School of Medicine
University of Crete

Analysis of NGS data for disease understanding

Master of Science in Bioinformaticstics

Danae Yiannakou, 2022

Acknowledgements

I dedicate this work to my family, Nikos, Sylvana, Christos, Vanessa and Pavlos. I must convey my profound gratitude to them for providing me with unfailing support and continuous encouragement throughout my years of study. This would not have been possible without them. Thank you. I would like to thank my supervisors Dr Pavlos Pavlidis and Dr Maria Grigoriou. They consistently allowed this research to be my own work, but steered me in the right direction whenever I needed it. I am also grateful to my friends who have supported me along the way, Kyriakos, Dimitra, Marialena, Chryso, Maria, Stephanie and Constantina. Thank you for all your encouragement!

Abstract

Systemic lupus erythematosus (SLE) is a systemic autoimmune disease facilitated by aberrant immune responses directed against cells and tissues, resulting in inflammation and organ damage. The majority of cells involved in the pathogenesis of SLE originate from bone marrow (BM) derived haematopoietic stem and progenitor cells (HSPCs). Previous research suggests that in SLE, haematopoiesis is dysregulated, with a skewing toward the myeloid lineage at the expense of lymphopoiesis. Also, HSPCs acquire a primed phenotype with a "trained immunity" signature, which may contribute to inflammation and flare risk. However, whether the epigenetic profile is implicated in this particular dysregulation, is still unknown. DNA methylation can regulate gene expression by inhibiting the binding of transcription factors (TFs) to DNA. Such TFs can be cis-regulatory elements (CREs) that typically regulate gene transcription by binding to other transcription factors as CCCTC-binding factor (CTCF). Herein, we show that differentially expressed genes (DEGs) and more specifically genes that are involved in myeloid-related pathways are possible regulated by distant methylated regulatory factors. We found that CREs, specifically enhancers that are up to 2Mb away from DEGs are methylated, while other transcription factors in those distances such as CTCF binding sites are also altered by methylation. Here, we show these modifications along with the expressions of DEGs and at the same time, we indicate the significance of their relationship. The findings suggest that interactions of distant methylated CREs and TFBS with DEGs cause HSPCs to reprogramme towards myeloid lineage, which may contribute to increased immunological responses and flares in SLE.

Contents

Acronyms

BM bone marrow. 4

cDNAs complementary DNAs. 6

ChIP-seq chromatin immunoprecipitation followed by sequencing. 40

CLPs common lymphoid progenitors. 3

CMPs common myeloid progenitors. 3

CREs Cis-regulatory elements. 7

CTCF CCCTC-binding factor. 8

DEGs differentially expressed genes. 9

DMRs differentially methylated regions. 9

DR-DM downregulated-hypomethylated. 12

DR-UM downregulated-hypermethylated. 12

FDR False Discovery Rate. 10

GMPs granulocyte-macrophage. 3

HSCs haemopoietic stem cells. 2

HSCT haemopoietic stem cell transplant. 1

HSPCs Hematopoietic stem and progenitor cells. 3

log2FC log two fold change. 10

LT-HSCs long term HSCs. 3

MDS Multidimensional Scaling. 10

MEPs megakaryocyte-erythrocyte progenitors. 3

MPP multipotent progenitor cell. 3

NGS next-generation sequencing. 5

NZB New Zealand Black. 1

NZW New Zealand White. 1

PCA Principal Component Analysis. 10

RRBS Reduced representation bisulfite sequencing. 7

SLE Systemic lupus erythematosus. 1

snRNA small nuclear RNA. 5

ST-HSCs short term HSCs. 3

TADs Topologically associating domains. 40

TFBS transcription factor binging sites. 7

TFs transcription factors. 3

tRNA transfer RNA. 5

TSS transcription start site. 8

UR-DM upregulated-hypomethylated. 12

UR-UM upregulated-hypermethylated. 12

VST variance stabilizing transformation. 10

WGBS Whole genome bisulfite sequencing. 7

1. Introduction

1.1 Systemic lupus erythematosus

Systemic lupus erythematosus (SLE) is a complicated autoimmune disease with a chronic relapsing–remitting history with a wide range of symptoms that can range from moderate to life-threatening illness. SLE develops clinically as a result of a combination of genetic susceptibility, environmental, immunological, and hormonal variables, with a high preference for women of reproductive age Basta et al. (2020). The exact molecular mechanisms that cause SLE clinical symptoms are yet unknown. In order to better understand these molecular and genetic mechanisms, various mouse models of spontaneous lupus have been used, including the classic mouse model, F1 hybrid of the New Zealand Black (NZB) and New Zealand White (NZW) strains, called NZB/W F1 Pathak and Mohan (2011). During the last two decades, research employing multiple mouse strains of spontaneous and inducible lupus has shed light on the immune system's role, including innate immunity and adaptive immunity in the disease's pathogenesis Pan et al. (2020). B and T lymphocytes have been shown to play a central role in adaptive immune response of SLE while the role of innate immune components has been only recently addressed and also found to play an important part in the disease Herrada et al. (2019); Pan et al. (2020). Now, we know that a complex network of innate and adaptive immune cells interactions occurs during SLE Herrada et al. (2019). On that note, it has been also shown that epigenetic changes in immune cells play a major role in the disease pathogenesis via gene expression dysregulation Hedrich et al. (2017).

SLE treatment is complicated and requires a multidisciplinary approach. The first line of defense is pharmacological treatment in the form of immune suppression Davis and Reimold (2017). Despite all of these treatment options, a significant number of patients continue to have high disease activity and relapse often, causing organ damage. Because of the inherent variety of illness causes, it has been challenging to develop medicines that work for the vast majority, in most if not all SLE patients Davis and Reimold (2017). New treatment options for people with severe or refractory illness are thus required. Over the last two decades, haemopoietic stem cell transplant (HSCT) has been tried in patients with SLE de Silva and Seneviratne (2019).

1.2 Stem Cells

Stem cells are in multicellular that can differentiate into many types of cells and multiply endlessly to produce additional stem cells. No other cell in the body has the natural ability to generate new cell types while in a cell lineage, they are the earliest form of cell.

- Embryonic stem cells: stem cells that come from embryos that are 3 to 5 days old. At this stage, an embryo is called a blastocyst and has about 150 cells. These are pluripotent stem cells, meaning they can divide into more stem cells or can become any type of cell in the body. This versatility allows embryonic stem cells to be used to regenerate or repair diseased tissue and organs.
- Adult stem cells: stem cells that are found in small numbers in most adult tissues, such as bone marrow, brain and skin. Compared with embryonic stem cells, adult stem cells have a more limited ability to give rise to various cells of the body. Transplantation of haemopoietic stem cells (HSCs), can now be used to treat most inherited blood cell diseases.

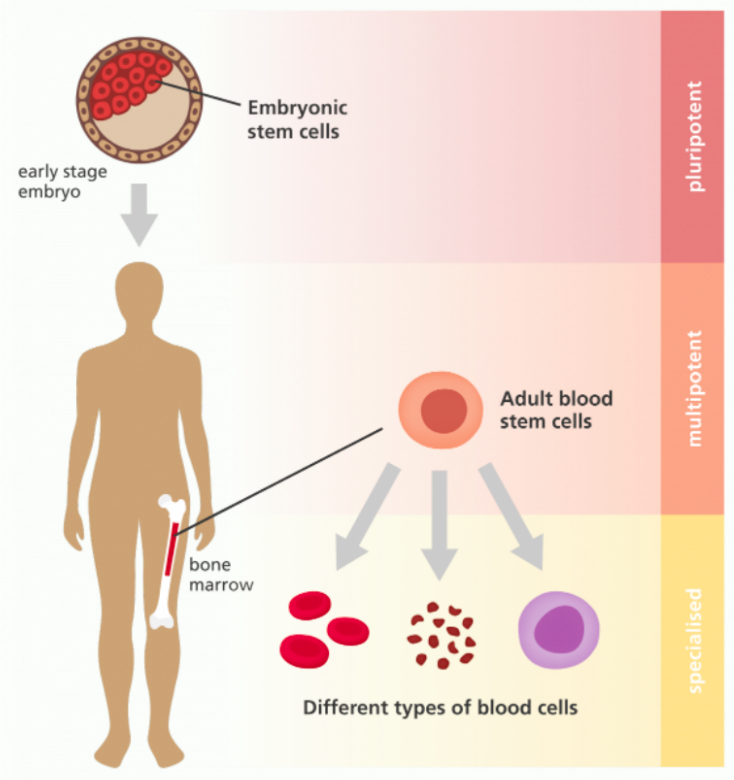


Figure 1.1: Different types of stem cell in the body (Credit: Genome Research Limited.)

1.2.1 Haematopoietic Stem and Progenitor Cells

Hematopoietic stem and progenitor cells (HSPCs) are a rare population of precursor cells that possess the capacity for self-renewal and multilineage differentiation Schulz et al. (2009). HSPCs are dormant cells that live in the BM niche that also proliferate and differentiate in response to stress or inflammation to replenish any progeny required Baldridge et al. (2011); Trumpp et al. (2010); Zhao and Baltimore (2015). It is also known that all peripheral blood cells, both the myeloid and lymphoid lineage, derive from HSPCs Suda et al. (2011); Gunsilius et al. (2001). The immunophenotype-based tree-like hierarchy model is the most well-known model for illustrating the link between an HSC and its progenies, as well as the stepwise differentiation process Kondo et al. (1997); Morrison et al. (1997); Akashi et al. (2000); Manz et al. (2002). The pool of HSPCs can be broken down into two subsets of cell types: the HSCs and the multipotent progenitor cell (MPP). HSCs can be separated into two subpopulations based on their CD34 expression: CD34 long term HSCs (LT-HSCs) and CD34+ short term HSCs (ST-HSCs). LT-HSCs are a rare, quiescent bone marrow population with full long-term ($> 3\sim4$ months) regeneration capacity, whereas ST-HSCs have only a short-term (usually 1 month). LT-HSCs differentiate into ST-HSCs, and subsequently, ST-HSCs differentiate into MPPs, which have no detectable self-renewal ability Yang et al. (2005). Subsequently, two distinct progenitor cell populations emerge from the MPP pool: the common lymphoid progenitors (CLPs) and the common myeloid progenitors (CMPs). The second branch point at CMPs segregates bipotent granulocyte-macrophage (GMPs) and megakaryocyte-erythrocyte progenitors (MEPs). CLPs further form T, B, NK and dendritic cells, while GMPs differentiate into granulocytes/monocytes and MEPs generate megakaryocytes/erythrocytes. All these populations form a tree-like and balanced hierarchy model, within which key transcription factors (TFs) and cytokines precisely conduct the stepwise differentiation of HSCs to mature blood cells Zhu and Emerson (2002); Robb (2007); Metcalf (2008); Zhang and Lodish (2008); Seita and Weissman (2010).

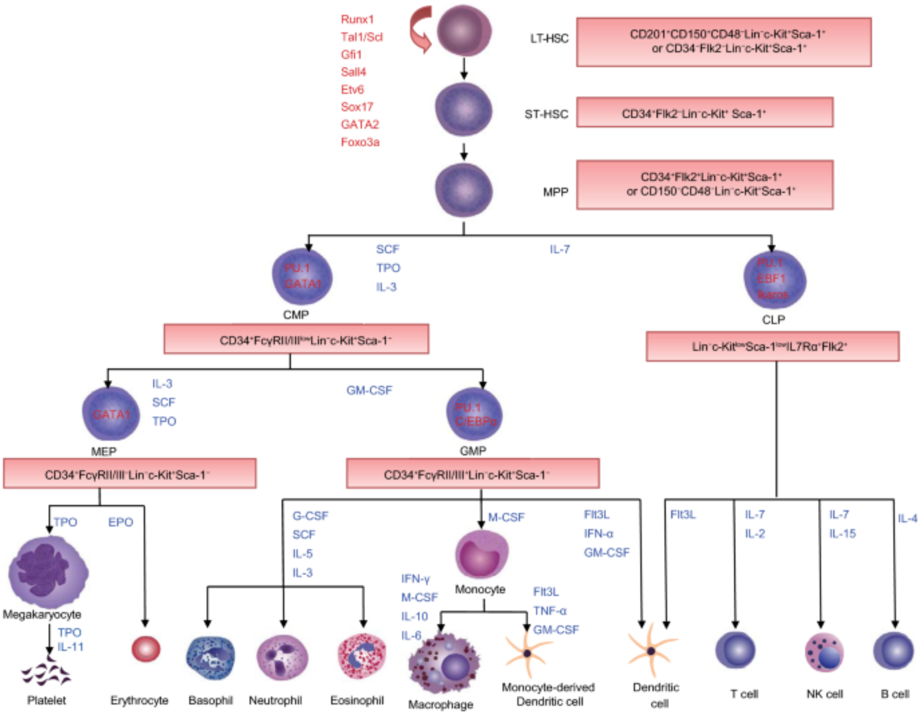


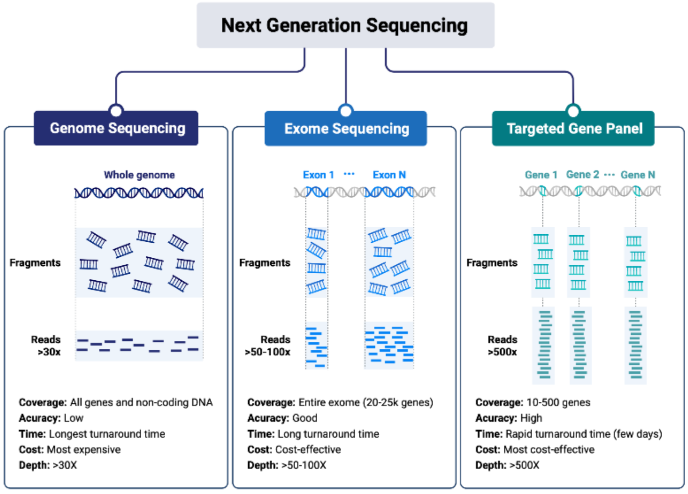
Figure 1.2: The immunophenotype-based tree-like hierarchy model. LT-HSCs differentiate through ST-HSCs and then MPP stages into lineage-restricted progenitors such as lymphoid, myeloid, or megakaryocyte/erythroid progenitor. Cheng et al. (2020).

