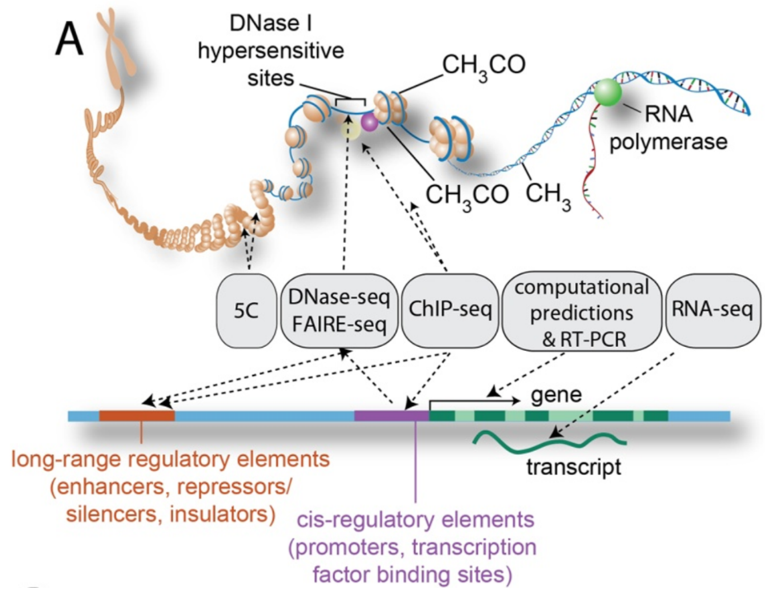


Figure 1.1: Regulatory regions and elements that interact in the regulation of gene expression. Image credits from the ENCODE Consortium [4]

find a number of epigenomic features that are shared mostly in all of them and which are well-known indicators of promoter function [6, 11, 12]. These features are histones modifications (HM), DNA methylation DNase hypersensitivity and and transcription factors binding sites.

The combination of the epigenomic features along with the transcriptomics quantification provides a complete blueprint of the regulatory machinery and helps understand the downstream effect of those regions. For this reason, four main marks are going to be deeply studied in this project (**Table 1.1**).

Table 1.1: Description of abbreviations for the features to study in the project.

Feature	**Experimental method ***	**Description**
DNA Methylation	WGBS	Addition (or lack) of the methyl group in the 5C residue of cytosine by DNA methyltransferase
Histone modification (HM)	ChIP-seq	Post-translational modifications of nucleosomal histones consisting mostly in the methylation and acetylation of its amino acids
Chromatin Accessibility	DNAse-seq	Regions of open chromatin (DNase I hypersensitivity regions)
Transcriptomics	RNA-seq	Depth of RNA reads for each position in a given region

* (Detailed in BOX 1)

1.1.1. DNA Methylation

Methylation of the DNA is the process of addition of a methyl group to the C5 carbon residue of cytosine by DNA methyltransferase (DNM T3L) [13]. This process is more likely to happen in CpGs, and can occur in both DNA strands to maintain methylation at CpGs during replication. On the other side, most cytosines are non-methylated in differentiated mammalian tissues.

Regulatory regions are mostly characterized by a state of hypomethylation. Thus, in the promoter region of an active gene we usually expect to see a pattern of hypomethylation, whereas in the promoter of a repressed gene we will expect it to be highly methylated [14].

CpG islands

Many promoters coincide with regions of high CpG content. These regions are called CpG islands (CGIs) and are regions of 300-3000 pb long with a proportion of CG higher than 50% and a long proportion of CpG sites (about 60 %).

CpG islands at promoter regions are usually highly unmethylated when the gene is expressed [15]. Promoters can be divided in 'high CpG-content promoters (HCPs)' or 'low CpG-content promoters (LCPs)'. HCPs are usually regulating active genes, whereas LCPs are inactive by default. Both promoters have similarities in other features, which means that other features such as presence of HM or DNA methylation are more precise in classifying high or low activity at promoters [12].

1.1.2. Histone modification

Eukaryotic chromatin structure DNA is based in a series of organized layers. DNA is wrapped in a region of 147 bp by nucleosomes, which is formed by a histone octamer, and forming the primary structure of chromatin in mammalian genomes [12]. In the nucleosome, histones H2A, H2B, H3 and H4 can be modified by post-translational modification.

There are more than 100 post-translational possible HM [16], and those distinct marks act in combinatorial way to change the conformation of chromatin, interacting with the transcription machinery and facilitating or repressing transcription. The HM are predictive of gene expression and have been used to annotate regulatory regions through predictive models and unsupervised methods, for instance, the hidden Markov Models [17,35].

The most useful HM for identification of functional elements are methylation of H3K4 and acetylation of H3K27. Histones modifications H3K4me3 and H3K36me3 are associated with active genes. The most predictive of promoter regions is the methylation mark H3K4me3 (histone 3 lysone 4 trimethylation), which is well known associated with the TSS regions of transcribed genes [11]. H3K36me3 might be well informative of transcription as well, occurring along the gene body. Other HM related to gene activation are H3K27ac, H3K9me1, H3K27me1, H4K20me1 and H2BAc. On the other hand, HM H3K9me3 and H3K27me3 are markers of inactive transcription [12].

Histone mark H3K4me1 is usually associated with enhancers, which are helpers of promoters for the gene expression. Enhancers are highly cell type-specific and are located far away from the target genes. Both promoters and enhancers can be flanked by nucleosomes with the modifications H3K4me3 and H3K4me1. Usually, the ratio of those marks is used to differentiate promoters and enhancers, whether H3K4me3 or H3K4me1 is higher, respectively [19]. Although, high peaks the promoter-specific mark H3K4me3 has been found in active enhancers, leading to ambiguity around this marker. Therefore, the distinction between enhancers and promoters can be difficult, and new models have de-constructed such distinctions and created a unifying model for both regulatory elements [18].

1.1.3. Chromatin accessibility

Chromatin accessibility is a marker of regulatory regions in the DNA. These are regions where chromatin is altered resulting in hypersensitivity to cleavage by the DNAse I nuclease enzyme [4], thus the DNA is exposed and accessible. DNase I is the endonuclease that catalyzes the hydrolytic cleavage of DNA in order to digest the double-stranded DNA.

As in the case of HM H3K4me3 peaks, the presence of DNase hypersensitivity sites is a good method for identifying the location of regulatory elements, such as promoters, enhancers, silencers, and replication origins [21]. Chromatin accessibility is highly related to gene expression and there are evidences that is cell type-specific [20].

BOX 1 | *Experimental methods*

ChIP-seq (Chromatin immunoprecipitation). *This method is based in selection of DNA chromatin complexes by using antibodies to specific epitopes. After, the sample is sequenced by high-throughput technologies to determine the regions in the genome most often bound by the protein. There are different antibodies used (for transcription factors, chromatin binding proteins, histone proteins, etc) [22].*

***DNAse-seq*.** *Employs DNase I enzyme to cut live chromatin preparations at sites where there are specific proteins. Those cut points are sequenced using high-throughput technologies to determine open chromatin regions at genome-wide level [20].*

RNA-seq. *Analysis of the transcriptomics and expression profile of the sample to study. Process based in isolation and purification, followed by high-throughput sequencing. The abundance of RNA is calculated by the alignment to a sequence from the raw reads.*

WGBS (Whole Genome BiSulfite Sequencing). *Consists in a treatment of the DNA sequence with sodium bisulfite, which converts cytosines to uracils, whereas methyl-citosines remain unmodified. Then, it is amplified by PCR and resultant sequences will have methylated cytosines unmodified and unmethylated cytosines converted to thymines. The comparison of the modified DNA with the original sequence allows to infer the methylation state of each position of the original DNA [13].*

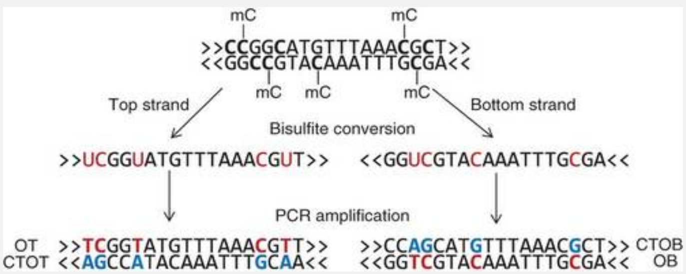


Figure 1.2: Process of bisulfite treatment for whole genome methylation sequencing. Adapted from Krueger F, et al.,2012.

1.2 The Blueprint Consortium

Despite extensive studies of regulatory elements have successfully allowed to create a start point in the study of the epigenome, there is still a lot to discover and understand. This is why in the last years big consortia and international projects are being set up for a better understanding of the epigenome and the regulation of gene expression.

One of those projects is the European Blueprint Database [1], which started in 2011 and was aimed to study the epigenome of both healthy and diseased hematopoietic cells [23, 24]. There are currently more than 100 epigenomes and data from different blood cell types. There are also available blood-based diseases, such as Acute Myeloid Leukemia (AML), Acute Promyelocytic Leukemia (APL), B cell myeloma, and Burkitt Lymphoma.

Hematopoiesis is the process by which blood cells are created in all vertebrate organisms. The hematopoietic stem cells (HSCs) are a pool of pluripotent cells which give rise to all the cells in blood, including the lymphoid lineage (T and B cells) and myeloid lineage (neutrophils, eosinophils, basophils, monocytes, macrophages, megakaryocytes, platelets and erythrocytes) [25].

Mutations and errors during hematopoiesis process are the cause of leukaemia and myelomas, involving genetics and epigenetics. For this reason it is of extreme importance the study of this process to better understand the normal process, as well as providing a start point to study the molecular basis that leads to blood malignancies and cancers. One of the best approaches to this study is in the epigenome, as it plays a key function in the differentiation from HSCs to the different cell types.

[1] www.blueprint-epigenome.eu

1.3 Objectives

The objective of this project is the exploration of regulatory regions of the genome, and the development of a pipeline for whole-genome discovery and identification of novel regulatory regions based in the prediction of active regions from epigenomic data integration.

Pointed objectives of the master project are:

1. Exploration of the Blueprint Database
2. Characterization of differentially methylated regions at TSS regions and CpG islands
3. Statistical approach for prediction of promoter functionality from epigenomic features.
4. Creation of a pipeline for discovery of cell type-specific novel promoters by integration of epigenomics and transcriptomics data.

2 Methodology

2.1 Blueprint Database

The Blueprint Database is one of the biggest databases worldwide related to epigenomics. It contains a complete set of epigenomic information from blood cells samples and there are available many cell types from bone marrow, peripheral blood, cord-blood, etc. Blueprint data can be explored through the portal DDC Blueprint[2] [26]. The total number of samples in the dataset for this analysis was 104, of those 92 are healthy donors and 12 from cancer donors.

2.2 Genomic and Epigenomic data

Blueprint data is open and free to download from the Blueprint ftp Data portal [3]. An index file was downloaded from Blueprint database with all samples information related (sample_ID, donor_ID, laboratory, experiment type, cell type, tissue, etc), and sub-setted for the data of interest (Appendix C script `index_samples.R`).