1.2.2 Haematopoietic Stem and Progenitor Cells in SLE

The majority of cells involved in the pathogenesis of SLE come from bone marrow (BM) derived HSPCs which are the most primitive multipotent population that gives rise to all blood cell types King and Goodell (2011). As mentioned, lymphopoiesis and granulopoiesis are part of hematopoiesis, specifically derive from lympoid and myeloid lineage respectively. Lymphopoiesis is the process in which lymphocytes (B cells, T cells and NK cells) develop from progenitor cells while granulopoiesis leads to the production of granulocytes which are neutrophils, eosinophils and basophils. In previous studies, evidence was presented from gene exppression analyses, the upregulation of genes linked with cell death and granulopoiesis, providing further evidence of the apoptosis and granulocytes role in the pathogenesis Grigoriou et al. (2020). Specifically, there is evidence of haematopoiesis dysregulation in SLE, with a skewing toward the myeloid lineage which is associated with epigenetic tinkering, at the expense of lymphopoiesis, and priming of HSPCs. Additionally, this evidence exhibits a trained immunity signature which is mainly based on the epigenetic and metabolic reprogramming of cells, that may also contribute to inflammation and flare risk Itokawa et al. (2022).

1.3 Next Generation Sequencing Technologies

Sanger sequencing has been replaced by next-generation sequencing (NGS) as DNA sequencing technology has improved. NGS is a massively parallel sequencing technique that provides ultra-high throughput, scalability, and speed. The method is used to determine the nucleotide order of entire genomes or specific DNA or RNA portions. NGS has changed biology, allowing labs to undertake a wide range of applications and analyze biological systems at a level never before feasible. A full human genome can be sequenced in a single day using NGS, while the prior Sanger sequencing technology took more than a decade to complete the final draft Behjati and Tarpey (2013). NGS is an exceedingly flexible set of techniques and approaches and can be adapted to: (a) sequence alterations in the entire genome, exome, or any subset thereof (DNA), (b) provide DNA copy number information, (c) sequence the entire transcriptome (transcribed RNA) or any subset thereof, (d) identify translocations, (e) demonstrate gene expression levels and many more.



Template adapted from: Dr. Roshini Abraham
Clinical Immunologist at Nationwide Children's Hospital

Figure 1.3: Next Generation Techniques including: Genome Sequencing, Exome Sequencing and Targeted Gene Panel.

1.3.1 RNA sequencing

Gene expression is the process by which the information encoded in a gene is used to either make RNA molecules that code for proteins or to make non-coding RNA molecules that serve other functions such as transfer RNA (tRNA) and small nuclear RNA (snRNA). Gene expression acts as an "on/off switch" to control when and where RNA molecules and proteins are made and as a "volume control" to determine how much of those products are made. The process of gene expression is carefully regulated, changing substantially under different conditions. In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype, *i.e.* observable trait. In past years, hybridization-based approaches such as microarrays, were the most used solutions for gene expression profiling and DE analysis, thanks to their high throughput and relatively low costs Wang et al. (2009). Despite their widespread usage in quantitative transcriptomics, these approaches have several limitations Wang et al. (2009) Wang et al. (2009); Roy et al. (2011). The introduction of NGS has transformed transcriptomics, quickly establishing RNA-seq as the preferred approach for studying gene expression Wang et al. (2009) Shendure (2008). The standard workflow of an RNA-seq experiment goes as follows. The fragmented RNAs in the sample of interest are reverse-transcribed into complementary DNAs (cDNAs). The cDNAs are then amplified and subjected to NGS analysis. The millions of short reads generated can then be mapped onto a reference genome, and the number of reads aligned to each gene, referred to as "counts", provides a digital assessment of gene expression levels in the sample under inquiry.

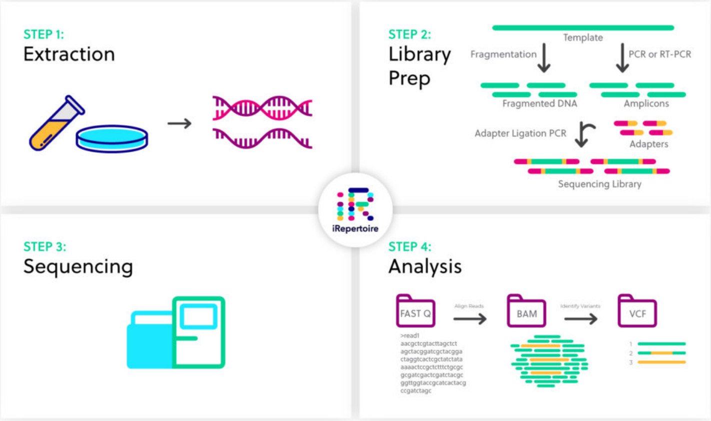


Figure 1.4: RNA sequencing workflow consists of four main steps: 1) extraction of RNA, 2) Sample preparation and library construction, 3) Next-generation sequencing of the library and 4) Bioinformatic analysis.

1.3.2 DNA methylation

Modifications to DNA that control whether genes are turned on or off are known as epigenetic alterations. These changes are made to DNA and do not alter the sequence of the DNA building blocks. A common type of epigenetic modification is called DNA methylation. DNA methylation involves the attachment of small chemical groups called methyl groups to DNA building blocks. DNA methylation regulates gene expression by recruiting proteins involved in gene repression or by inhibiting the binding of transcription factor(s) to DNA Moore et al. (2013). In the burgeoning field of epigenetics, there are several methods available to determine the methylation status of DNA samples, including bisulfite conversion, digestion with methylation-sensitive restriction enzymes, and antibody- or 5-methylcytosine binding protein–based purification of methylated DNA. The technique of bisulfite sequencing is considered to be the “gold standard” method in DNA methylation studies. The bisulfite treatment of DNA mediates the deamination of cytosine into uracil, and these converted residues will be read as thymine, as determined by PCR-amplification and subsequent Sanger sequencing analysis. 5 mC residues, on the other hand, are unaffected by this change and will continue to be interpreted as cytosine. When a Sanger sequencing read from an untreated DNA sample is compared to the same sample after bisulfite treatment, the methylated cytosines can be seen Kurdyukov and Bullock (2016). Bisulfite-conversion based sequencing techniques can be divided into three methods: (a) Whole genome bisulfite sequencing (WGBS), (b) Reduced representation bisulfite sequencing (RRBS) and (c) targeted bisulfite sequencing. For the purpose of this study, we will focus and elaborate more on the RRBS method. RRBS is a cost-efficient method for genome-wide DNA methylation profiling. Genomic DNA is first digested by a methylation-insensitive restriction enzyme (*e.g.*, BglII, MspI) and size selected to produce a small subset of the genomic DNA enriched for CpG sites in most of the promoters and CpG islands. Bisulfite conversion is performed and the sequencing library is constructed subsequently in order to be used in NGS analysis Baubec and Akalin (2016).

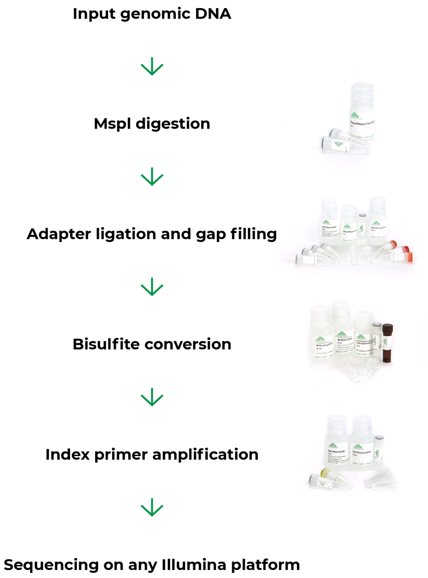
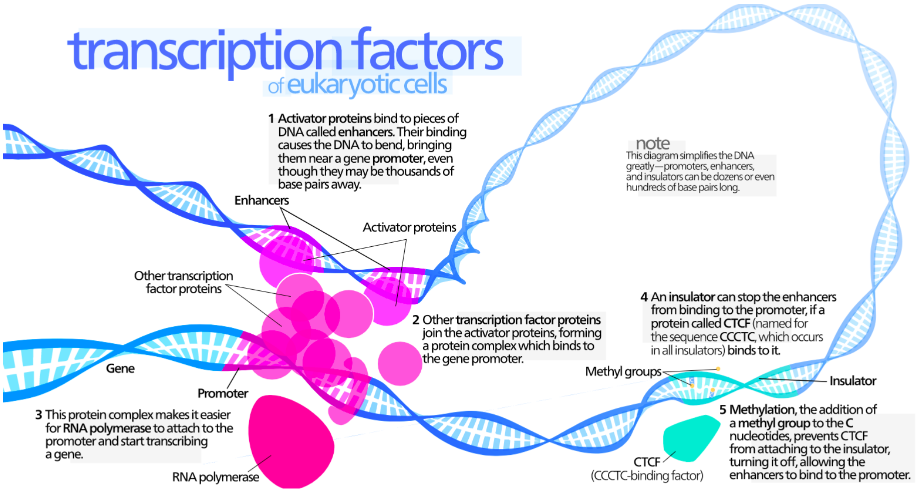

Input genomic DNA
MspI digestion
Adapter ligation and gap filling
Bisulfite conversion
Index primer amplification
Sequencing on any Illumina platform

Figure 1.5: Workflow of DNA methylation analysis, specifically RRBS technique.

1.4 Transcriptional regulatory factors

Transcription factors are proteins involved in the process of converting, or transcribing, DNA into RNA. One distinct feature of TFs is that they have DNA-binding domains that give them the ability to bind to specific sequences of DNA called transcription factor binging sites (TFBS). In the last two decades, it has been firmly demonstrated that epigenetic events can change the accessibility of DNA to transcription factors which can regulate gene expression profiles in immune cells contributing to the pro-inflammatory phenotype in SLE Itokawa et al. (2022); Hedrich et al. (2017). Cis-regulatory elements (CREs) such as promoters and enhancers, are also known to affect gene expression Wang et al. (2019). Promoters are DNA regions located within 1–2 kilobases (kb) of a gene's transcription start site (TSS) and include small regulatory elements (DNA motifs) required for the RNA polymerase transcriptional machinery to assemble (Doane and Elemento (2017)). However, without the participation of distant regulatory elements, such as the enhancers, transcription is negligible, in general. Enhancers are position-independent DNA regulatory elements that interact with site-specific transcription factors to determine cell type identity and control gene expression Doane and Elemento (2017). Additionally, previous studies revealed that enhancers can be found upstream, downstream, within the introns, or even relatively far away from the gene they regulate, up to 2 million base pairs away van Heyningen and Bickmore (2013). Enhancers can loop over extended genomic ranges to engage distant promoters, whereas promoters guide gene transcription in a position and orientation-dependent way Kim et al. (2015b). CREs typically regulate gene transcription by binding to TFs. CCCTC-binding factor (CTCF) is a highly conserved zinc finger protein, it is best known as a TF and its binding sites are found in intergenic regions Kim et al. (2007). It can function as a transcriptional activator, a repressor or an insulator protein, blocking the communication between enhancers and promoters Kim et al. (2015a).



note
This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be dozens or even hundreds of base pairs long.

1 **Activator proteins** bind to pieces of DNA called **enhancers.** Their binding causes the DNA to bend, bringing them near a gene **promoter,** even though they may be thousands of base pairs away.

2 Other **transcription factor proteins** join the activator proteins, forming a protein complex which binds to the gene promoter.

3 This protein complex makes it easier for **RNA polymerase** to attach to the promoter and start transcribing a gene.

4 An **insulator** can stop the enhancers from binding to the promoter, if a protein called **CTCF** (named for the sequence **CCCTC**, which occurs in all insulators) **binds to it.**

5 **Methylation,** the addition of a **methyl group** to the **C** nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.

Figure 1.6: A detailed explanation of how transcription factors of eukaryotic cells interact and regulate gene expression Wikipedia contributors (2022).

1.5 Purpose of the study

We asked if the HSPCs in the BM were responsible for the basic molecular abnormalities in SLE (genetic or epigenetic). To this end, we employed the NZBW/F1 mouse model of SLE to explore the transcription and epigenetic profile of HSPCs that could cause HSPCs to reprogramme towards myeloid lineage, which may contribute to increased immunological responses and flares in SLE.

In this study, we mainly investigate the epigenetic profile of murine lupus

HSPCs in two perspectives: (a) Inspection of differentially methylated regions (DMRs) in murine lupus HSPCs (b) Examination of the CREs and TFBS of differentially expressed genes (DEGs) as those involved in pathways of interest.

2. Materials & Methods

2.1 RNA sequencing pipeline

HSPCs were isolated from the bone marrow of NZB/W F1 mice using FACS ARIA III. RNA was extracted and libraries were generated using the Illumina TruSeq Sample Preparation kit v2. Single-end 75-bp mRNA sequencing was performed on Illumina NextSeq 500. Quality of sequencing was assessed using FastQC software Andrews (2010). Raw reads in fastq format were collected and aligned to the mouse genome (mm39 version) using STAR 2.7 algorithm Dobin et al. (2013). Gene quantification was performed using HTSeq Dobin et al. (2013); Putri et al. (2022) and differential expression analysis was performed using DESeq2 v3.15 Huber (2017) package in R R Core Team (2021). Genes with a False Discovery Rate (FDR) 0.05 and log two fold change (log2FC) 0.5 or -0.5 were considered statistically significantly up- and downregulated, respectively. BiomaRt package was used in order to get access into the Ensembl database and retrieve the coordinates and strand of DEGs. Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) plots were created using plotPCA from DESeq2 and ggplot2 tools correspondingly. Using a variance stabilizing transformation (VST) function the count data were transformed in a way that can be used for visualization, clustering or other machine learning tasks. Here we used VST and heatmaps with hierarchical tree clustering were designed. Finally, volcano and barplots were created in R with an in-house developed script which is based on the ggplot2 package.

2.2 DNA methylation analysis pipeline

After using FastQC software to evaluate the quality of sequencing, raw reads were trimmed with Trim Galore! (v0.6.4) using the options –rrbs, –illumina and the default value for –quality (Phred score: 20) Andrews (2010). This way, quality trimming will be performed first, and adapter trimming is carried in a second round. Further analysis of fastq files was performed with Bismark (v0.23.1) using the Bowtie2 dependency in three individual steps, genome preparation of mouse genome (mm39), alignment and methylation information extraction. The bismark.cov.gz files produced from the latter step were then used in R with the methylKit package for further analysis, specifically to detect DMRs. Here, the minimum coverage to read was set to 10 while an extra function was performed where it tiles the genome with a window and step size of 1000bp length and summarizes the methylation information Liu et al. (2020). Finally, for a region to be labeled as DMR, its corrected p value with SLIM method needed to be less than 0.05 and its absolute value of differential methylation level greater than 25% Li et al. (2021). Regions with value of differential methylation level >25% were considered hypermethylated and regions with value <25% were considered hypomethylated. Hierarchical clustering using default values of hclust and euclidean distance was performed followed by CpG methylation PCA analysis with autoplot and barplot with ggplot2 showing the average methylation percentage in each sample, all done on methylated regions and DMRs. Additionally, a volcano plot and a heatmap using ggplot2 and heatmap.2 tools in R, was built showing the hypomethylated and hypermethylated regions in each sample. Pie plots were designed using in-build functions of methylkit showing the percentage of differentially methylated regions overlapping with exon/intron/promoters but also the CpG island annotation. Finally, the percentage of hypo- and hyper- methylated regions per chromosome was shown with barplots using the ggplot2 tool in R.

2.3 Pipeline of combined analysis of RNA-sequencing and DNA-methylation

A custom script was built in R R Core Team (2021)combining the RNA sequencing and the DNA methylation results. Chipseeker was used to retrieve the information about the annotation of DMRs and their closest TSS Yu et al. (2015). To illustrate venn diagrams together with their statistical value, we used the tool Venn Diagram along with phyper Chen (2022). Subsequently, using 'bedtools closest', we retrieved the closest distances between DEGs and DMRs while for each distance a statistical test was done with pbinom to get the significance in their linkage Quinlan and Hall (2010). For each case of the lateral analysis, circos plots with OmicCircos were designed showing the expression of DEGs and the methylation of DMRs. Furthermore, to identify the methylated regulatory features of DEGs within the distances found significant from the last analysis, we constructed a function that combines regulatory features and DMRs. Specifically, biomaRt package was used in order to get access into the Ensembl database and retrieve the regulatory features of DEGs, while 'bedtools intersect' was needed to get the DMRs that overlap with the bound regions of those regulatory features Quinlan and Hall (2010); Durinck et al. (2009), Durinck et al. (2005), Quinlan and Hall (2010). All other plots were designed with ggplot2 and the pipeline is also able to produce some useful csv files in case of further analysis Wickham (2009).

2.4 Statistics

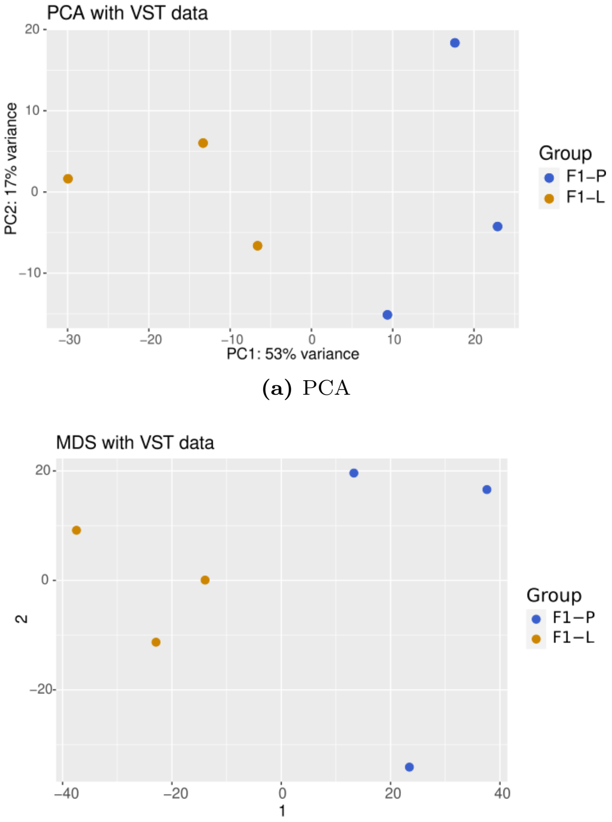
Statistical analyses were performed using hypergeometric (phyper) and probability distributions (pbinom). Hypergeometric test was used to model the association between DEGs and TSS of genes found closest to DMRs, while also for genes of interest and DEGs. Probability distribution-pbinom calculates the cumulative density function of the binomial distribution and it was used to calculate the probability of a variable X (relation, inverse relation) following a binomial distribution taking values greater than or equal to x (probability of success). We interpret inverse relation as the affiliation of DEGs with DMRs that are upregulated-hypomethylated (UR-DM) and downregulated-hypermethylated (DR-UM) while relation as the ones that are upregulated-hypermethylated (UR-UM) and downregulated-hypomethylated (DR-DM). Null hypothesis (H0): Results do not differ significantly from what is expected. Alternative hypothesis (H1): Results are significantly greater from what is expected. Setting a threshold of significance at 0.05 or 5%.

3. Results

The results are divided as follows: (a) we show the results from the RNA sequencing analysis where we indicate that most genes lean towards overexpression. (b) We present the results of methylation analysis and point out the annotation of DMRs where the majority of them are hypomethylated. (c) Additionally, we report the results from the combination analysis of DEGs and DMRs, starting off by inspecting if DEGs overlap with any of the genes where their TSS was found closest to DMRs, (d) evaluating the distances between DMRs and DEGs and (e) retrieving the methylated transcriptional regulatory factors for each DEG. (f) Finally, we show the results of the investigation of DEGs that are involved in pathways.

3.1 The transcriptional profile of F1-L demonstrates overexpression

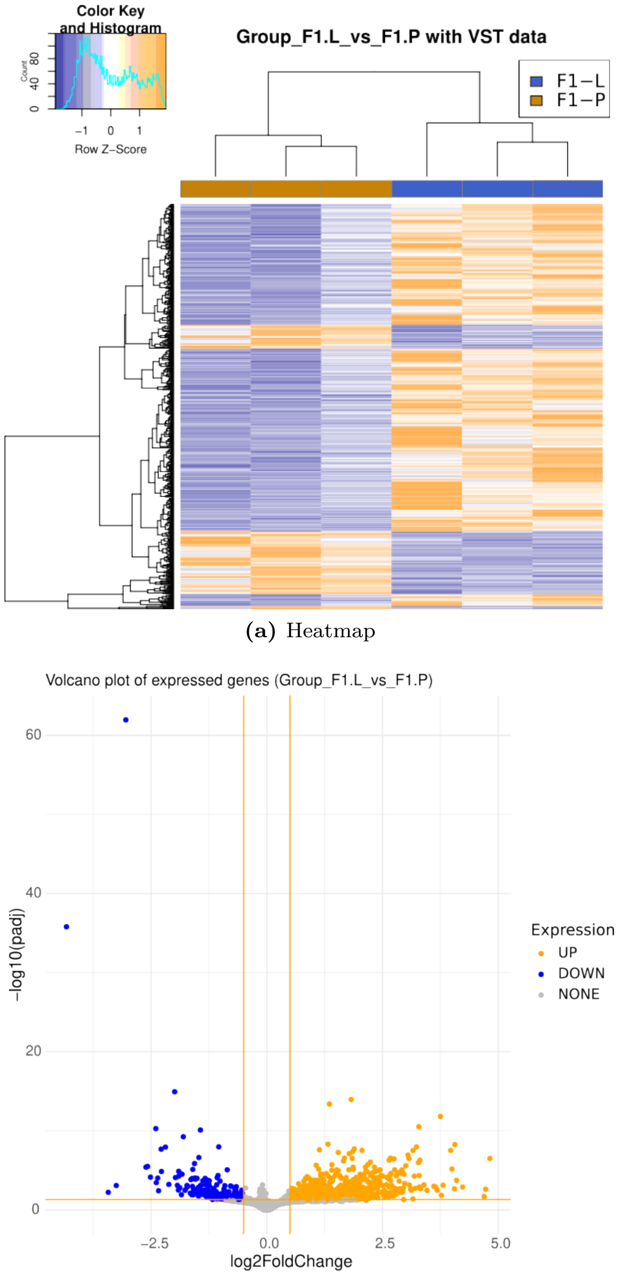
For this project we only need to retrieve information about DEGs such as their log2FC values and their coordinates in order to examine them together with DMRs and find any association. In order to study whether the transcriptional profile of HSPCs in SLE is altered, we used the spontaneous mouse model NZBW/F1 at two time points: preclinical stage (F1-P, n=6) and clinical stage (F1-L, n=4), defined as the point with proteinuria of >100 ng/dL. Gene profiling was performed in murine LSK compartment—representing HSPCs in mice—sorted by flow cytometry from BM of NZBW/F1. A gene is declared as differentially expressed when its corrected p value with FDR is <0.05 while the absolute value of log2FC is >0.5. PCA and MDS was performed and showed that the transcriptional profile in F1-L HSPCs differs from F1-P (Figure 3.1). A total of 809 DEGs between F1-P and F1-L were identified, of which 181 were found downregulated and 628 upregulated (Figure 3.2). Finally, the barplot in Figure 3.3 shows the distribution of log2FC values of DEGs across the chromosomes, which we can clearly see that overexpression prevails downregulation even at a chromosome level. Collectively, these data indicate that most DEGs in F1-L HSPCs are overexpressed.



(a) PCA

(b) MDS

Figure 3.1: a) Principal component analysis and b) Multidimensional scaling showing the differentiation of transcription profile in F1-L from F1-P.



(a) Heatmap

(b) Volcano

Figure 3.2: a) Heatmap of differentially expressed genes between F1-L and F1-P. Orange/blue gradient represents the row Z-score of overexpression/downexpression in F1-L compared to F1-P mice. b) Volcano plot of differentially expressed genes between F1-L and F1-P. The horizontal line indicates the threshold of corrected p value which is 0.05 while the two vertical lines pinpoint threshold of the absolute value of log2FC=0.5 (Left line=-0.5, Right line=0.5. Orange, blue and grey dots represent the upregulated, downregulated and the unchanged regions respectively.