One donor with the full epigenome set available was selected to perform the exploration of the data; "CD14+, CD16- classical monocyte" with DONOR_ID C001UY. This donor was used to perform the pipeline and for first results. In order to have a reproducibility of the pipeline in more individuals, a series of samples were selected which had as available data: methylome, Chip-seq data(H3K3me3 and H3K27ac), open chromatin data and RNA-Seq. Bash-based commands were used for downloading the data, code is detailed in **Appendix C** script `download_data.sh`.

Reference genome used for annotated regions is assembly hg38/

[2] dcc.blueprint-epigenome.eu/#/home
[3] ftp://ftp.ebi.ac.uk/pub/databases/blueprint/

GRCh38. Annotated TSS regions were downloaded from the Blueprint ftp site (gene annotation format from Ensembl Gencode v22 [27]), and CpG islands regions were downloaded from the UCSC annotation hg38. TSS regions were established by selecting a region from 500 pb upstream to 500 pb downstream from the TSS (a total region length of 1000 base pairs per TSS).

For gene expression data the Reference transcripts from Ensembl Gencode was downloaded also from the Blueprint fpt data website. The TPM (Transcripts per Million) for each gene was computed. TPM is a metric of quantification of RNA-seq which is obtained from the RPK (Reads Per Kilobase) of the transcripts divided by million scaling factor. The TPM transcripts quantification method normalizes for sequencing depth and gene length. The value of the gene expression for one gene is the sum of the values of all the transcripts that map within this gene.

2.3 Description of methylation

For the study of methylation between cell types, we studied 92 healthy samples of 12 different cell types from both the myeloid and lymphoid lineages. The 5mC data from bisulphite sequencing (BS-seq) at whole genome level was analyzed. For a given CpG position, methylation value is calculated by the following estimation:

$$\beta_i = \frac{y_i}{x_i + y_i}$$

where x_i is the number of converted reads and y_i is the number of non-converted reads at position i. The β value statistics is a number between 0 and 1, and reflects the methylation status of all the cells in the sample. For instance, in ideal conditions a value of 0 indicates that all copies in that CpG site in the sample are unmethlated and a value of 1 indicates that all copies of the site are methylated.

First, we created an algorithm called `make_combine.py` to recover the methylation values for the CpGs per TSS region. This algorithm is developed in Python language [28], code is detailed at **Appendix B.1**. Followed, we calculated the statistics per TSS ; median number of CpGs per TSS, number of CpG sites which have methylation value, methylation values statistics and the standardized score (z value) for neutrophils vs monocytes. Same statistics were performed fo CpG islands. Statistics were calculated with R software [29].

2.4 Exploration of the data

As pointed above, for the discovery of new promoters the studied cell type was a "CD14+, CD16- classical monocyte"(DONOR_ID C001UY). We did the exploration of the data by using Bedtools software [30], which was used to find the closest distances between two different features, as well as to find intersections between features. Intersections between features were performed for TSS against peaks of H3K4me3, DNase hypersensitivity, and CpG islands.

2.5 Principal Components Analysis

We performed Principal Components Analysis (PCA) in order to see the variability explained by median methylation in TSS regions at whole-genome level. The Principal Components Analysis is unsupervised method based in dimensional reduction of the correlated variables to a small number of uncorrelated variables called principal components, which will be used for data reduction and visualization. Each of those resultant components is associated with a linear sum of variables [31]. PCA was performed with R software [29] with function `prcomp` from the package `stats`. Samples were coloured by lineage and by cell type. The same proceeding was done for CpG islands regions.

In order to find out the functionality of genes that explain more variability in the PCA, we performed annotation analysis for the to p 5 genes (which explain more variability in PC1)with the `biomaRt` package [32] from R software. In this step, the ENSEMBL_ID of each transcript was used in order to get the chromosome, start position, gene name and gene functionality.

2.6 Predictive model

We built a logistic regression model, which is a method to study the association of a set of variables (covariates) X1,...Xk with a response or dependent variable Y. The output variable in the logistic model is a dichotomous factor [33]. Logistic regression model was built with function `glm` from `stats` package [29].

With this model we want to predict if a gene is active or inactive from the information of the regulatory marks in the regions. The co-variables we are going to use as predictors are the epigenetic marks of the promoter regions (TSS $\pm$ 500 bp region), and the output of the model is the active (expressed) or inactive (repressed) state of the gene. Predictors:

- DNase : Dichotomous variable for DNase hypersensitivity hotspot intersection (coded 1) or no intersection (coded 0) in the TSS region.
- HM : Dichotomous variable for H3K4me3 peak intersection (coded 1) or no intersection (coded 0) in the TSS region.
- Meth : Continuous variable for methylation median value of CpG sites at the TSS region.
- RNAseq : Continuous variable for median read counts in TSS region.

Output: A dichotomous outcome variable to indicate active gene (coded 1) or inactive gene (coded 0).

The data was randomly split in two sets, one set of training data and another set of testing data. Thus, half of the TSS were used to fit the model, and the other half were used to test it. The training set of TSSs (28,688 TSS regions) was used to build the model. To do so, we first transformed the gene expression data in TPM values to dichotomous variable by choosing as cut-off the third quartile (q_3). Once we had all the information of the covariates and the output for each TSS, we built the model formulated as:

$$logitP(Y = 1) = \beta_0 + \beta_1 * DNase + \beta_2 * HM + \beta_3 * Meth + \beta_4 * RNAseq$$

After the model was fitted, we tested its predictive power. To do so, we used the testing data for prediction of the outcome. The probability of y=1 is:

$$P(Y = 1) = \frac{1}{1 + exp\{-(\beta_0 + \beta_1 * DNase + \beta_2 * HM + \beta_3 * Meth + \beta_4 * RNAseq)\}}$$

where π is the probability of y being 1, β_0 is the basal value, β_1 to β_k are the covariates to fit the model. In this case, we chose a threshold of 0.5, so whenever P(Y)>0.5 we will predict it as P(Y = 1). A confusion matrix was constructed then with the real data and the predicted data, and classification accuracy was estimated with calculation of precision and recall:

$$precision = \frac{TP}{TP + FP}; \quad recall = \frac{TP}{TP + FN}$$

2.7 Target regions pipeline

The pipeline is Bash command-line and uses programming languages GNU awk, GNU datamash, Bedtools software and R software. (**Appendix B.2**). We use BEDTools software [30] for defining intersected regions. The methodology used for my purpose was `closestBed`, which returns the closest feature to each entry in a BED/GFF file.

For performing closestBed, two input files in BED (Browser Extensible Data) format must be given. A BED file consists of a tab-delimited columns with three required fields: (1) chromosome, (2) Start Position, (3) End Position, and it is usually followed by (4) unique ID for each region. Given two BED format inputs files A and B, BEDTools closest returns the regions where file A has its closest feature in file B and the genomic distance to it (**Figure** 2.1). The result is in form of standard output.

```
Usage:  $ closestBed [OPTIONS] -a <BED> -b <BED>
```

To visualize the code for the *Target regions pipeline* refer to **Appendix** B.2.

Figure 2.1: Graphical explanation of usage of closestBed. In this case, feature B2 would be reported as the closest feature to feature A. With the parameter `-d` the output will tell us the distance to that region in base pairs. Adapted from Aaron R. Quinlan et al., 2010.

The selected regions were explored and visualized using several software and tools, such as the Integrative Genomics Viewer, the UCSC Genome Browser and the `Gviz` package from R. The Blueprint Consortium has provided a Data Track Hub in UCSC [4] browser in order to visualize the data from samples in the genome browser. New tracks can be uploaded from the local and be integrated with the data in the Hub. This way, I could integrate and visualize my target regions with the other elements available in UCSC (CpG islands, genes, transcripts, regulatory marks, etc)

[4] Blueprint data set Track Hub

2.8 Exploration of the transcriptomics

We estimated transcripts values per region by checking the number of reads in each region (1,000 bp). RNA-seq is estimated by recovering values of raw reads from the `bam` file for the regions of interest, with an algorithm called `check_rna_seq.pl` created in advance by the group. This script uses SAMtools to take the coverage values in a region of the genome that has to be indicated. The algorithm is written in `Perl` language and pipeline is written in command line. RNA-seq was estimated for TSS regions (i), list of all target regions (ii) and random regions (iii) for comparison of results.

For each of those regions, we carried out a summary of the RNA-seq values per region. Because of the fact that median values could be drag by the potentially high number of zeroes, we took the third quartile (q_3) per region for further analysis and plots. For instance, statistic analysis output is a summary with minimum, q1, median, mean, q3 and maximum per each TSS. For a given TSS i I would take the $q3_i$. This value was estimated for all regions and plotted using R. To visualize the code used refer to **Appendix** B.3.

3 Results

3.1 Methylation analysis of Blueprint Samples

DNA methylation can give important clues about the cell differentiation of the cells during hematopoiesis [3]. The methylation state of the TSS at whole-genome was analyzed for 92 samples. For this analysis, the number of TSS taken for the analysis is 57,376, which are all annotated TSS regions from chromosome 1 to chromosome 22 (autosomic chromosomes).

After filtering the regions with "Not Available" methylation values the number of TSS goes down to 39,949 TSS. Median values methylation per sample are in a range of 0.60-0.95, and a median of 0.91. Therefore, TSS are highly methylated at whole-genome level.

The PCA was performed with 92 samples of 12 different cell types. Plot of the PCA results is shown in **Figure 3.1**. It shows the clustering of the cell types mostly by PC1 (36% of the variance) and by PC2(15%). First principal component separates populations by median methylation value; all myeloid cells are aggregated whereas erythrocytes, HUVEC cells and lymphoid lineage (CD4 T cells, CD8 T cells, B cells and NK cells) are more segregated. B cells are known to differ in their methylation level during maturation, having a demethylation at the late stages [3]. In this PCA we can clearly see that they follow this pattern.

The second component separates the cells by lineage, as we can see that a the left all myeloid are aggregated (monocytes, neutrophils, macrophages, eosinophils and dendritic cells) and lymphoid at the right. Finally, PC3 clusters HUVEC cells very away from the rest of cell types. HUVEC cells (Human Umbilical Vein Endothelial Cells) are used in order to see if they act as outlier cell types and check that the experiments are correct and we can confirm it in this analysis.