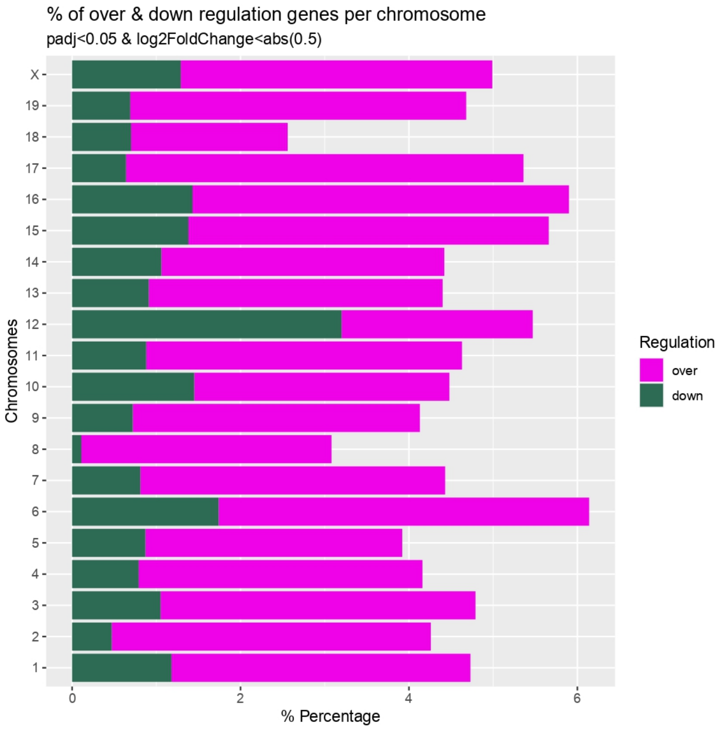
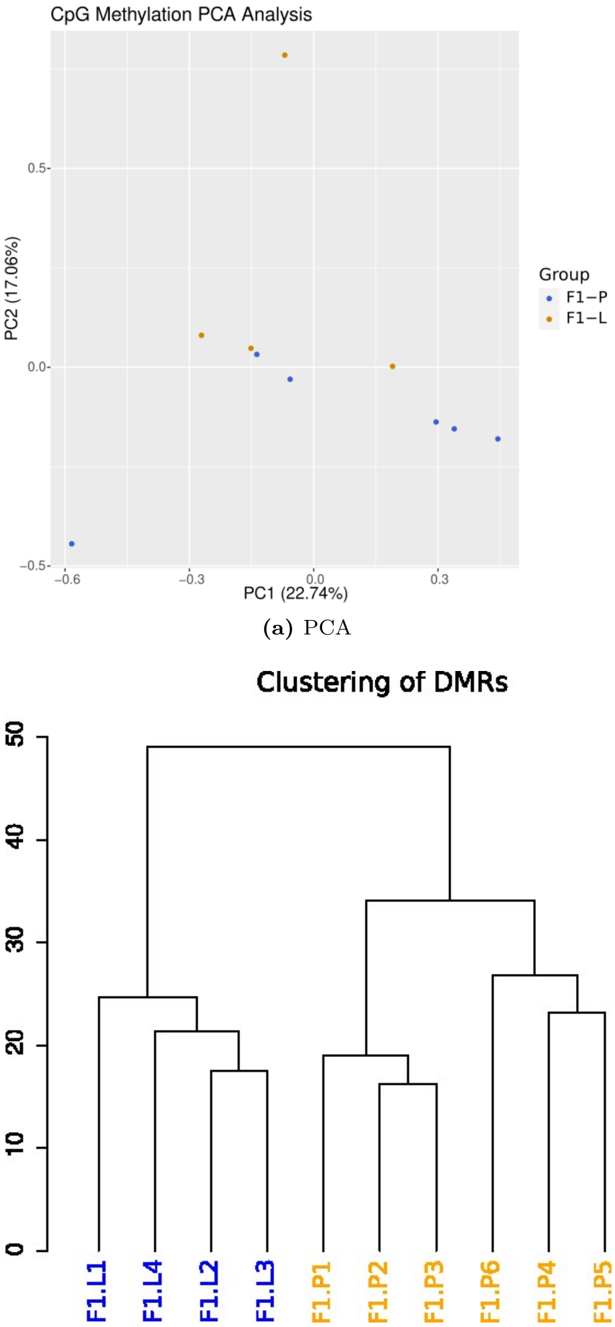


Figure 3.3: Distribution of differentially expressed genes accross each chromosome.

3.2 The epigenetic profile of F1-L demonstrates hypomethylation on CpG islands, introns, distal intergenic regions and promoters

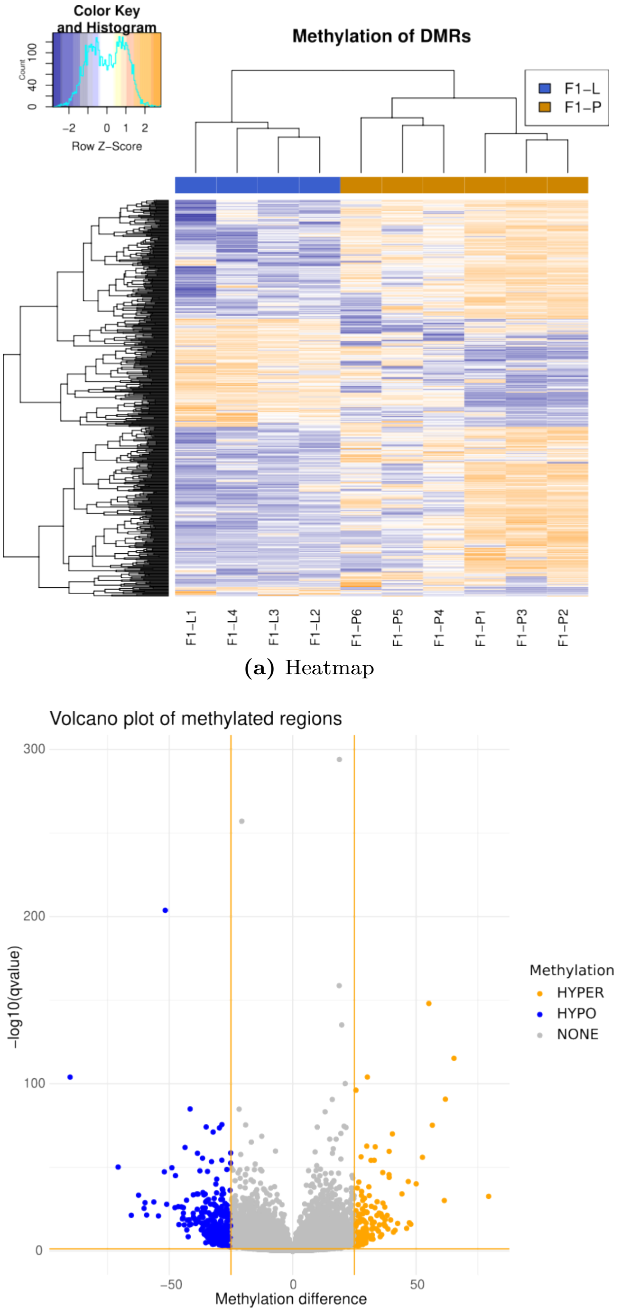
The alterations that occur in the epigenetic profile of F1-L HSPCs were studied using the same mouse model at two time points as mentioned above. PCA and clustering analysis using DMRs revealed that in F1-L HSPCs show different epigenetic profiles from F1-P (Figure 3.4). After summarizing the methylation information of the genome with window and step size of 1000bp length, 533 DMRs were identified using thresholds of 0.05 and 25% for the p value and the absolute value of differential methylation level respectively (Figure 3.5) Liu et al. (2020). Based on these results, we found that hypomethylated DMRs (377) prevails in relation to hypermethylated DMRs (156) in F1-L compared to F1-P. A similar outcome is also observed in the plot Figure 3.6 which represents the distribution of the methylated regions at each chromosome. Target regions are enriched for promoters, introns, intergenic regions and CpG Islands. Specifically, these results showed that 61% of DMRs are CpG islands while 22.89% were promoters, 30,02% were introns and 33.02% were intergenic regions (Figure 3.7). Some of those DMRs overlap in those regions which is clearly represented on the upset plot (Figure 3.8), where x-axis indicates the annotated regions and y axis indicates the number of DMRs that those regions share. This plot confirms that most methylated regions are intergenic and in fact are distal, meaning that they exist >3kb from the nearest TSS and some of them overlap with promoters. Also, many of the DMRs exist on introns and at the same time only few of them are on promoters, exons and 5'UTR regions. The distribution of DMRs relative to the TSS (Figure 3.9) shows that almost 80% of DMRs exist in regions >3kb upstream and downstream from TSS confirming the previous results, that a significant amount of DMRs are distal intergenic. Hence, we expect to find methylated CTCF binding sites and enhancers since we know that they exist within intergenic regions and introns respectively Kim et al. (2007), Park et al. (2014). Collectively, hypomethylation overcomes hypermethylation in HSPCs of F1-L mice, giving also that the majority of DMRs are on introns and distal intergenic regions, suggesting that we expect to find methylated enhancers and CTCF binding sites.



(a) PCA

(b) Clustering

Figure 3.4: a) PCA and b) Clustering using euclidean distance of DMRs showing the differentiation of epigenetic profile in F1-L from F1-P.



(a) Heatmap

(b) Volcano plot

Figure 3.5: a) Heatmap of differentially methylated regions between F1-L and F1-P. Orange/blue gradient represents the row Z-score of hypermethylation/hypomethylation in F1-L compared to F1-P mice. b) Volcano plot of differentially methylated regions between F1-L and F1-P. The horizontal line indicates the threshold of corrected p value which is 0.05 while the two vertical lines pinpoint threshold of the absolute value of methylation difference=25% (Left line=-25%, Right line=25%. Orange, blue and grey dots represent the hypermethylated, hypomethylated and the unchanged regions respectively.

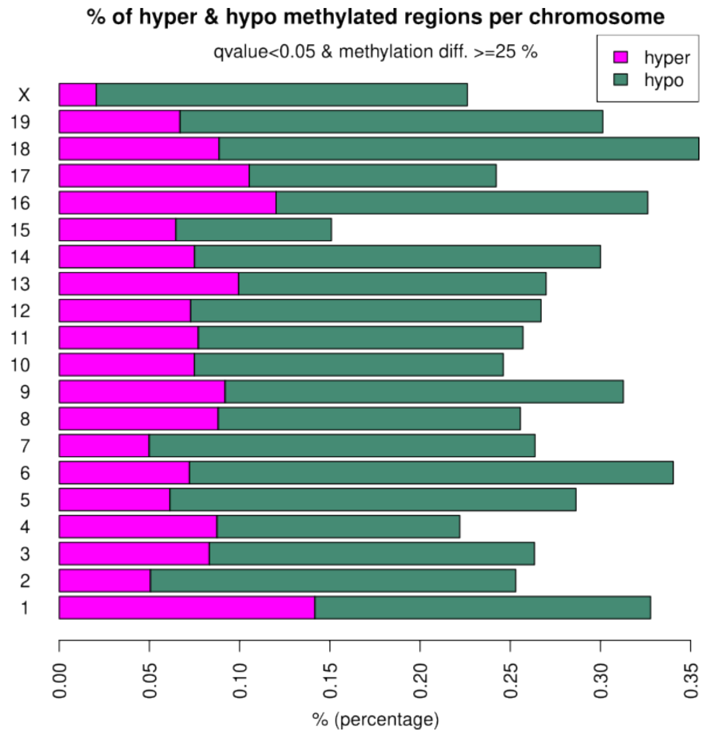
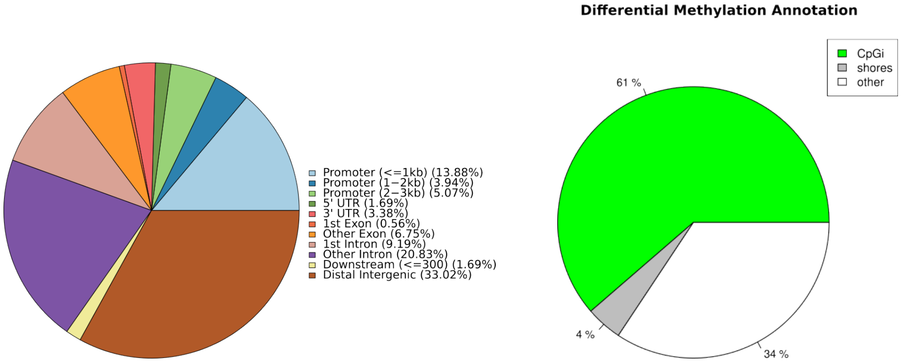


Figure 3.6: Distribution of differentially methylated regions accross each chromosome.



(a) CpG Island Annotation

(b) Annotation with exon/intron/promoters

Figure 3.7: a) CpG Island Annotation of DMRs and b) percentage of differentially methylated regions overlapping with exon/intron/promoters.

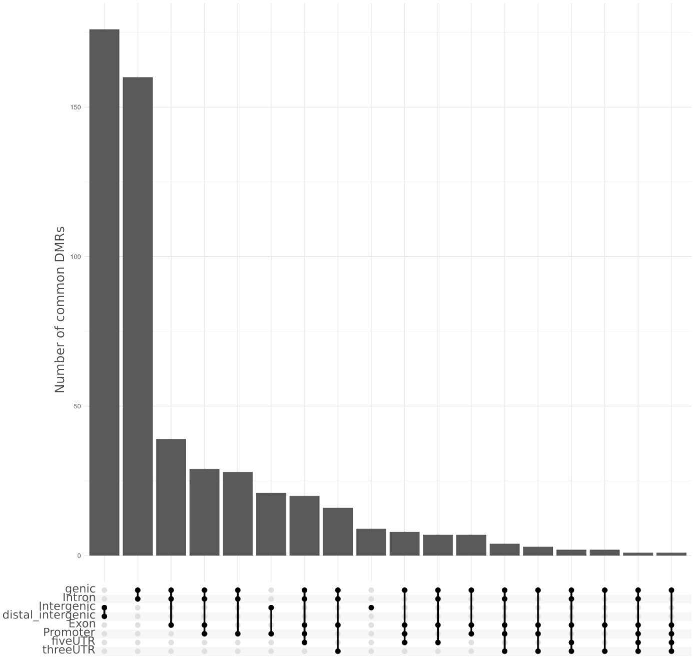


Figure 3.8: Upset plot. Number of common DMRs accross the annotated regions.

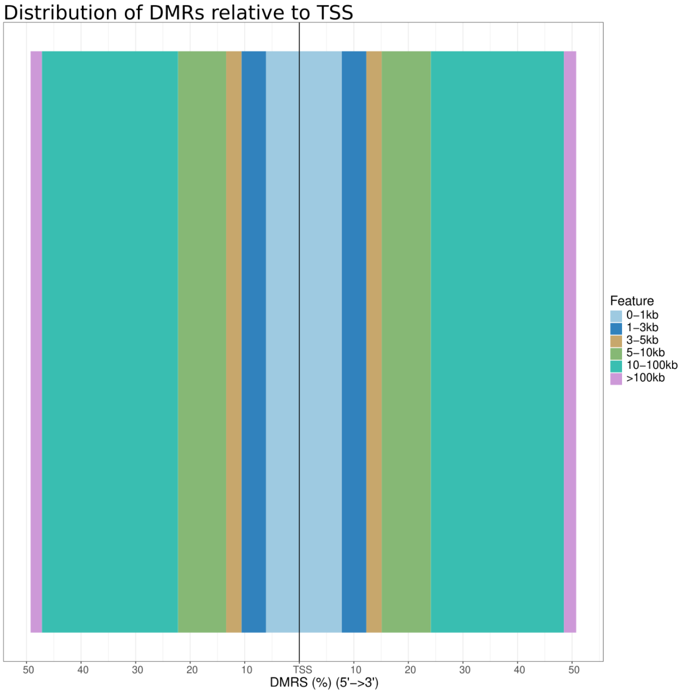


Figure 3.9: Distribution of DMRs relative to TSS. The colors indicate the distance of DMRs from their closest TSS.

3.3 Very few DEGs have near-by DMRs

Given that some DMRs are on promoter regions and some were found up to 3kb near to TSS known as core promoters, it was worth investigating the fact that DEGs may have close-by DMRs that possibly are also on their core promoters. In order to evaluate and examine this, a crossover was done between the DEGs and the genes that their TSS was found to be closest to the DMRs, extracting the common genes (Figure 3.10). Here, we report that there is no significance in the number of common genes found, concluding that the majority of DEGs does not have any near-by DMR. The dot plot shows the genes found common with their expression values along with their closest DMRs, specifying their position and methylation level, the distance from TSS and the annotation (Figure 3.11). Here it is worth mentioning that the *Ccnd1* and *Cebpe* myeloid marker genes were found upregulated which is consistent with a previous study in F1-L HSPCs Grigoriou et al. (2020). Moreover *Ccnd1* gene has an upstream hypomethylated distal intergenic region, while *Cepbe* has an upstream hypomethylated near-by exon. Together all these data, suggest that very few DEGs have near-by DMRs while the majority of the reported nearest gene found for each DMR, might play another role such as regulating other genes.

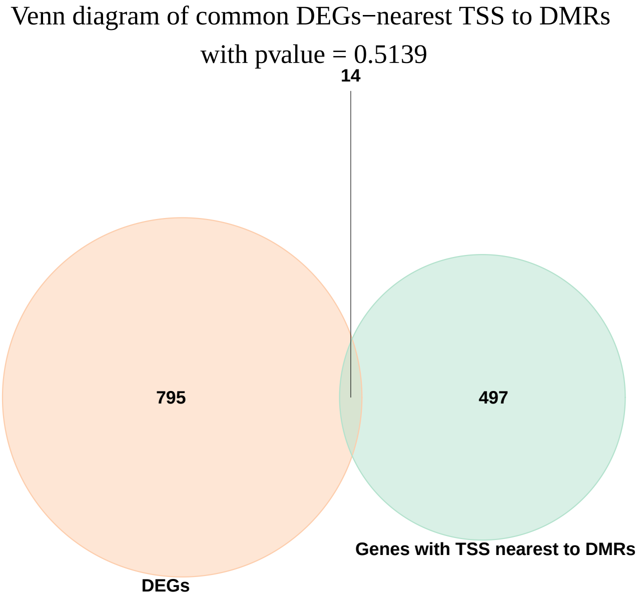


Figure 3.10: Venn diagram showing the common genes between DEGs and TSS closest to DMRs. The orange circle represents the number of DEGs while the green circle the number of TSS found closest to DMRs. The significance in the number of common genes(14) is reported with a p value of 0.5139.

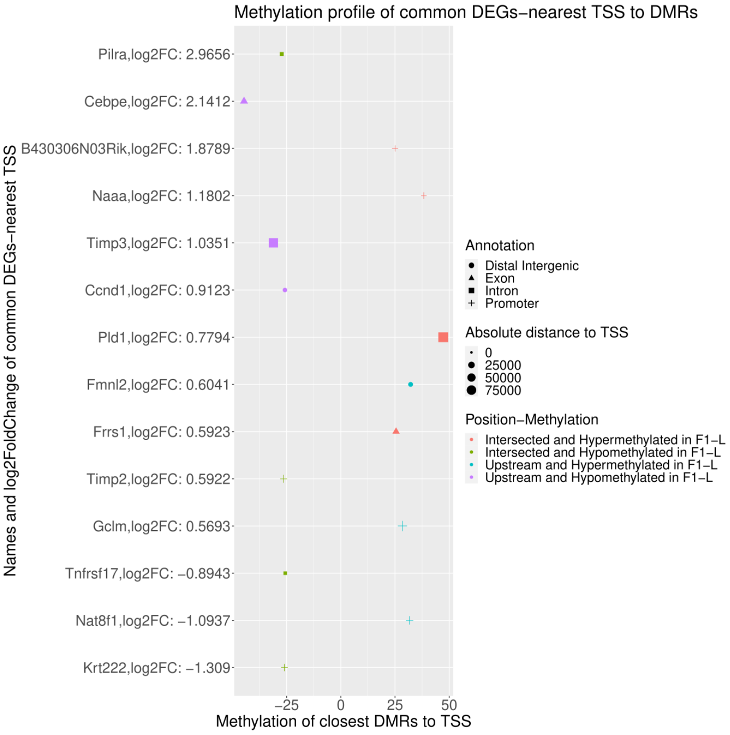


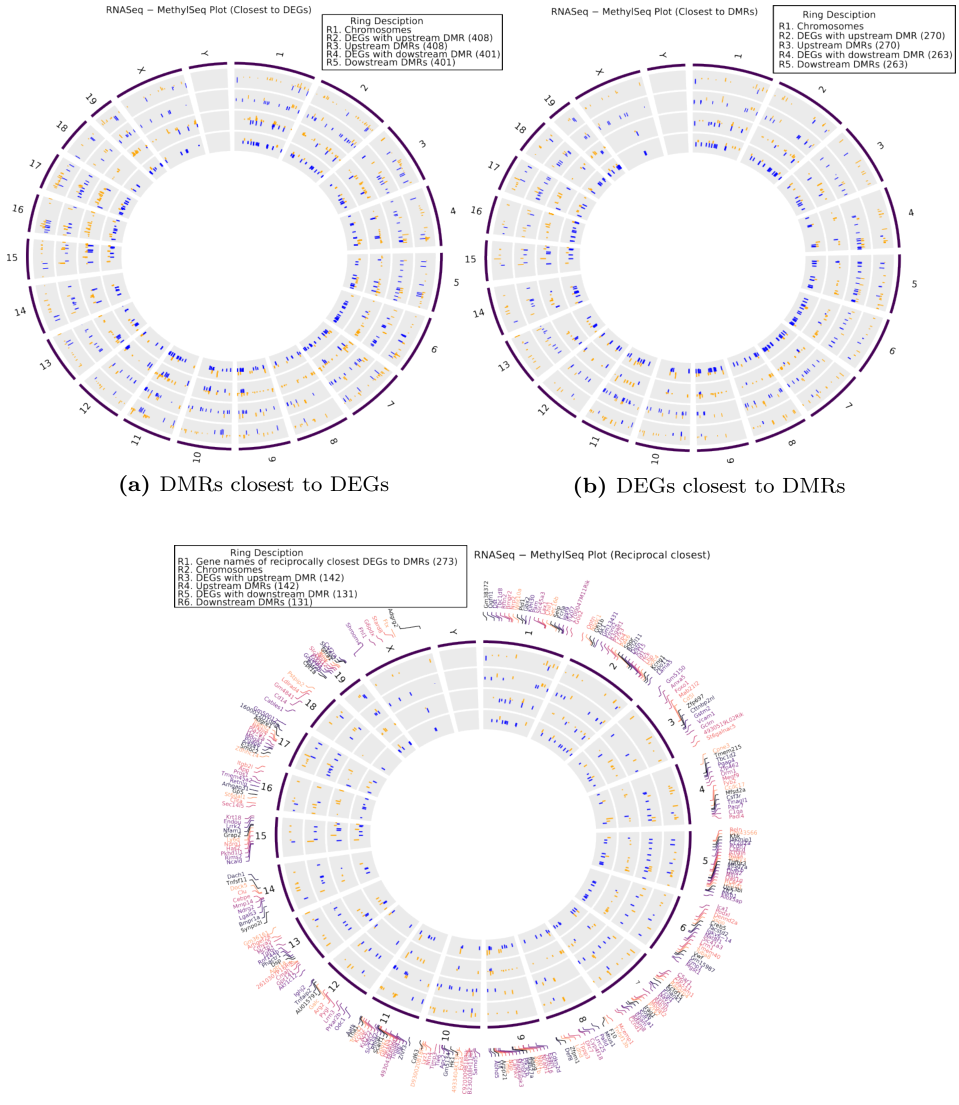
Figure 3.11: Dotplot where in its y-axis are the common gene names with their log2FC value in descending order(up to down) and in its x-axis are the methylation values of the DMRs found closest to each common gene. Each dot represents the closest DMR found, where the shape, size and colour indicate the annotation of each DMR, its distance and its position from the common gene respectively.

3.4 The distances between DMRs and DEGs indicate that DMRs are on CREs and TFBS of DEGs

Since barely any of the DEGs had close DMRs in a distance <3kb and also being on promoters, we next asked what are the distances between them and is there any significant linkage? To this end, we retrieve information in the distance which each DEG has a closest DMR (DMRs closest to DEGs), each DMR has a closest DEG (DEGs closest to DMRs) and the distance that it is shared by both of the cases above (Reciprocal closest). In all three cases, the downstream and upstream DMRs from DEGs were found.

In Figures 3.12a, b and c, are circos plots that represent all cases mentioned, showing the expression and methylation values for DEGs and DMRs respectively, that occur at each distance. Here, in Figure 3.12a and 3.12b the results show that multiple different DEGs have the same closest DMR and conversely, multiple DMRs have the same closest DEG, respectively. Figure 3.12c shows the distance that is shared by both of the cases above which is also the smaller one, meaning that for each DEG there is only one closest DMR and vice versa. Additionally, there is a strong contrast between the expression and epigenetic profile, but in order to confirm that, the cumulative density function (pbinom) was used to calculate the significance of the relationship between DEGs and DMRs at each distance (Table 3.1). Table 3.1 shows that there is a significant inverse relation, more specifically in the DR-UM affiliation in the downstreams at the distances of DMRs closest to DEGs and the reciprocal.

Given that the reciprocal case shows the smallest distance between DEGs and DMRs and were also found up to 17Mb, we suspected that DMRs may overlap with some CREs or TFBS of DEGs, hence we wanted to investigate their linkage at those distances. It is known that enhancers can be located up to 2 million base pairs away from the affected genes, thus we examined the relationship of DMRs found closest to DEGs specifically within the distances of 50kb, 100kb, 500kb, 1Mb and 2Mb van Heyningen and Bickmore (2013). In order to evaluate the significance of the relationship between DEGs and DMRs for each case, a statistical test analysis was done in the same way as mentioned above. Table 3.2 gives the p values for the relationship of multiple DMRs found closest at each DEG within the distances of 50kb, 100kb, 500kb, 1Mb and 2Mb with a statistical significance in the inverse relation, specifically those that are downstream and DR-UM in the last two distances. Figures 3.13a and b display the circos plots and show the expression values for each DEG along with their closest DMRs within the distances found significant together with their methylation status. Concluding, all these data suggest that CREs -enhancers and distant TFBS of DEGs are possibly methylated contributing in their expression.