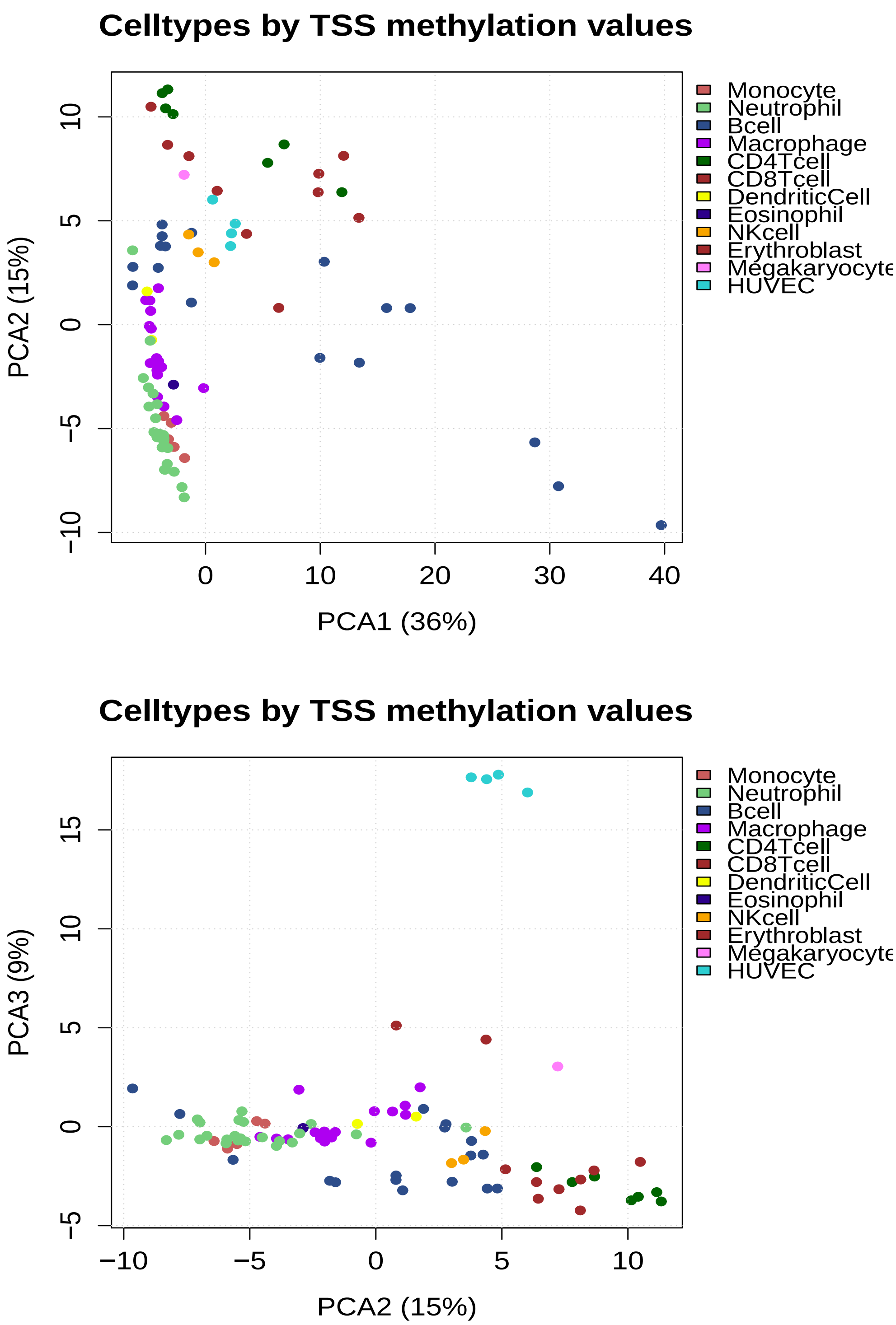


Figure 3.1: PCA of 92 healthy samples TSSs methylation mean clustered by cell type.

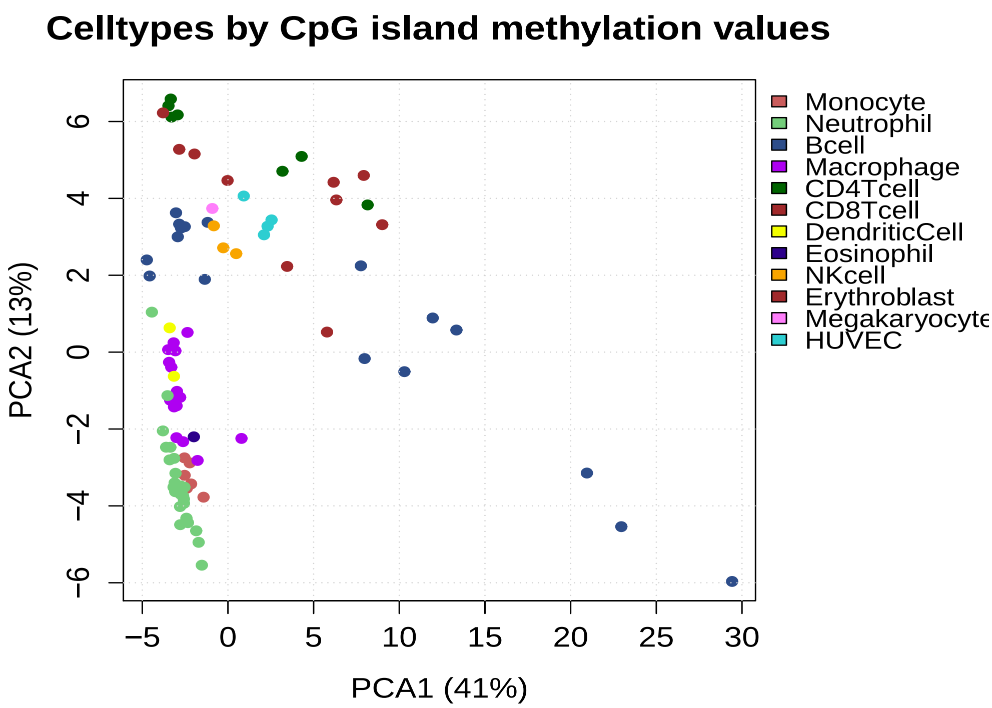
Table 3.1: Annotation of top 5 genes of first principal components. These five TSS are the most explanatory of the variability between both lineages.

Chr	TSS position	Gene name	Description
1	28,579,764	SNORA61	small nucleolar RNA, H/ACA box 61
1	12,065,002	Metazoan_SRP	Metazoan signal recognition particle RNA
11	9,578,028	snoU13	Small nucleolar RNA U13
14	22,507,666	TRAJ34	T cell receptor alpha joining 34
16	30,471,587	Y- RNA	-

The annotation of the "top 5" genes of first principal component is shown in **Table 3.1**. T cell receptor alpha gene is a protein-coding gene related to the molecular signalling process of T cells. It is clearly changing its methylation pattern between myeloid cells and lymphoid cells. The other genes which explain more of the variability explained by first component are RNA and snRNA.

For the CGIs a total of 17,356 regions were used for Principal Components Analysis in the 92 healthy samples. As we can see in **Figure 3.2**, samples are clustered by PC1 and PC2 following the same pattern than with the TSS methylation medians. One possible reason could be that TSS and CGIs overlap because promoters are regions CpG island-rich. In this case, explained variability by first component is 41% and explained variability by second component is 13%.

The methylation retrieval and PCA in CGIs is ought to be repeated with a new set of CGIs coordinates from the hg38/GRCh38 genome because we suspect that analysis were performed using the CGIs from hg19 and, as a consequence, median methylation values we have been using for this analysis would be from random genomic regions instead of from CGIs. In any case, the results obtained would suggest that methylation values are cell type-specific whereas we look at TSS regions or other regions.



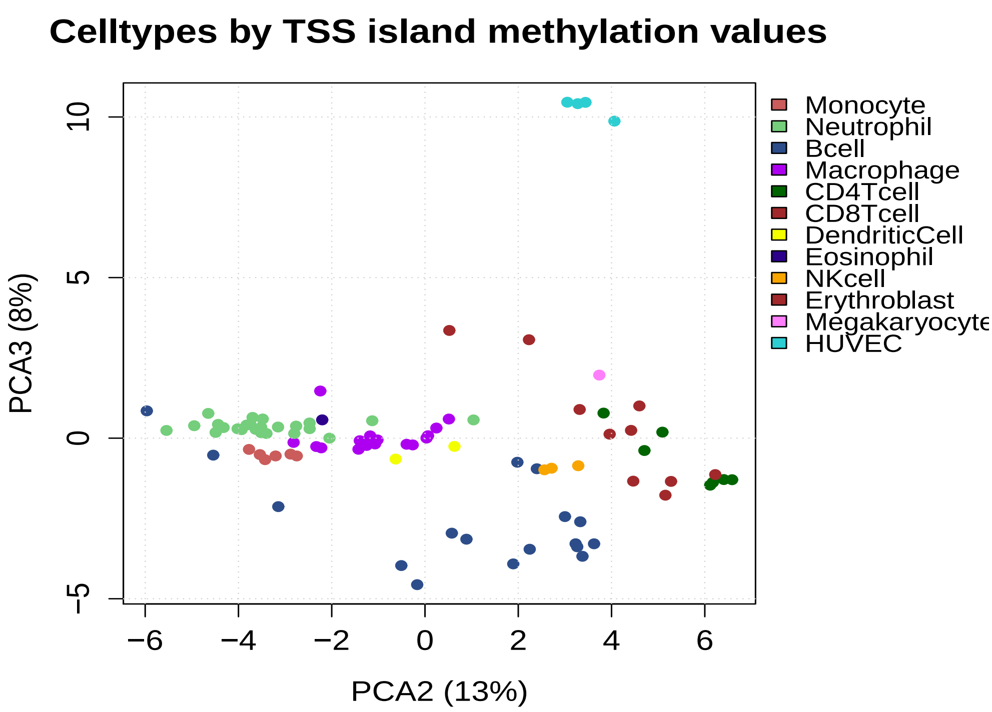


Figure 3.2: PCA of 92 healthy samples CGIs methylation mean clustered by cell type.

3.2 Epigenomic analysis of sample to study

First of all, we have performed a description of the epigenomic features of sample "CD14+, CD16- classical monocyte" (DONOR_ID C001 UY), in order to have a clear view of the sample to study. Features are described in **Table** 3.2. Epigenomic features are calculated taking the whole genome of the sample; autosomic chromosomes (chr 1 to chr 22), mitochondrial DNA (chr M) and sexual chromosomes (chr X and chr Y).

Table 3.2: Epigenomic features of sample C001UY. Number in whole genome, region median length and fraction of the genome covered.

Epigenomic Feature	Number	Median length	Fraction of genome
5mC methylation	31,289,743	1	0.97 %
DNase hypersensitivity hotspots	224,435	297	2.06 %
H3K4me3 peaks	31,580	741	0.72 %

First approach is to characterize the epigenomic features in TSS in order to have a start point on the characteristics of regulatory regions in the genome. Total number of TSS regions for analysis in this sample is 57,376 whole-genome. In average, there is a mean of 16 CpG sites per TSS region. The intersection between TSS regions and CGIs is of 17,267, which is the 30.1% of TSSs. The median distance of TSS to CGIs is 12,820 base pairs **Figure** 3.3.

Data from previous studies describe that the number of promoters that have CpG islands is near 40% [6, 11], so the results in this analysis are quite good taking into account that we have selected regions of 1000 base pairs around the TSSs, but CGIs could be located in a wider region around the promoters.