(a) DMRs closest to DEGs

(b) DEGs closest to DMRs

(c) Reciprocal closest

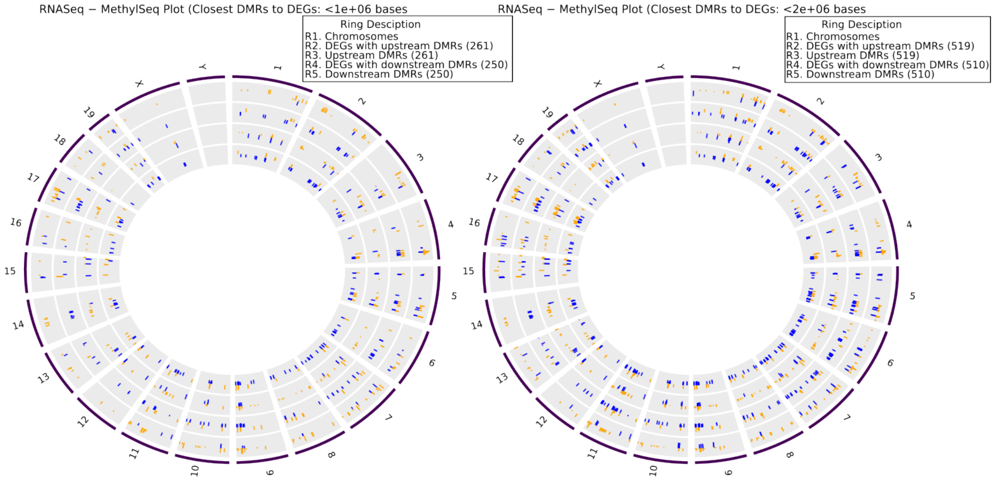
Figure 3.12: Circos plots representing the expression and methylation values of DEGs and DMRs at each distance (a) DMRs closest to DEGs, (b) DEGs closest to DMRs and (c) Reciprocal closest. In circos plot (c), ring one(R1) indicates the name of DEGs, where ring two to six (R2-R6) show the same information respectively in circos plots (a) and (b) having them as ring one to five (R1-R5). R2-R6 or R1-R5 show the expression values of DEGs that have upstream DMRs, the methylation values of upstream DMRs, the expression values of DEGs that have downstream DMRs and the methylation values of downstream DMRs at each case of distance.

	UR_DM	UR_UM	DR_UM	DR_DM
DMRs closest to DEGs (upstreams)	0.560	0.512	0.569	0.497
DMRs closest to DEGs (downstreams)	0.092	0.982	0.028	0.929
DEGs closest to DMRs (upstreams)	0.456	0.687	0.414	0.614
DEGs closest to DMRs (downstreams)	0.250	0.911	0.140	0.805
Reciprocal (upstreams)	0.188	0.954	0.089	0.873
Reciprocal (downstreams)	0.148	0.982	0.041	0.902

Table 3.1: P-values showing the significance of the relationship between DEGs and DMRs at each distance.

DMRs closest to DEGs below:	UR_DM	UR_UM	DR_UM	DR_DM
50kb(upstream DMRs)	0.534	0.831	0.466	0.822
50kb(downstream DMRs)	0.667	0.805	0.584	0.714
100kb(upstream DMRs)	0.382	0.873	0.313	0.801
100kb(downstream DMRs)	0.735	0.658	0.721	0.552
500kb(upstream DMRs)	0.421	0.768	0.357	0.677
500kb(downstream DMRs)	0.354	0.849	0.258	0.737
1Mb(upstream DMRs)	0.433	0.712	0.384	0.638
1Mb(downstream DMRs)	0.062	0.993	0.013	0.958
2Mb(upstream DMRs)	0.592	0.454	0.623	0.459
2Mb(downstream DMRs)	0.075	0.988	0.019	0.942

Table 3.2: P-values showing the significance of the relationship between DEGs and DMRs at each distance.



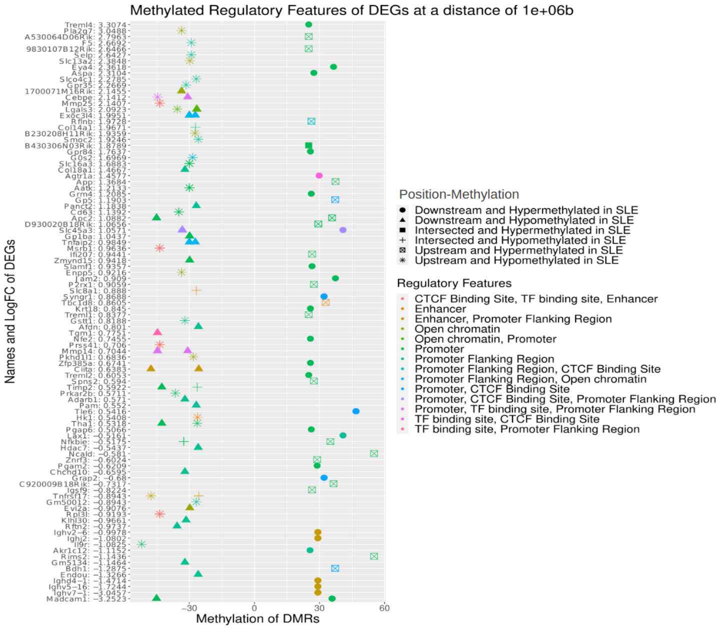
(a) DMRs closest to DEGs in a distance of 1Mb

(b) DMRs closest to DEGs in a distance of 2Mb

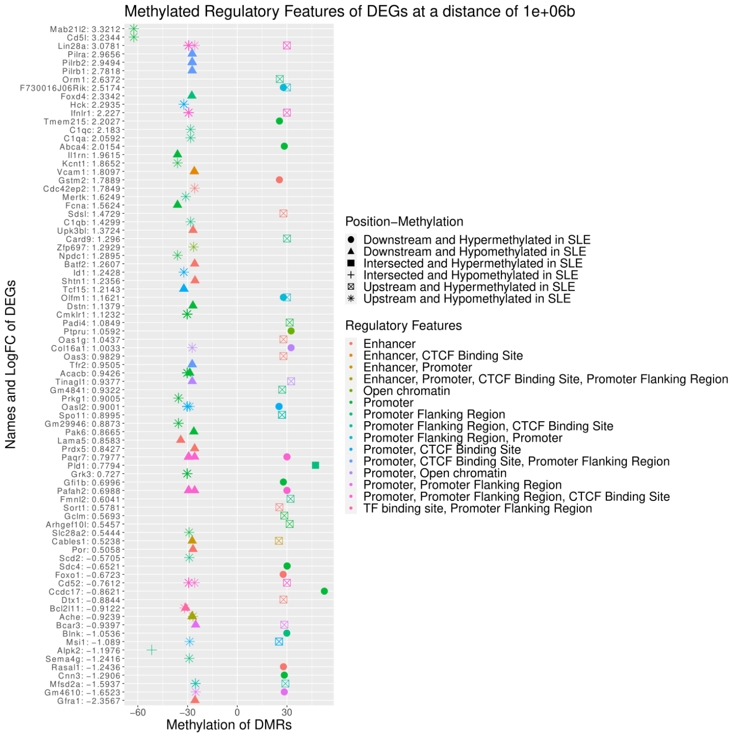
Figure 3.13: Circos plots representing the expression and methylation values of DEGs and DMRs respectively, at the cases of (a) DMRs closest to DEGs in a distance of 1Mb and (b) DEGs closest to DMRs in a distance of 2Mb. Ring one to five (R1-R5) show the expression values of DEGs that have upstream DMRs, the methylation values of upstream DMRs, the expression values of DEGs that have downstream DMRs and the methylation values of downstream DMRs at each case of distance.

3.5 Distant regulatory features of DEGs are methylated

To evaluate our last suggestion, that CREs and TFBS of DEGs could be methylated, we performed an overlapping analysis of those with the DMRs. We then retrieved the expression values of DEGs and methylation values of their closest regulatory features below the distances of 1Mb and 2Mb (Figures 3.14-3.15). Given that we have already shown that the majority of DMRs are not on core promoters of DEGs and that we expect that most of them would be on enhancers and CTCF binding sites, it is important to focus on the methylation status of this type of transcription factors. In order to find the significance of the relationship between DEGs and their methylated regulatory factors found in the distances of 1Mb and 2Mb, we performed a statistical analysis with pbinom where the results showed an importance in their inverse relation, specifically in the DR-UM (Figure 3.16). Collectively, these data indicate that the expression of DEGs, specifically downregulated, is probably mediated by the hypermethylation that occurs in their distant CREs-enhancers and TFBS, especially, CTCF binding sites.

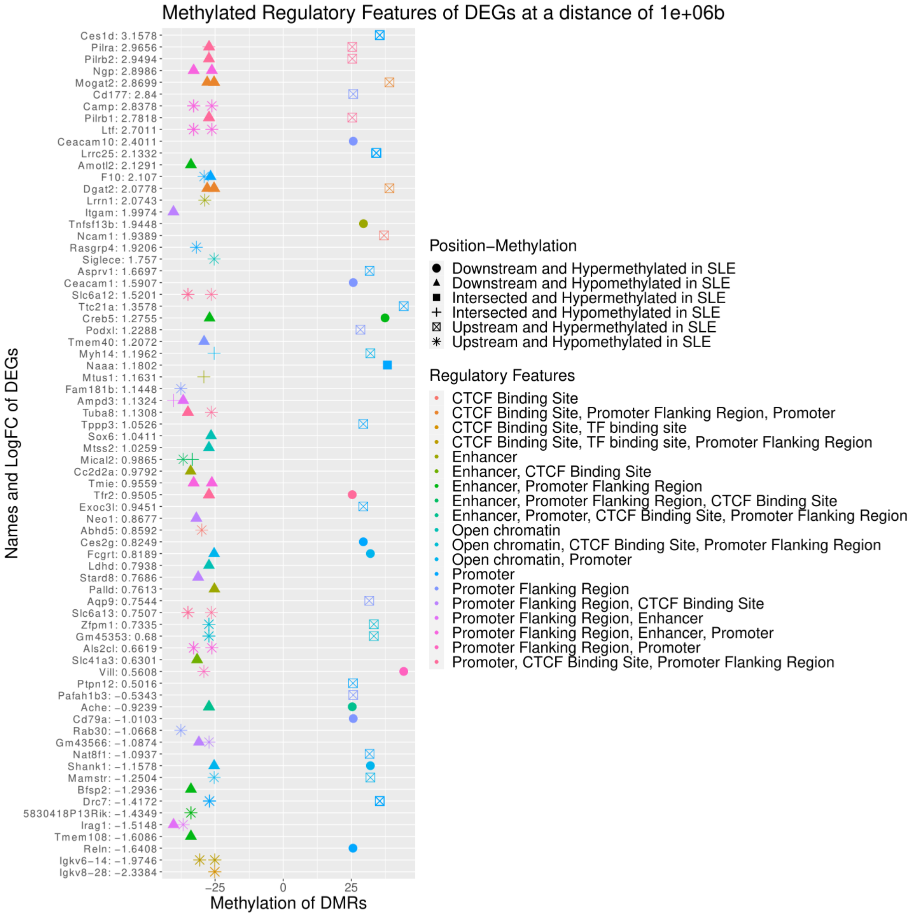


(a) Methylated RFs of DEGs in a distance of 1Mb



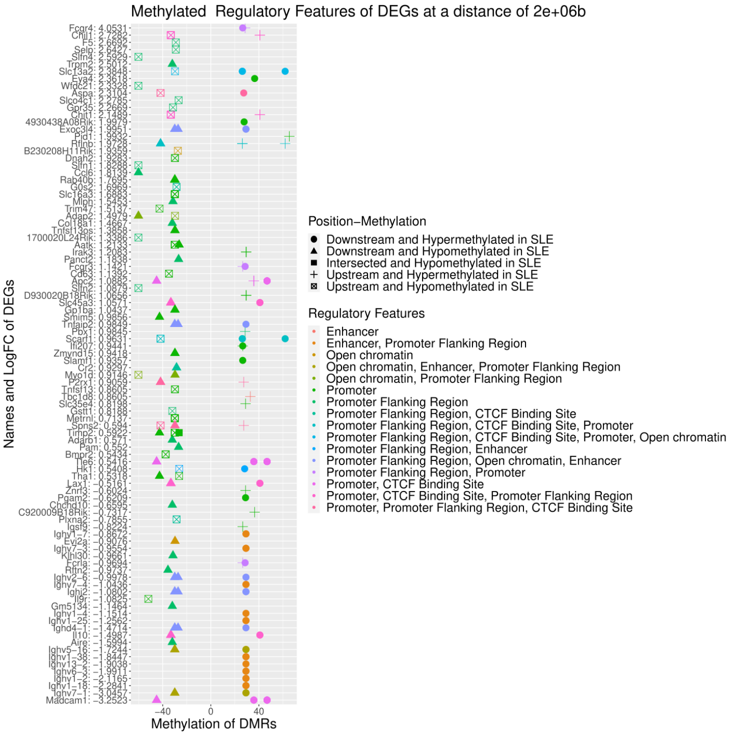
(b) Methylated RFs of DEGs in a distance of 1Mb

See the next page for complete description

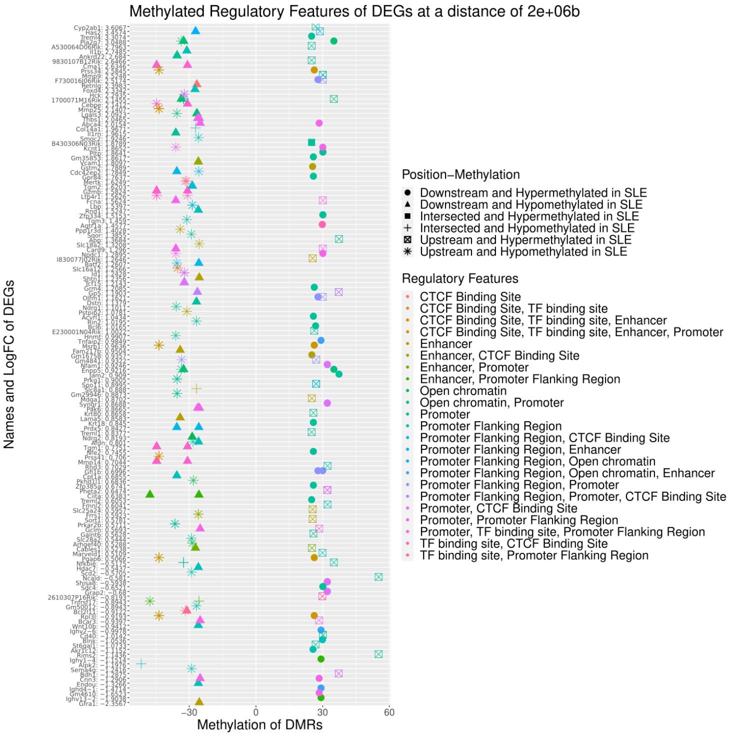


(c) Methylated RFs of DEGs in a distance of 1Mb

Figure 3.14: Dot plots where in y-axis are the DEGs with their logFC values in descending order (up to down) that their RFs in a distance of 1Mb are methylated. X-axis shows the methylation values of those RFs and at the same time the shape and colour of each dot gives the position of the RFs from DEGs and their annotation status respectively.