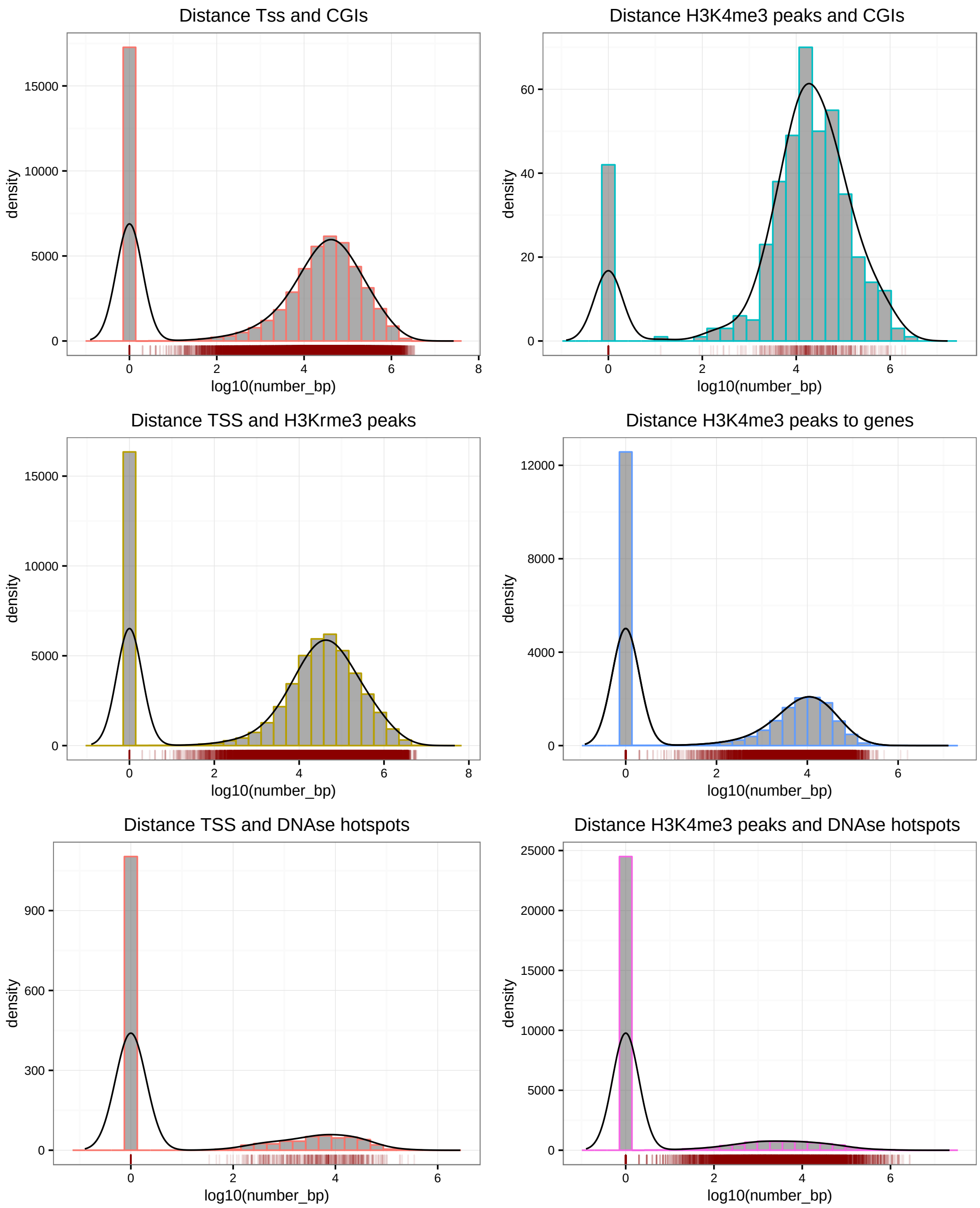


Figure 3.3: Histograms of distance distribution between features and genomic elements. 1) Distance TSS with CGIs. They intersect in a high density; 2) H3K4me3 peaks with CGIs. Distances are mostly in the 10 e4; 3) Distance TSS and H3K4me3. High density of intersection; 4) Distance H3K4me3 with genes. High density of intersection (marker of active regions); 5) Distance TSS and DNase hotspots. 6) Distance H3K4me3 and DNAse. High percentage of intersection.

Epigenomic signature has been reported in first exon regions [8], with a strong association of HMs with the regulation of splicing at first exon [14]. Thus, we have intersected data of H3K4me3 peaks with genes to see how many overlap. The median distance between genes and peaks is of 3,700 base pairs, and the number of intersections at whole genome level is 18,873 (87.6 %) (**Figure 3.3**). This high percentage indicates that HM H3K4me3 is a good indicator of active regions.

We have also studied the intersection of TSSs with epigenomic features (**Table 3.3**). The median distance from features to TSS region is much larger for H3K4me3 than that for the DNase, probably because of the total number of regions in the genome is much lower. This result indicates that in a great percentage of promoters we can find those epigenomic features and those could be the active promoters.

After this exploration of distances, we want to see the implication of regulatory elements in the gene expression. For that purpose we checked the gene expression of genes whose TSS regions intersect with regulatory regions versus genes whose TSS regions do not intersect with regulatory regions. For this analysis we used a total of 21,539 genes from chromosome 1 to 22. This analysis was performed for H3K4me3 peaks as well as for DNase hotspots.

In **Figure 3.4** we show the differences between both in the $\log_{10}$ scale expression. Gene expression is about 10 times higher in genes

Table 3.3: Epigenomic features in TSS regions

Epigenomic Feature	**Number of Intersections**	**Fraction of TSS**	**Median distance feature to TSS**
DNase hotspots	19,095	39.5 %	3,625
H3K4me3 peaks	16,847	27.8 %	17,940
Both features	15,425	26.9 %	-

where its TSS has intersection with the selected marks. This observation indicates that open chromatin and HM could be good predictors of transcriptionally active promoters and that they it could be accurate to choose such features to start genome-wide identification of active regulatory regions.

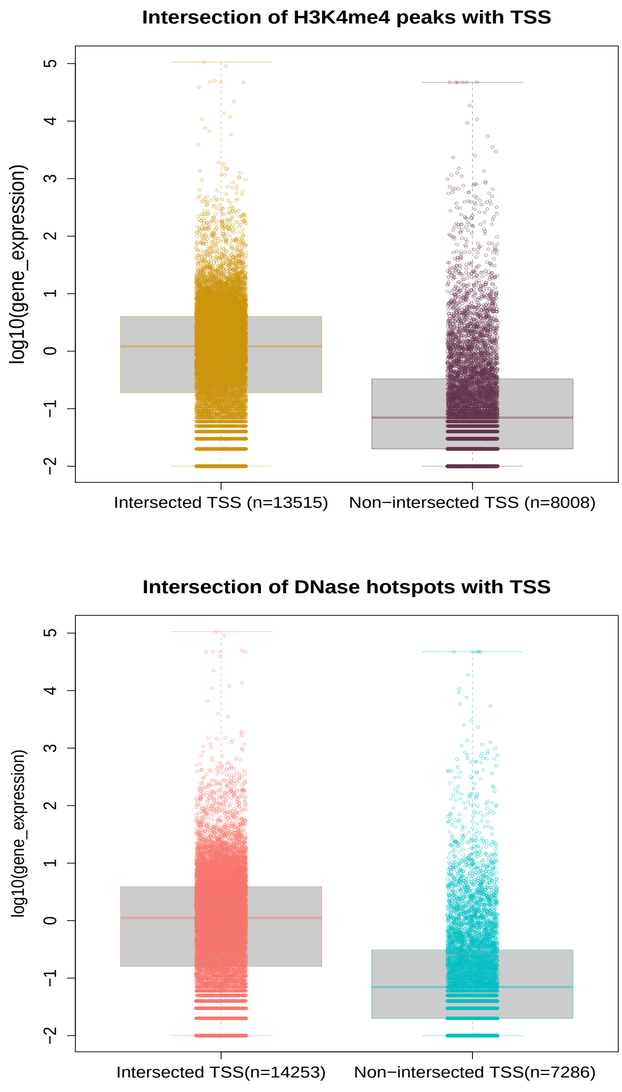


Figure 3.4: Boxplots of values of gene expression for genes whose TSS have intersection with epigenomic features versus genes which TSS have no intersection. a) Gene expression differences for TSS and HM peaks b) Gene expression for intersections of TSS and open chromatin regions.

3.3 Prediction model for promoter function

To give an statistical power to the observations above, supervised analysis based in logistic regression model was performed (deeply described in the Section 2.6). The TPM values varies between active and inactive TSS so this analysis is aimed to test the predictive power of regulatory elements and RNA-seq in promoters to predict the activity of genes. As we can see in the **Table 3.4**, best predictors of gene expression are DNase hypersensitivity and H3K4me3 peaks intersections. Thus, those features are the ones to have into account for doing prediction of novel unannotated promoters with more robustness.

Table 3.4: Logistic Regression Analysis of 28,688 TSSs based in regulatory elements. Notice that methylation value and RNA-Seq are not statistically significant in explaining the output value, whereas H3K4me3 peaks and DNase hypersensitivity are strongly significant.

Predictor	**Variable type**	β	**SE** β	**p-val**
5mC methylation	Continuous	-2.5972	0.0952	0.895
DNase hypersensitivity hotspots	Dichotomous	1.4455	0.0568	$< 2e\text{-}16$
H3K4me3 peaks	Dichotomous	1.981	0.0808	$< 2e\text{-}16$
RNA-Seq	Continuous	-0.0001	0.0002	0.430

For testing the predicting model we tried to predict whether the testing TSSs were active or inactive based in the input variables. Of all testing TSSs, when the probability given by the model was higher of 0.5 we transform it to 1 and we would consider it as predicted to be expressed.

The results of the prediction are compared with the observed value, shown in **Table 3.5**. The measure we used to validate the predictive model are precision (positive predictive value) and recall (sensitivity) [33]. Precision of the model is 0.25 and Recall is 0.27. This values would be due to the fact that we are using only one histone

Table 3.5: Observed and predicted frequencies for prediction of gene expression TPM by logistic regression model.

		Predicted	
		Expressed(1)	Repressed(0)
Observed	Expressed(1)	1,670	4,390
	Repressed(0)	4,872	13,166

mark and the DNAse hotspots to predict the gene expression. Other studies using a higher number of histone marks [35] have been able to predict gene expression levels based on the model with a larger number of predictors.

3.4 Discovery and identification of novel promoters

3.4.1. Target regions pipeline

The pipeline for finding novel promoters was created using sample C001UY epigenomic features, by looking at regions of intersection between peaks of H3K4me3 and DNase hypersensitivity hotspots which fall more than 10,000 base pairs far from the nearest annotated TSS (**Figure** 3.5). The pipeline was first carried out for chromosome 1 and then to whole genome level (chromosomes 1 to 22, Chr M , Chr X and Chr Y). This pipeline is purposed to be carried out in any sample to study in order to have a big dataset of target regions along the genome for multiple samples to compare.

In the first approach of the pipeline on sample C001UY, we selected a total of 24,492 regions from chromosome 1 to chromosome Y. From these regions, 12,572 intersect with annotated TSSs. So, 51.3% of the regions with those features correspond to annotated promoters. This number is similar to expected, as we explored the number of TSS which intersect with those features previously in Section 3.2 and this number corresponds to expectations.

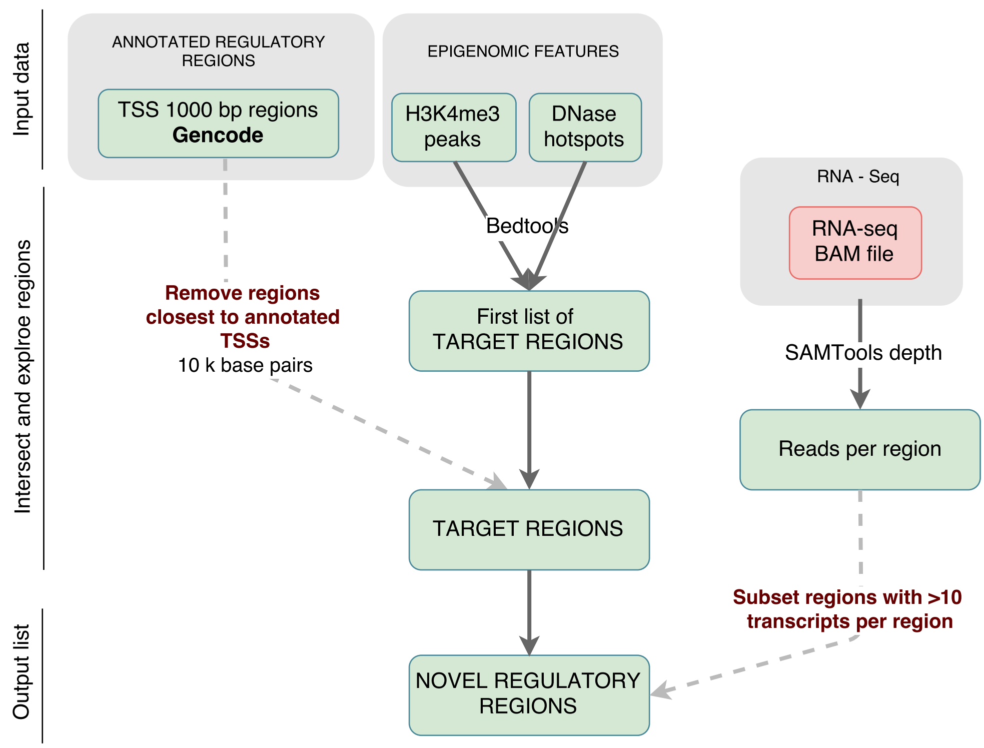


Figure 3.5: Simplified graphic of pipeline for target regions.

From this list of target regions, we select the ones that are at least 10,000 base pairs away from already annotated promoters, whether they would fall in intragenic or intergenic regions. The number of regions according to those features filtered is 6,538. So, we could say those are the total number of regions which find described epigenomic features and are not annotated, thus could be specific regulatory regions of the monocyte we are studying.

3.4.2. Exploration of target regions

Methylation values of the selected regions were studied. From the list of target regions, only 510 regions have available methylation data. The median methylation value for target regions is 0.056. This value agrees with the low methylation values that we could find in active regulatory regions [6].

In order to visualize how a target region would look like in the genome, I have plotted the first region in my list of target region (**Figure** 3.6). Each track represents one of the features to study. This region is found in the second intron of the gene AGRN, a protein-coding gene related to development of neuromuscular junction.

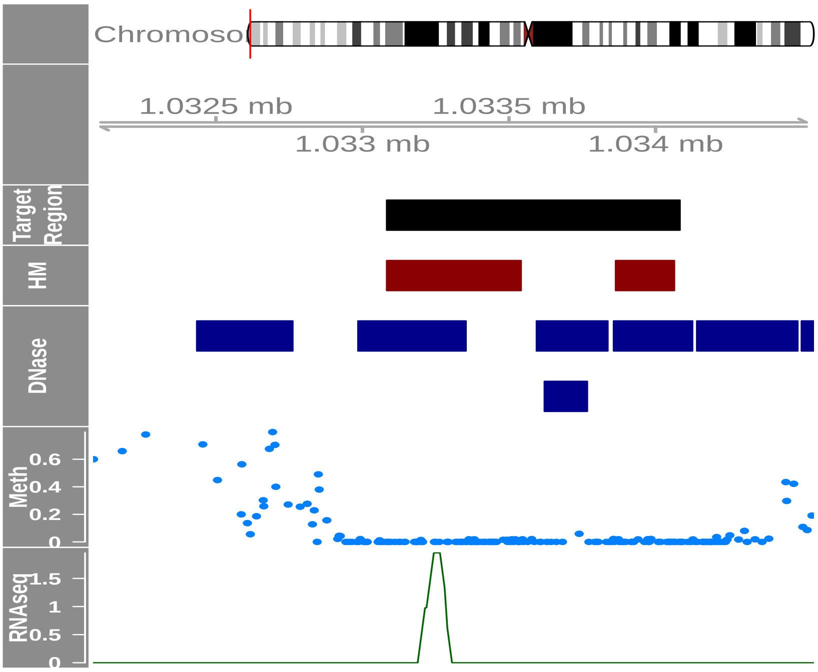


Figure 3.6: Representation a of target region. Track 1) Ideogram of chromosome with a red line in the position of the target region; Track 2) Zoom in the 2000 pb region (1.0325 Mb-1.0345 Mb); Track 3) Region selected as novel regulatory region (1000 bp) ; Track 4) H3K4me3 Peaks at the region; Track 5) DNAse hypersensitivity hotspots; Track 6) CGs methylation values in a range from 0 to 1. Track 7)RNA-seq values per base (depth).

3.4.3. Exploration of the transcriptomics

Transcripts quantification is expected to be high around active regulatory regions, for example in active promoters [11,35]. For the purpose of checking the transcripts level around given regions, the RNA-seq pipeline was performed. This pipeline was carried out for annotated promoters (i); target regions list (ii); and random regions (iii).

Total number of TSSs analyzed for RNA-seq is 33,443 (58.2%) after removing the regions where there was Not Available RNA-seq data. Number of target regions with available RNA-seq data is 5,057 (77.5%). In order to compare to random RNA-seq values, the pipeline was also performed in 10,000 random positions of 1000 base pairs length from Chr 1, for which only 4,172 (41.2%) had RNA-seq reads. Distribution of the q_3 of the values in the region was computed and plotted in the log_{10} scale and is shown in **Figure 3.7**. Notice that the RNA-seq median value of target regions is higher respective to TSSs with low TPM gene expression, as well as higher than random regions.

Comparison between target regions and random regions was performed with a test for equality of means for impaired data. Re-

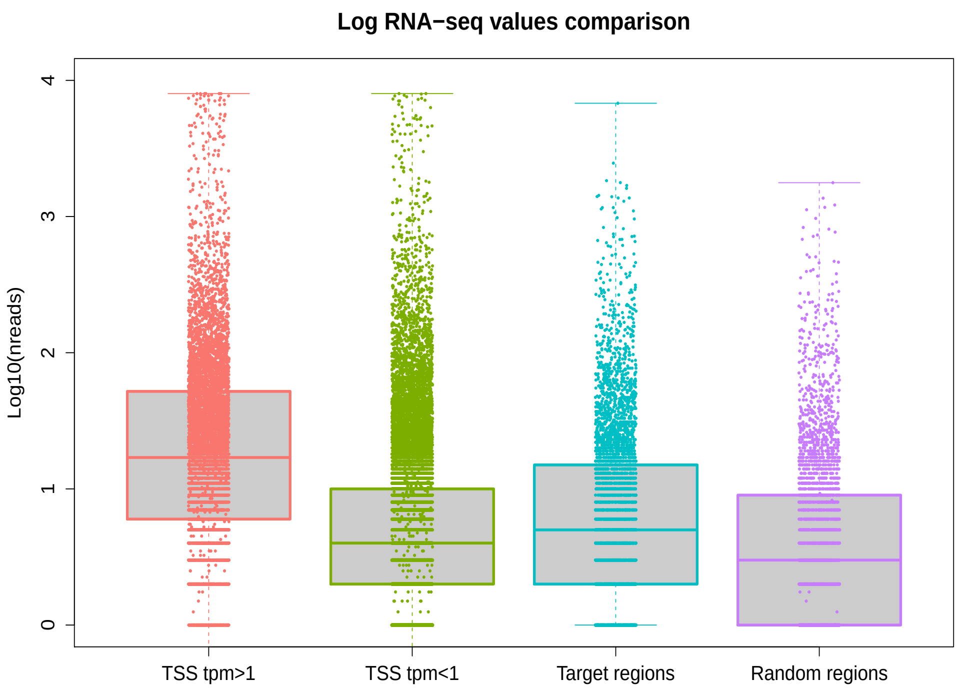


Figure 3.7: Boxplots of comparison of RNA-seq quantification between regions. Transcripts are calculated as the Log10 of the q3 value of reads per region. Transcripts distribution for: 1) TSS with gene expression >1 TPM, 2) TSS with gene expression <1 TPM, 3) Target regions regions 4) Random regions in chr 1.

sults of the test shows that, with a p-value of $< 2e\text{-}16$, the differences between the means are statistically significant. This observation indicates that the selected regions in our pipeline have significantly higher transcripts abundance than it would be expected by chance.

On the other hand, TSSs RNA-seq distribution was compared with the gene quantification (TPM) for protein-coding genes. This correlation is performed to visualize the association between transcripts abundance in our data with the gene expression profiles. This is important to know because in the list of the target regions they have not genes associated to see if there is gene expression as we did in the TSS regions, then we have to use the RNA-seq data as a quantification of the gene expression.

Figure 3.8 shows the correlation between RNA-seq we found around the promoter region and the gene expression of related genes.

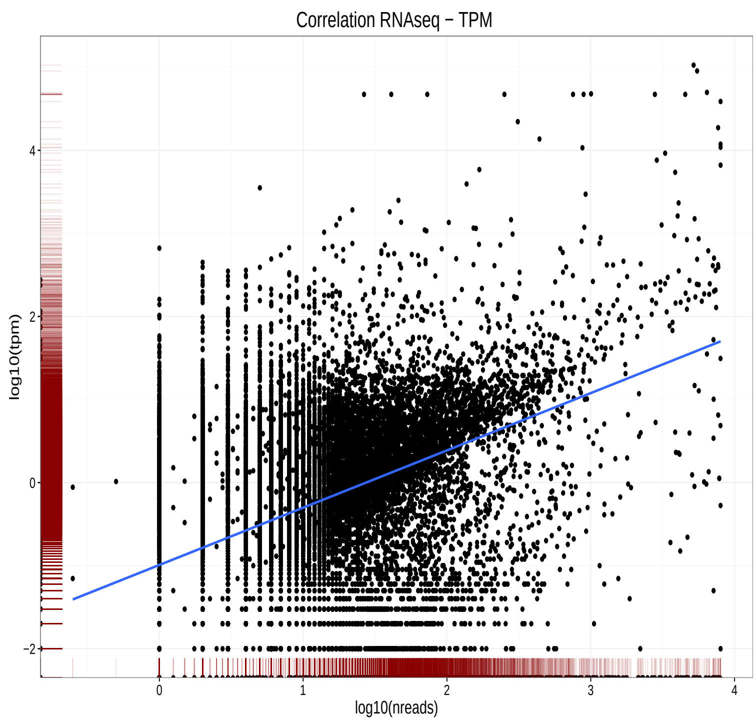


Figure 3.8: Correlation plot of gene expression (TPM) and RNA-seq at TSS regions.

Data was also fitted in a linear regression model with the RNA-seq as predictor and the TPM values as outcome. The p-value of the test is $< 2e\text{-}16$ and the adjusted r-squared of the regression is 0.04. With these results we can say that RNA-seq can be a good quantification predictor of gene expression at the target regions.