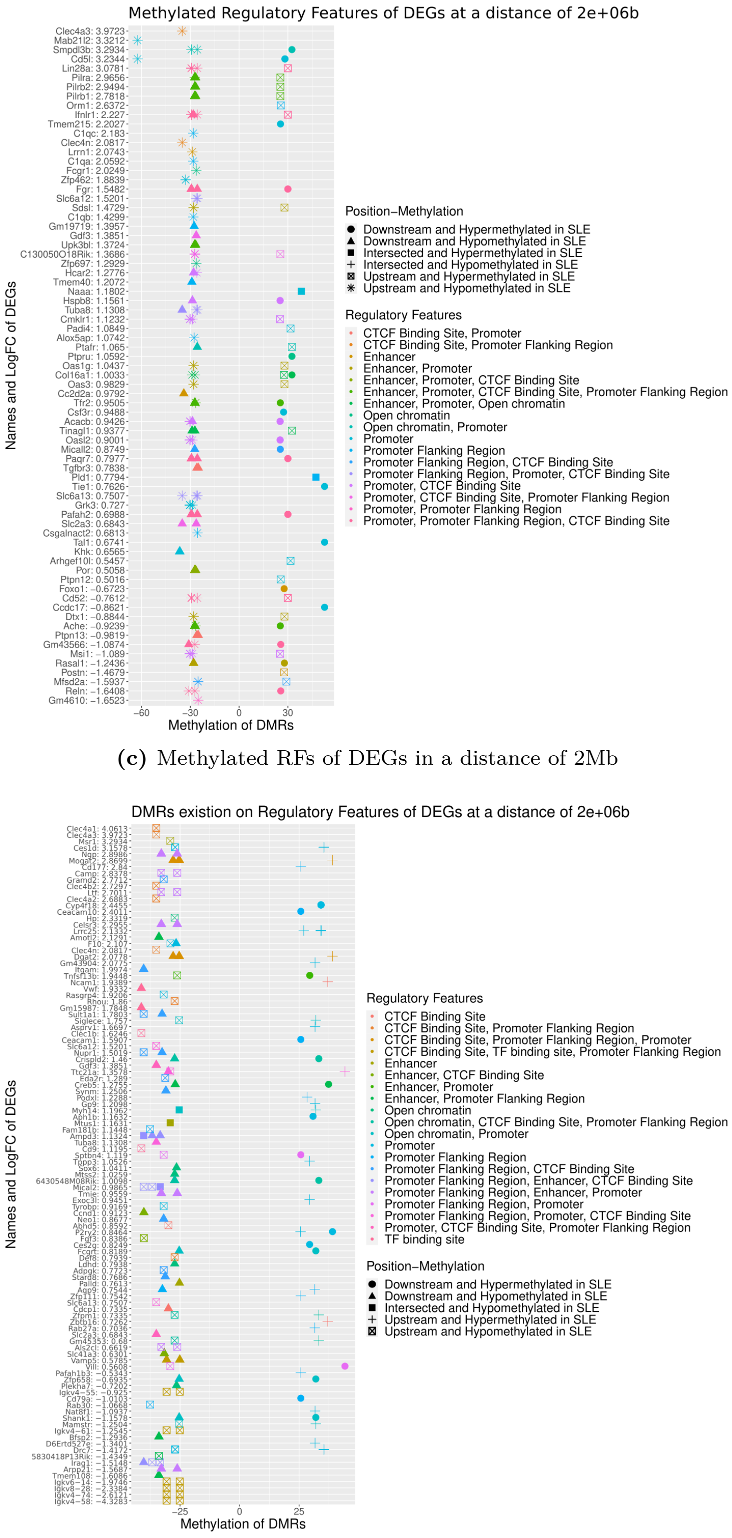


(a) Methylated RFs of DEGs in a distance of 2Mb



(b) Methylated RFs of DEGs in a distance of 2Mb

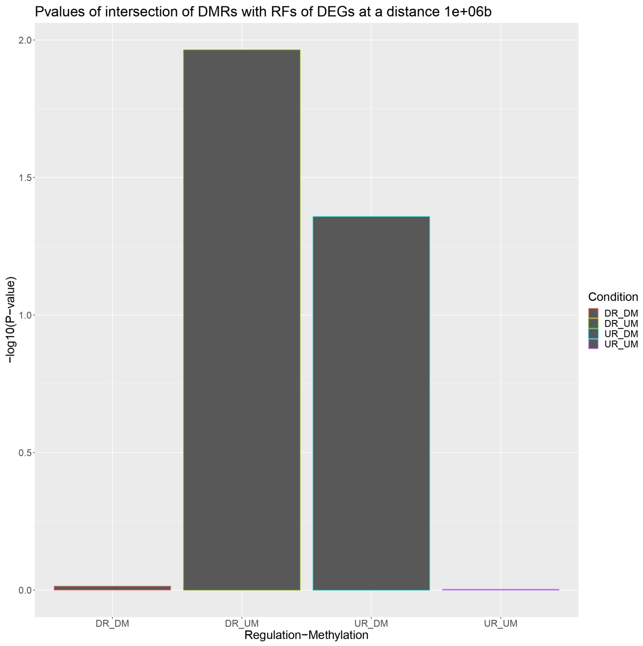
See the next page for the next subfigures



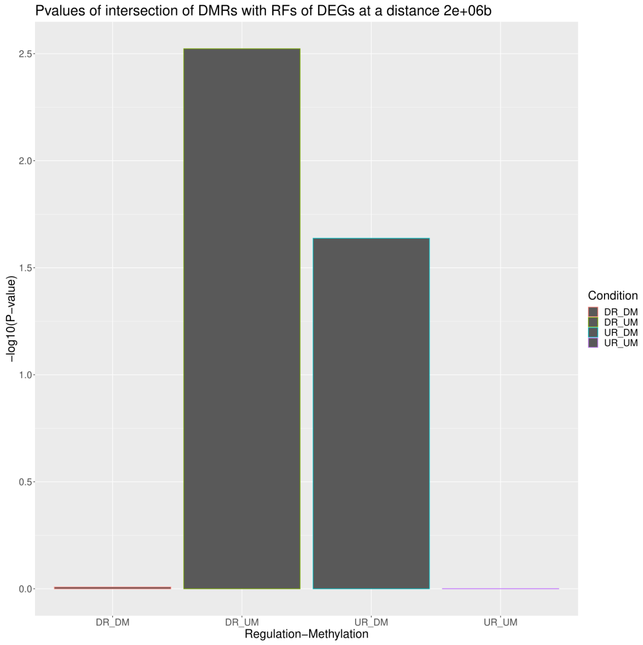
(c) Methylated RFs of DEGs in a distance of 2Mb

(d) Methylated RFs of DEGs in a distance of 2Mb

Figure 3.15: Dot plots where in y-axis are the DEGs with their logFC values in descending order (up to down) that their RFs in a distance of 2Mb are methylated. X-axis shows the methylation values of those RFs and at the same time the shape and colour of each dot gives the position of the RFs from DEGs and their annotation status respectively.



(a) Pvalues showing the significance in the linkage of methylated RFs with DEGs in a distance of 1Mb



(b) Pvalues showing the significance in the linkage of methylated RFs with DEGs in a distance of 2Mb

Figure 3.16: Bar plots where y-axis indicates the -log10(p-value) found for each relation of methylated RFs and DEGs represented in the x-axis. A value above 1.3 in y-axis represents p-values<0.05.

3.6 DEGs involved in pathways of interest in F1-L HSPCs have distant methylated TFBS

As mentioned, B and T lymphocytes have been proven to play a key role in SLE's adaptive immune response, while innate immune components have only recently been studied and discovered to also have a significant impact in the disease Herrada et al. (2019); Pan et al. (2020). Grigoriou et al. (2020) presented evidence for deregulation of hematopoiesis, with a skewing towards myeloid lineage at the expense of lymphopoiesis and priming of HSPCs, which may contribute to persistent inflammation in SLE and risk for flare once the disease is in remission. Given all this, we wanted to investigate the epigenetic profile of DEGs related to pathways of interest, more specifically to myeloid-related pathways, in F1-L HSPCs. Since it was previously shown that the majority of genes found closest to DMRs are not DEGs, we then asked if our genes of interest are the ones nearest to DMRs (Figure 3.17). Following on, these results did not report any significance, concluding that our genes of interest do not have a near-by DMR/methylated core promoter. This result is expected, since most of the genes of interest are DEGs. In order to investigate the epigenetic profile of the genes involved in pathways of interest, we used the results found previously of DEGs having DMRs within the significant distances found from the previous results, 1Mb and 2Mb. More specifically, we retrieved the DEGs which are markers for the pathways of interest that also have DMRs within the distance mentioned above and designed two dot plots (Figure 3.18). Here, the results show the expression values of DEGs which are in contrast with the methylated values of their DMRs found in the distances of 1Mb and 2Mb, an outcome that was awaited since it was previously shown using all DEGs and we know that most genes of interest are DEGs. To find the significance in the relationship of DMRs and DEGs of interest, we executed a statistical test which revealed a significant DR-UM affiliation in the distance of 2Mb (Figure 3.19). Subsequently, we performed a similar analysis to investigate the methylated regulatory factors of DEGs involved in pathways of interest. Figure 3.20 shows the expression values of DEGs and what type of markers are, along with the epigenetic profile of their regulatory factors in the distances (a) 1Mb and (b) 2Mb. Here, we notice some important genes, such as *Blnk1, Cd79a, Cebpe, Vwf, Csf3r, Ccnd1* and *Ciita*, that were previously reported significant in F1-L to have methylated regulatory elements Grigoriou et al. (2020).

Grigoriou et al. (2020) showed the increase of neutrophils in F1-L BM suggesting deregulation of homeostatic mechanisms in the level of CMPs with priming of HSPCs towards the granulocytic differentiation at the expense of lymphopoiesis. More specifically, *Blnk1* and *Cd79a*, which are lymphoid markers were found downregulated in murine F1-L HSPCs which is consistent with Grigoriou et al. (2020) results, adding the fact that their regulatory features are methylated. Specifically a promoter in a distance of 1Mb, downstream from Blnk1 was found hypermethylated and the same applies for *Cd79a* (Figure 3.20).

Other genes such as *Cebpe, Vwf, Csf3r, Ccnd1* and *Ciita*, where the first three are myeloid markers, the fourth is a gene involved in proliferation and the last one is an IFN stimulated gene, were all found upregulated. This is also consistent with Grigoriou et al. (2020), with the majority having hypomethylated regulatory features (Figure 3.20). *Cebpe* has two hypomethylated promoters, one upstream and one downstream, *Vwf* has a downstream hypomethylated TF binding site and *Csf3r* a downstream hypermethylated promoter. *Ciita* has two downstream hypomethylated regulatory features, an enhancer and a promoter together. Finally, we can also detect that in the distance of 2Mb, there are two downstream hypomethylated regulatory features from *Ccnd1*, an enhancer and a CTCF binding site together.