3.4.5. Subsetting target regions

After the performed transcriptomics analysis, we filtered the target regions list to those which find a high gene expression around, as transcripts are expected to be found next to promoters and regulatory regions. The threshold for selection is 10 reads per site as a median in the region of the target region. After carrying out the filtering, we obtain a total of 1,616 regions out of 6,538 (24.7 %) which have high transcripts expression around.

3.5 Reproducibility of the pipeline

3.5.1. Description of samples to study

For reproducibility of the pipeline a first approach was the description of samples to study. Not all donors have the full set of data available, and most of them have only a few.

For what concerns to DNase hotspots, only 28 samples of Venous Blood donors had availability for this sort of data. Table of DNase description is refereed in **Appendix D**. As we can see in **Figure 3.9**, there is few variation between cell types for the median length of the regions (270 to 300), whereas for the number of regions, the range is wider, being the higher number for three samples of macrophage that act as out layers.

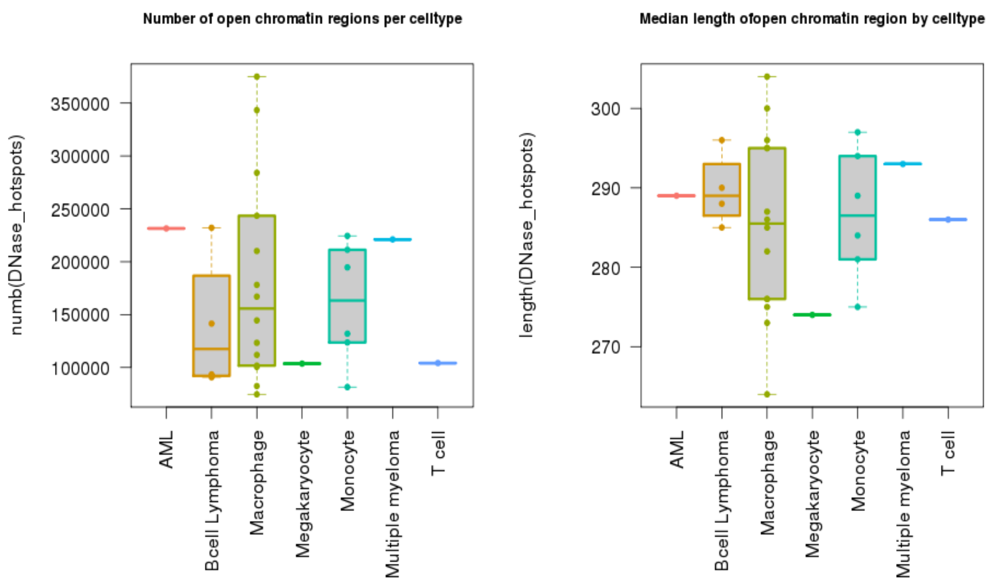


Figure 3.9: a) Number of regions of open chromatin per cell type. Range from 100 K to 350 K regions at whole genome level. b) Median length of open chromatin regions. Range from 250 to 350, and means at 290. Available cell types for DNase info are macrophages, megakaryocyte, monocyte, T cell. Cancer samples; Acute Mieloid Leukemia (AML), B cells lymphoma (includes Sporadic Burkitt Lymphoma samples), Multiple Myeloma

Same analysis was performed for H3K4me3 peaks. From venous blood donors, there was a total of 80 samples with available ChIP-seq data. The descriptive table of the feature description per sample is in **Appendix D**. In the boxplots **Figure 3.10** we can see that the number of peaks along the genome is higher in lymphoma and myeloma samples (up to 70,000 peaks along the genome) than that for the rest of samples (20,000-50,000 peaks).

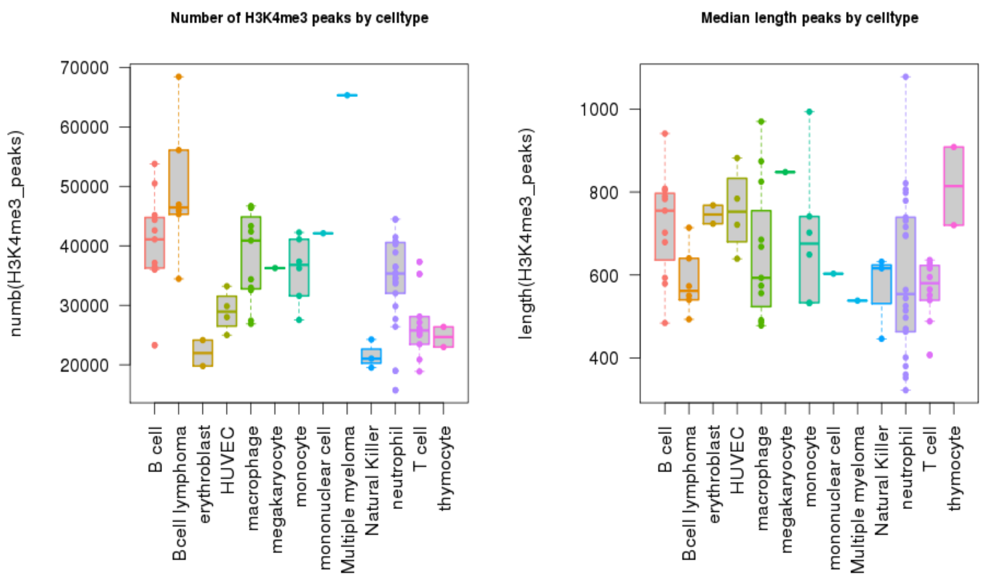


Figure 3.10: a) Number of regions of H3K4me3 peaks per cell type. Range from 20000 to 70000 peaks in a whole genome scale. b) Median length of H3K4me3 peaks regions. Range from 400 to 800 base pairs long approximately. Cell types with HM peaks are: lymphoid lineage(B cells, T cells, thymocites), myeloid lineage (erytroblasts, macrophages, megakaryocyte, monocytes, mononuclear cells, natural killers, neutrophils), and cancer samples cells (B lymphoma cells and one multiple myeloma sample)

From the results about number of regions we can see that there is a high variability in the number of regions between samples, even for the same cell type (i.e the number of DNase hotspots regions of macrophages), leading to suggestions that there could be a batch effect.

The Blueprint consortium has many different laboratories associated which sequence the data of the varying types of experiments. Thus, there could be a batch effect that affect the results of samples that are extracted and sequenced in the different laboratories. To assess whether there is batch effect that would affect the biological variability, we have described the data for different laboratories and visualized it in box-plots by laboratory.

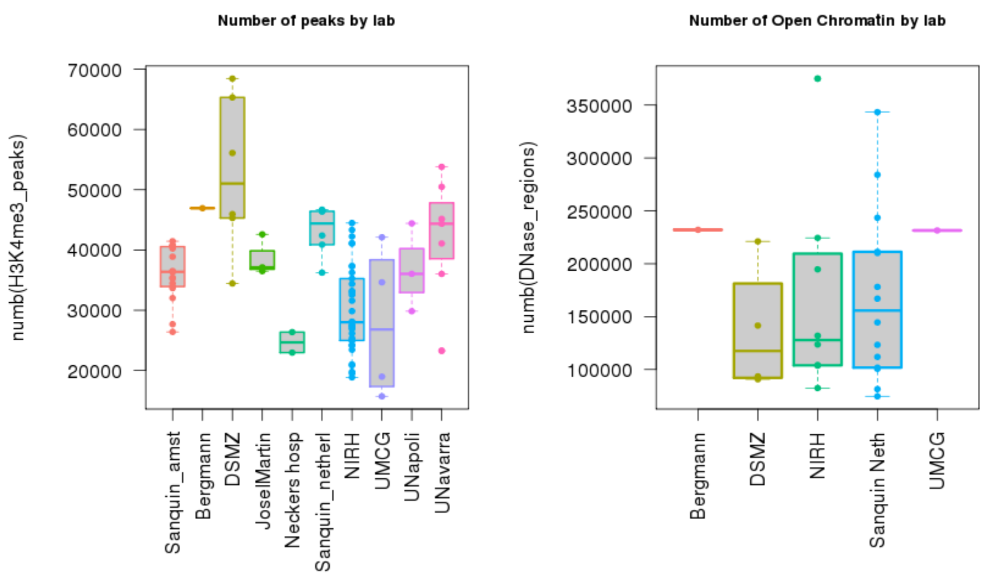


Figure 3.11: a) Number of regions of H3K4me3 peaks by lab. b) Number of regions of open chromatin by lab.

In **Figure 3.11** it is represented the number of both epigenomic features at the whole-genome level in the different laboratories. Laboratory DSMZ is the one with more difference, and taking a look at the sample this laboratory has provided we see that the cell types are Sporadic Burkitt lymphoma, Mantle cell lymphoma, Germinal centre B cell and Multiple myeloma. The same for laboratory Bergmann, whose only one sample is Multiple myeloma type.

Thus, this observation leads to deduce that differences in those laboratories cold be not due to a batch effect, as biological variability is confounded with the change of laboratory. From this graph we could also infer that cancer samples differ from healthy cell types in the number of H3K4me3 peaks rather than in the number of open chromatin regions.

3.5.2. Pipeline reproducibility on other cell types

To perform the pipeline in other cell types we selected four samples that have all dataset of information for DNase hotspots, H3K4me3 peaks, H3K27ac, DNA methylation, and RNA-seq. The target regions pipeline was carried out for all the samples selected and the number of regions is described with the samples description in **Table 3.6**. To notice is the high number of regions selected with the pipeline for the macrophage N00031319896021, which is not due to a batch effect as it was seen earlier.

Table 3.6: Description of samples to study with information of donor ID, cellular type, laboratory where it comes from and number of target regions

DONOR_ID	Cell type	Laboratory	# target regions
C001UY	monocyte	NIHR Cambridge	6,538
C005VG	macrophage	NIHR Cambridge	2,374
C0066P	T cell	NIHR Cambridge	4,697
S001MJ	macrophage	NIHR Cambridge	2,660
N00031319896021 untreated	macrophage	Sanquin Netherlands	12,300

4 Discussion

The methylation state of the genome is an indicator of the cell state in differentiation during development in early life, and is indicator as well of the hematopoiesis stage. The hematopoiesis process entails a wide range of cell types with very different functions, molecular properties and lifespan, thus is a great opportunity to study the methylation differences along the genome between cell types. One large scale study carried out by the CNAG group, among others, studied widely the methylome in B cell differentiation. They found changes in the methylation state in 30% of CpG sites [3].

In this project, I have looked for methylome changes in TSS regions and CpG islands, instead of all CpG sites.Unsupervised analysis segregated cell types according to methylation state. Principal Component Analysis of methylation level across TSS and CpG sites clustered cell populations in a similar way, clustering by lineage. First principal component separates by methylation value, probably from higher methylation state to lower methylation state. B cells differentiation is characterized by high variation of their methylation state from naive B cells to plasma B cells, with a demethylation at the late stages [3], thus the first component is clearly segregating the sub populations of B cells.

The study of the epigenome of the sample to study has given important information in order to have a global view of the features at a whole genome level. The number of open chromatin regions covers the 2 % of the genome, very similar to studies performed in immune cell types in open chromatin assays for DNase I and FAIRE [9]. The great number of open chromatin regions (224,435) indicates that it is a good marker of functional elements, although it provides few information on the role of those regions (i.e it is not distinctive). For being able to identify the promoters from the DHS regions, we would have to look for TFBS within those regions, as well as knowing the nucleosomes position map by MNase-seq [34].

Epigenomic features have a high density of intersection with TSSs, in both cases around the 30 % of TSS intersect with the features. This information along with the box-plots of gene expression at intersected TSSs show the implication of both features in activation of transcription at promoter level. Supervised analysis based in logistic regression with epigenomic features and RNA-seq confirmed the predictive power of H3K4me3 peaks and DNAse hotspots as predictors of mRNA transcript abundance. Thus, with all of the analysis performed we could suggest that HM and DNase are good predictors of gene expression, as reported in previous studies [17,19–21,35,36]. On the other hand, we have observed that the prediction model should be improved by integrating more marks.

After all the analysis performed on exploration of regulatory regions and all information obtained above, we have been able to stablish a pipeline for novel promoters discovery, in which we use the chromatin signature and histone modification data to select target regions. Methylation data was not taken for the pipeline, as methylation state information is redundant to the chromatin state features. The same case is for RNA-seq transcripts, as p-value for both predictors is not statistical significant in the model created based in a logistic regression model.

The study of the RNA-seq in the regions has allowed us to confirm the list of target regions as possible new regulatory elements. Comparisons with random regions results in significant differences , so we can say that the transcripts found at target regions regions are higher that those for random region of the same length, thus they are not due to chance.

4.1 Further work

One first approach to further work is the analysis of the target regions in the samples selected for reproducibility. By comparing the results of the pipeline among different cell types from the hematopoietic branches we could have a better characterization on the regulatory machinery of the different lineages and cell types. It is also necessary to look at the locations where those regions are, whether they are in gene bodies (introns or exons), intergenic regions or repetitive regions, among others.

It would be of consideration to integrate more marks of histone modifications to the pipeline. For example, the HM H3K27me3 to differentiate whether the promoters are active or poised [12]. As well, it would be interesting to add the HM H3K36me3 which is found in high levels in the gene body of the active genes.

One last mark which would be interesting to integrate is the HM H3K4me1 to differentiate enhancers from promoters,as usually higher peaks of HM H3K4me1 correspond to enhancers. Enhancers are more dynamic and less conserved among cell types than promoters, so we would expect that many of the novel regulatory regions defined in this project could be enhancers [18].

5 Conclusions

The work we have carried out in this project is a first approach to characterization of promoters and provides comprehensive information which can be useful for novel cell type-specific regulatory region discovery from the Blueprint database.

Methylation analysis have shown that methylation values at TSS and CpG islands are good markers of differentiation between hematopoietic cell types. With the analysis of HK4me3 and DNase hotspots we have been able to demonstrate the role of epigenomic features in gene expression, and to predict the functionality of those regions based in the epigenomic features.

We have created a pipeline based in defining regions of intersections between the two epigenomic features studied that could predict regulatory regions, and further we have looked at the transcripts at selected regions. The strength of this strategy is that epigenomic features are combined with RNA-seq data in the definition of putative novel regulatory regions, in order to have a complete outline of the regions.

This work could be a start point in the study of cell type-specific promoters in hematopoietic cells from the Blueprint database.

References

[1] Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, et al. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. Nucleic Acid Research. 2014 October; 43(5): 2575–2766.

[2] Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007 February; 128(4): 669–681.

[3] Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. Nature Genetics. 2015 July; 47(7): 746–756.

[4] The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012 September; 489(7414): 57–74.

[5] Agirre X, Castellano G, Pascual M, Heath S, Kulis M, Segura V, et al. Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. Genome Research. 2015 April; 25(4): 478–487.

[6] Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics. 2012 April; 13(4): 233–245.

[7] Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, et al. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLOS Genetics. 2010 September; 6(9): e1001134.

[8] Curado J, Iannone C, Tilgner H, Valcárcel J, Guigó R. Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. Genome Biology. 2015 October; 16: 236.

[9] Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Research. 2011 October; 21(10): 1757–1767.

[10] Winter DR, Amit I. The role of chromatin dynamics in immune cell development. Immunological Reviews. 2014; 261(1): 9–22.

[11] Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, et al. A high-resolution map of active promoters in the human genome. Nature. 2005 August; 436(7052): 876–880.

[12] Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics. 2011 January; 12(1): 7–15.

[13] Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. Nature methods. 2012 January; 9(2): 145–151.

[14] Singer M, Kosti I, Pachter L, Mandel-Gutfreund Y. A diverse epigenetic landscape at human exons with implication for expression. Nucleic Acids Research. 2015 April; 43(7): 3498–3508.

[15] Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R, et al. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Research. 2011 July; 21(7): 1074–86.

[16] Shen X, Orkin SH. Glimpses of the Epigenetic Landscape. Cell Stem cells. 2009 January; 4(1): 80–93.

[17] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology. 2010 July; 28(8): 817–825.

[18] Andersson R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a uni-

fying model. Bioessays. 2014 December; 37(3): 314–323.

[19] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins D, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics. 2007 February; 39(3): 311–318.

[20] Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008 January; 132(2): 311–322.

[21] Wang YM, Zhou P, Wang LY, Li ZH, Zhang YN, Zhang YX. Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells. PLOS One. 2012 August; 7(8): e42414.

[22] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylationsin the Human Genome. Cell. 2007 May; 129(4): 823–837.

[23] Martens JHA, Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica. 2013 October; 98(10): 1487–1489.

[24] Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nature Biotechnology. 2012 March; 30(3): 224–226.

[25] Rice K, Hormaeche I, Licht J. Epigenetic regulation of normal and malignant hematopoiesis. Oncogene. 2007; 26(47): 6697–6714.

[26] Blueprint Accession Database;. http://dcc.blueprint-epigenome.eu/#/home.

[27] Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database: the journal of

biological databases and curation 2016. 2016; .

[28] Van Rossum G, Drake FL. Python Reference Manual. Virginia, USA; 2001. Available from: http://www.python.org/.

[29] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available from: http://www.R-project.org/.

[30] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 March; 26(6): 841–842.

[31] Van Belle G, Fisher LD, Heagerty PJ, Lumley T. Biostatistics: A Methodology For the Health Sciences. Wiley Interscience;. ISBN: 978-0-471-03185-7.

[32] Durinck S, Huber W. Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene); 2015. Available from: http://www.biomart.org/.

[33] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. Springer New York;. ISBN: 978-1-461-47137-0.

[34] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell. 2008 March; 132(5): 887–898.

[35] Karlic R, Chung HR, Lasserre J, Vlahoviček K, Vingrona M. Histone modification levels are predictive for gene expression. PNAS. 2010 February; 107(17): 2926–2931.

[36] Budden DM, Hurley DG, Joseph Cursons JFM, Davis MJ, Crampin EJ. Predicting expression: the complementary power of histone modification and transcription factor binding data. Epigenetics and Chromatin. 2014 November; 7(1).

A List of software and languages used in pipeline

Language	Tool/Package	Process
Bash	BedTools	Target regions pipeline
	SAMtools	RNA-seq pipeline
	wget	Download data
	datamash	Simple statistics
	awk	Parse files and outputs
R	biomaRt	Gene annotation
	ggplot2	Visualization of results
	stats	Statistical models and analysis
	Gviz	Target region diagram
Python, Perl		Parse files
		Extract methylation
		pipeline

B Scripts for workflow

B.1 Methylation retrieval

a) The script `parse_combine.py` retrieves the fields where the methylation value for CGs is found.

```
#!/usr/bin/python
import sys
infile = sys.stdin

cl=0
for line in infile:
  line=line.strip()
  fields=line.split("\t")
  cl+=1
  if (cl==1):
    sys.stdout.write("%s\t%s\t"%(fields[0],fields[1]))
    for i in range(8,len(fields)-1):
      sys.stdout.write("%s\t"%(fields[i]))
      sys.stdout.write("%s\n"%(fields[len(fields)-1]))
```

Listing 1: parse_combine.py

b) Script `make_combine.py` needs that I give it my list of samples and the list of regions. For each region will perform `parse_combine.py` script and the combine_cpg algorithm. The output for each region is stored in a file with the positions of each CG and the methylation values.

```
#!/usr/bin/python
import sys
infile=sys.stdin

cmd="/home/devel/heath/code/combine_cpg/combine_cpg sample.
    txt -r "
dir="/scratch/devel/sgarcia/put_prom_meth/"
cl=0
for line in infile:
```

```
  line=line.strip()
  fields=line.split("\t")
  cl+=1
  sys.stdout.write("%s\t"%(cmd))
  sys.stdout.write("%s:%s-%s\t"%(fields[0], fields[1], fields[2]))
  cmd1="%s | python parse_combine.py > dir%s.txt\n"
  sys.stdout.write(cmd1%(" -M ",fields[3]))
```

Listing 2: make_combine.py

c) For retrieval of methylation values at target regions, I call the two scripts explained above in the script `extract_meth_pp.sh`. As input it needs the list of target regions. From the files obtained I do the statistics of methylation values using R and GNU 'datamash' utility.

```
#!/bin/bash

### EXTRACTION OF METHYLATION VALUES AT GIVEN REGIONS ###

# example: /home/devel/heath/code/combine_cpg/combine_cpg sample.txt -r chr1:11369-12369 -M | python parse_combine.py > /scratch/devel/sgarcia/put_proms_meth/*.txt

# 1. Extract methylation
PP=$1
head ${PP}
cat ${PP} | python make_combine_pp.py  > make.combine.pp.sh & chmod +x make.combine.pp.sh
./launch.py mkcomb "./make.combine.pp.sh"

# obtain a file per region with all CGs and its methylation state

# 2. Statistical description of methylation state
Rscript put_proms_meth_means.R
cat put.prom.meth.means.txt | sed '1d' | awk -F'\t' 'BEGIN{OFS="\t"}{print $2}' | datamash median 1
```

Listing 3: extract_meth_pp.sh

B.2 Target regions pipeline

The pipeline needs three files as standard input: i) H3K4me3 peaks in `.bed` format, ii) DNase hotspots in `.bed` format, iii) regions of TSSs from gencode annotation in `gtf` format. Given these three files, the script intersects and creates a file with the list of target regions in bed format (tab-separated columns file; chr start end peak_ID).

```
#!/bin/bash
PEAKS=$1
DNAse=$2
TSS=$3
CMD="bedtools closest -d"
CURRENT=$(pwd)

#1. Intersect H3K4me3 peaks and DNase hotspots
${CMD} -a ${PEAKS} -b ${DNAse} | awk -F'\t' 'BEGIN{OFS="\t"}{ if ($NF==0) print $1,$2,$3,$4}' | sed -e "s/[0-9]*.macs2_peak_call_//" | uniq > tmp

#2. Intersect target regions with annotated TSSs
${CMD} -a tmp -b <(awk -F'\t' 'BEGIN{OFS="\t"}{print $1,$5,$6,$7}' ${TSS})| awk -F'\t' 'BEGIN{OFS="\t"}{print $1,$2,$3,$4,$8,$9}' > $(basename $CURRENT).closest.tss.putative.promoters.txt

#3. Select target regions >10k far from annotated tss
cat closest.tss.putative.promoters.txt | awk -F'\t' 'BEGIN{OFS="\t"}{ if ($NF>10000) print $1,$2,$3,$4}' > $(basename $CURRENT).putative.promoters.txt

#4. Message if file is created with exit
echo "List of Putative promoters file for DONOR_ID "$(basename $CURRENT)" created and saved at: " $CURRENT/$(basename $CURRENT).putative.promoters.txt
```

Listing 4: target_regions_pipeline.sh

B.3 RNA-seq retrieval

From the `Bam` file data is accessed with a script created previously by my group called `check_rna_seq.pl`. Target regions given as input in format chr1:start-end. The output of each region is saved in a file named as the region name. R scripts are then used to plot the distribution of the values across regions.

```
#!/bin/bash
### LOOK RNA SEQ AT GIVEN REGIONS ###
PP=$1
for chrom in $(seq 1 22) ; do  n=chr$chrom; echo $n; awk -v c=$n '{if ($1==c) print $0}' ${PP} > $(basename $1 .txt).$n.txt ;done
mkdir /scratch/devel/sgarcia/low_dim/

#1. Look at rna-seq coverage
for f in $(seq 1 22); do ./launch.py chr$f "./check_rna_seq.pl < ./putative_promoters/$(basename $1 .txt).$n.txt" ; done

#2. Do a summary of the rna seq data files obtained from rna seq
for f in /scratch/devel/sgarcia/low_dim/*/*.txt; do echo -n $f $'\t'; cat $f | datamash min 3 q1 3 median 3 q3 3 max 3 ; done > summary.rna.seq.*.txt

#3. Run R script to take the q3 from file created and plot
Rscript summay_put_proms.R

#4. Select regions with nreads>10
Rscript rna_seq_10k_putative_promoters.R
head 10k.putative.promoters.rna.expr.txt
```

Listing 5: rna_seq_target_regions.sh

C Data retrieval

Download the data needed for the analysis.

a) Index file is the Blueprint metadata file with all available information for each sample. Select by DONOR_ID the wanted samples to download and create a sub-setted table.

```
#!/usr/bin/env Rscript

### SELECT THE DATA I WANT TO DOWNLOAD ###

# Open Blueprint data index file
index <- read.delim("20150820.data.index",sep="\t", header=
    TRUE,stringsAsFactors = F)

# Subset donors I want to study
my_donors <- c("N00031319896021","C0066P", "C005VG","C001UY",
    "S001MJ")

my_samples_index <- index[which(index$SAMPLE_BARCODE %in%
    my_sample_barcodes | index$DONOR_ID %in% my_donors),]

# Subset columns I am interested in
my_samples_index <- data.frame(my_samples_index$SAMPLE_NAME,
    my_samples_index$SAMPLE_DESCRIPTION, my_samples_index$
    DONOR_ID,
my_samples_index$LIBRARY_STRATEGY, my_samples_index$
    EXPERIMENT_TYPE, my_samples_index$FILE, my_samples_index$
    TISSUE_TYPE)

colnames(my_samples_index) <- gsub("my_samples_index.","",
    colnames(my_samples_index))

my_samples_index <- my_samples_index[order(my_samples_index$
    SAMPLE_DESCRIPTION, my_samples_index$LIBRARY_STRATEGY),]
my_samples_index <- my_samples_index[grep("*.bed.gz",
    my_samples_index$FILE),]

```

```r
# Subset data I am interested in

my_samples_index <- my_samples_index[which(my_samples_index$EXPERIMENT_TYPE=="Chromatin Accessibility" |
my_samples_index$EXPERIMENT_TYPE=="H3K4me3" |
my_samples_index$EXPERIMENT_TYPE=="H3K27ac" |
my_samples_index$EXPERIMENT_TYPE=="DNA Methylation" |
my_samples_index$EXPERIMENT_TYPE=="mRNA-Seq"|
my_samples_index$TISSUE_TYPE=="Venous Blood"),]

# Write table in which last column is the link of file to download
write.table(my_samples_index,"./indexes/my_index_files.txt", sep="|", col.names = T, row.names=F, quote=F)
```

Listing 6: index_samples.R

b) From the table created, use the last column (link information) to download from the ftp site the files using GNU 'wget' utility.

```bash
INDEX=$1
cat ${INDEX} | awk $'{print "ftp://ftp.ebi.ac.uk/pub/databases/"$NF}' | while read item; do echo $item; wget -i $item; done
```

Listing 7: download_data.sh

D Description of the data

Table D.1: DNAse hotspots description. Information of sample, cell type, number of open chromatin regions along the genome, and median length of the regions for 28 samples with available data of open chromatin regions.

Sample	Cell_type	Number_regions	Median_length
C0010K46	Monocyte	194691	289
C001UY46	Monocyte	224435	297
C004084E	Monocyte	131956	275
C005PS4E	Monocyte	123652	281
C005VG45	Macrophage	82446	264
C0066P44	T cell	104198	286
C006NS47	CD42+ megakaryocyte	103659	274
C006UE47	Macrophage	375005	304
DG-75_d01	Bcell Lymphoma	93476	290
KARPAS-422_d01	Bcell Lymphoma	90741	285
S005FH41	AML	231471	289
S00BXV41	Macrophage	178108	287
S00BYT41	Macrophage	144526	286
S00C0J41	Macrophage	343325	300
S00CR241	Macrophage	243457	295
S00CS041	Macrophage	210165	295
S00CTZ41	Macrophage	284094	296
S00EPZ41	Monocyte	81454	284
S00HRJ41	Macrophage	74632	273
S00HSH41	Macrophage	111908	285
S00HTF41	Macrophage	123328	276
S00JPF41	Monocyte	211288	294
S00JQD41	Macrophage	100614	282
S00JRB41	Macrophage	101774	276
S00JS941	Macrophage	167006	275

SU-DHL-5_d01	Bcell Lymphoma	141531	288
U-266_d01	Multiple myeloma	221105	293
Z-138_d01	Bcell Lymphoma	232077	296

Table D.2: H3K4me4 peaks feature description. Sample, cell type, number of regions and median length of peaks for 80 samples vailable ChIP-seq data

Sample	Cell type	Number_peaks	Median_length
BL-2_c01	Sporadic Burkitt lymphoma	56095	550
C000S5H2	monocyte	41097	532
C0011IH1	monocyte	42247	533
C001UYH2	monocyte	31580	741
C00264H1	monocyte	27533	702
C002Q1H1	T cell	25377	629
C002TWH2	central memory T cell	20867	407
C002YMH1	T cell	37296	616
C003UQH1	effector memory T cell	18866	565
C004GDH1	neutrophil	44475	380
C00504H1	Natural Killer	19512	446
C0054XH3	effector memory T cell	23448	636
C005DFH1	T cell	27125	539
C005VGH1	macrophage	32554	668
C0062XH1	Natural Killer	21024	632
C0066PH1	T cell	35255	595
DG-75_c01	Sporadic Burkitt lymphoma	45302	640
JVM-2_c01	lymphoma	68432	540
KARPAS-422_c01	lymphoma	34447	714
S000RDH1	T cell	24993	551
S000RDH2	monocyte	37338	649

S0018AH1	T cell	26154	488
S0018AH2	macrophage	34367	556
S00198H2	macrophage	27419	486
S002R5H1	erythroblast	24145	768
S002S3H1	erythroblast	19778	723.5
S005CNH1	myeloid cell	29841	716
S005EJH1	myeloid cell	15718	360
S005FHH1	myeloid cell	18982	352
S005YGH1	Natural Killer	24249	616
S007SKH1	macrophage	32966	593
S008QKH1	myeloid cell	44417	322
S00BHQH1	macrophage	26872	685
S00BJMH1	endothelial cell of umbilical vein (proliferating)	28010	784
S00BJMH2	endothelial cell of umbilical vein (resting)	33205	639
S00BXVH1	macrophage	46690	478
S00BYTH1	macrophage	46419	491
S00C2FH1	T cell	28105	623
S00DCSH1	endothelial cell of umbilical vein (proliferating)	29848	721
S00DCSH2	endothelial cell of umbilical vein (resting)	24987	882
S00DFMH1	lymphocyte of B lineage	36041	804
S00FXFH1	neutrophil	26409	1078
S00FYDH1	neutrophilic myelocyte	36505	564
S00G11H1	neutrophil	40556	497
S00G3YH1	neutrophil	40679	463
S00JGXH1	neutrophil	41451	779
S00JHVH1	neutrophil	40246	401

S00JJRH1	neutrophilic myelocyte	35259	514
S00JMLH1	neutrophil	27680	525
S00JNJH1	neutrophil	32023	470
S00K5EH1	neutrophil	40730	544
S00K6CH1	neutrophil	36380	806
S00K7AH1	neutrophil	38884	694
S00K88H1	neutrophil	35429	731
S00NJBH1	monocyte	36240	993.5
S00NK9H1	macrophage	46395	825
S00NM5H1	macrophage	42403	874
S00NN3H1	macrophage	40881	970
S00T2LH1	macrophage	43327	574
S00VDSH1	neutrophilic metamyelocyte	33937	821
S00VEQH1	neutrophil	34474	798
S00VFOH1	neutrophil	33661	696
S00VHKH1	megakaryocyte	36255	848
S00VKEH1	B cell	41084	579
S00W0DH1	B cell	23281	941
S00W1BH1	B cell	36488	808
S00X9SH1	B cell	37124	783
S00XAQH1	B cell	42596	702
S00XCMH1	B cell	50483	755
S00XDKH1	B cell	53791	484
S00XXHH1	myeloid cell	34640	739
S00Y7SH1	B cell	36021	790
S00Y8QH1	B cell	45159	593
S00Y9OH1	B cell	44360	679
S010NDH1	thymocyte	22971	720
S010R5H1	thymocyte	26370	908.5
S013N1H1	mononuclear cell	42110	603

SU-DHL-5_c01	lymphoma	45955	573
U-266_c01	Multiple myeloma	65314	538
Z-138_c01	lymphoma	46931	493

Table D.3: Information of Donor ID, sample name, description of the cell type, and the epigenomic feature for five samples with all the information needed.

DONOR ID	SAMPLE NAME	LIBRARY STRATEGY
C001UY (CD14-positive, CD16-negative classical monocyte)	C001UYB4	DNA-Seq
	C001UYA3bs	Bisulfite-Seq
	C001UY46	DNase-Hypersensitivity
	C001UYH2	ChIP-Seq
C005VG (Macrophage)	C005VG11	RNA-Seq
	C005VG45	DNase-Hypersensitivity
	C005VG51	Bisulfite-Seq
	C005VGH1	ChIP-Seq
C0066P (CD8-positive, alpha-beta T cell)	C0066P44	DNase-Hypersensitivity
	C0066P51	Bisulfite-Seq
	C0066PH1	ChIP-Seq
S001MJ (inflammatory macrophage)	S001MJ12	RNA-Seq
	S001MJ48	DNase-Hypersensitivity
	S001MJ51	Bisulfite-Seq
	S001MJH1	ChIP-Seq
N00031319896021 (macrophage - T=6days untreated)	S00BXV11	RNA-Seq
	S00BXV41	DNase-Hypersensitivity
	S00BXVH1	ChIP-Seq