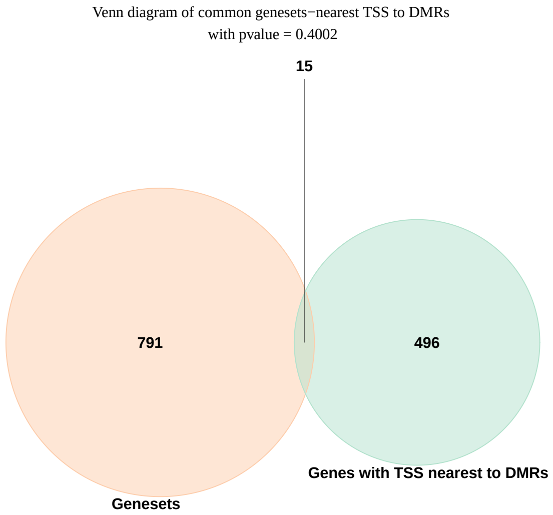
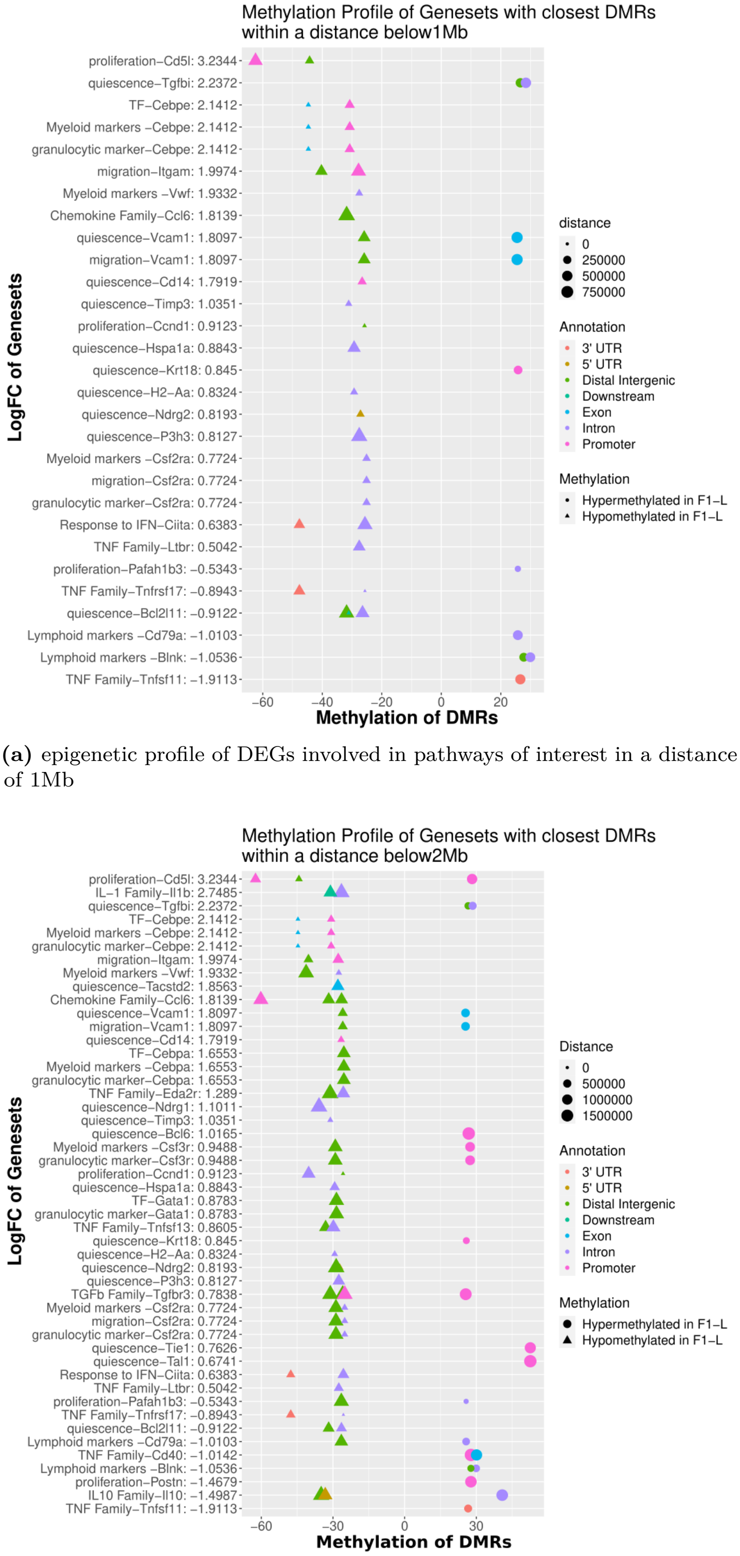


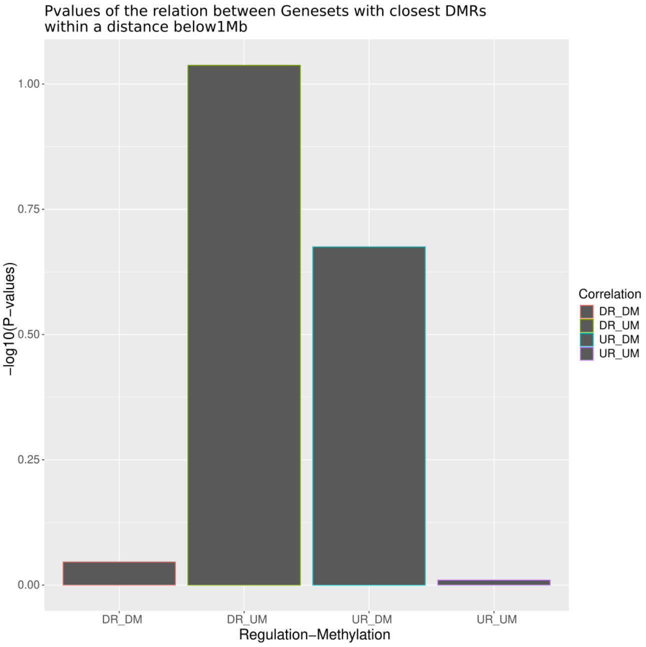
Figure 3.17: Venn diagram showing the common genes between DEGs that are involved in pathways of interest and TSS closest to DMRs. The orange circle represents the number of DEGs involved in pathways of interest while the green circle the number of TSS found closest to DMRs. The significance in the number of common genes(15) is reported with a p value of 0.4002.



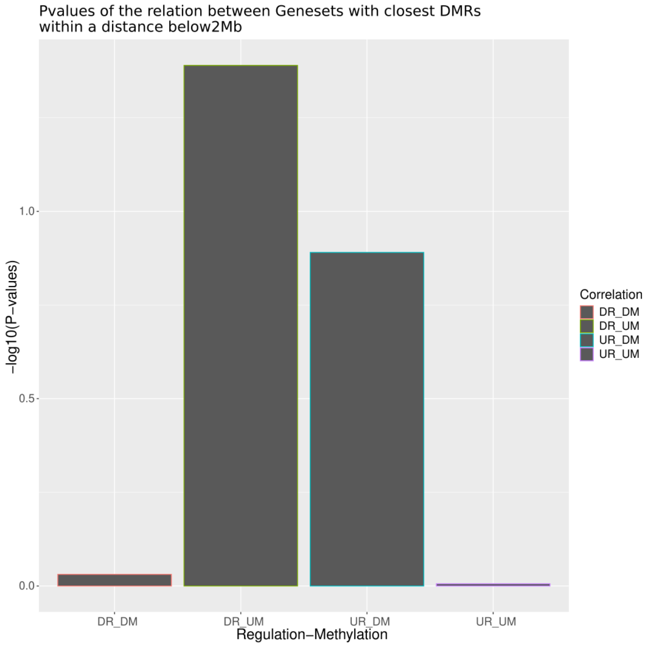
(a) epigenetic profile of DEGs involved in pathways of interest in a distance of 1Mb

(b) epigenetic profile of DEGs involved in pathways of interest in a distance of 2Mb

Figure 3.18: Dot plots where in y-axis are the DEGs with their logFC values in descending order (up to down) that have closest DMRs within a distance of (a) 1Mb and (b) 2Mb showing their methylation values in x-axis. Each dot represents a DMR found closest to each DEG at each distance, where the size, colour and shape show their distance from DEGs, their annotation and their methylation status.

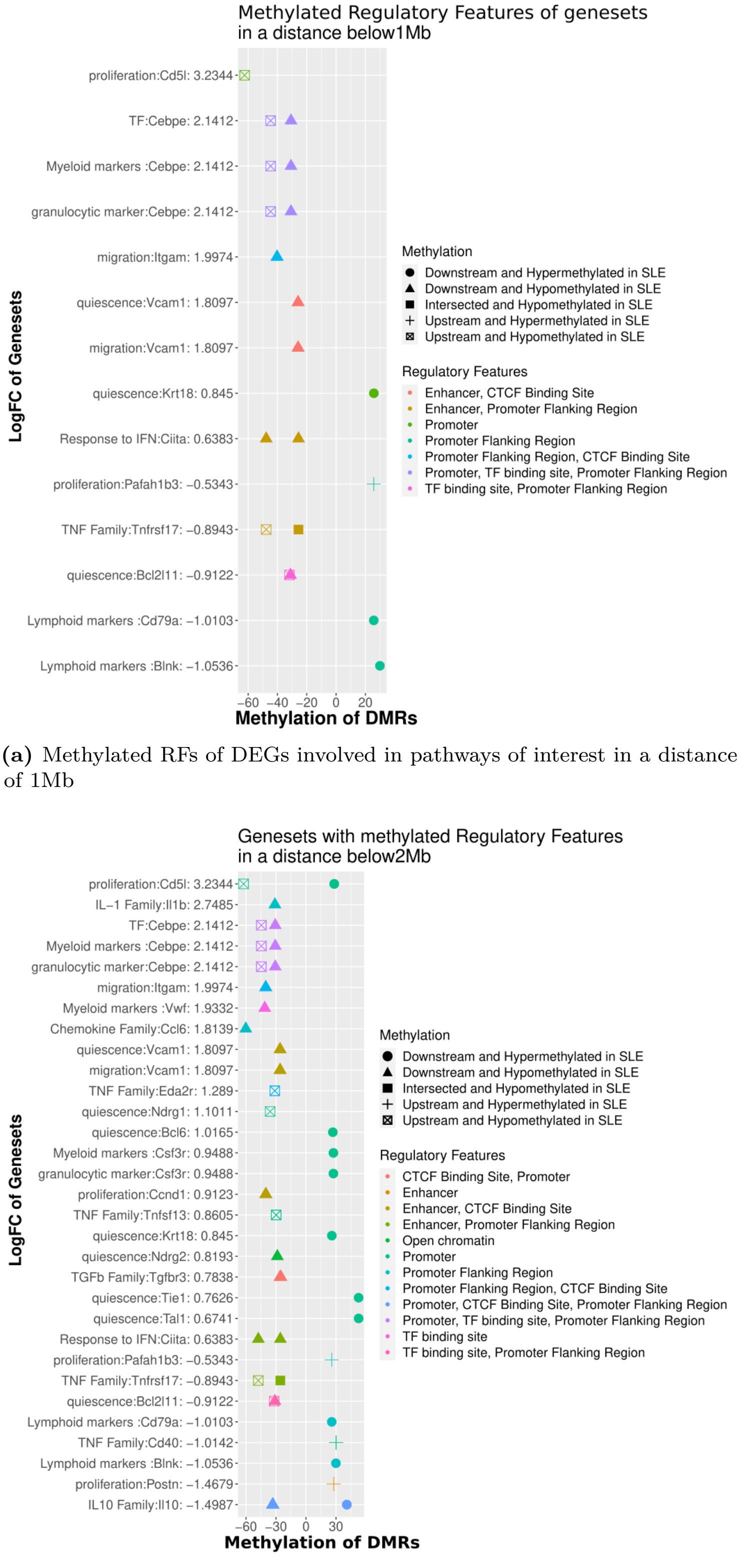


(a) Pvalues showing the significance in the linkage of DEGs involved in pathways of interest and their closest DMRs in a distance of 1Mb



(b) Pvalues showing the significance in the linkage of DEGs involved in pathways of interest and their closest DMRs in a distance of 2Mb

Figure 3.19: Bar plots where y-axis indicates the -log10(p-value) found for each relation of DEGs involved in pathways of interest and their closest DMRs in the distances of (a) 1Mb and (b) 2Mb. X-axis shows the different types of relation and a value above 1.3 in y-axis represents p-values<0.05.



(a) Methylated RFs of DEGs involved in pathways of interest in a distance of 1Mb

(b) Methylated RFs of DEGs involved in pathways of interest in a distance of 2Mb

Figure 3.20: Dot plots where in y-axis are the DEGs involved in pathways of interest with their logFC values in descending order (up to down) that their RFs in a distance of (a) 1Mb and (b) 2Mb are methylated. X-axis shows the methylation values of those RFs and at the same time the shape and colour of each dot gives the position of the RFs from DEGs and their annotation status respectively.

4. Discussion

HSPCs reside in the BM in a quiescent state, being ready to respond to stress, such as severe infection, systemic inflammation or iatrogenic myeloablation Carrelha et al. (2018). HSPCs are the most primitive multipotent population that give rise to all blood cell types, while the majority of cells involved in the pathogenesis of SLE come from BM HSPCs King and Goodell (2011). Grigoriou et al. (2020) showed evidence of haematopoiesis dysregulation in SLE, with a skewing toward the myeloid lineage which is associated with epigenetic tinkering, at the expense of lymphopoiesis and priming of HSPCs. In the same context with Grigoriou et al. (2020) that the HSPCs in the BM are responsible for the fundamental molecular abnormalities (genetic or epigenetic) in SLE, we investigated the epigenetic profile of F1-L HSPCs. We show that methylated distant CREs and TFBS regulate DEGs as those involved in pathways of interest in F1-L HSPCs. This could lead HSPCs to reprogramme into the myeloid lineage, perhaps contributing to enhanced immune responses and flares in SLE.

In order to examine DEGs alongside with the DMRs and see whether there was any linkage, we only needed a little information regarding DEGs such as their log2FC values and coordinates. A total of 809 DEGs between F1-P and F1-L were identified, of which 181 were found downregulated and 628 upregulated. Transcriptional analysis indicates that most DEGs in F1-L HSPCs are overexpressed. Moving on, methylation analysis results revealed that hypomethylation prevails in relation to hypermethylation in murine F1-L HSPCs even at each chromosome. Target regions are enriched for mostly CpG Islands, introns, intergenic regions and fewer promoters. Some of those DMRs overlap in those regions but still, most methylated regions are introns and intergenic and in fact distal, meaning that they are >3kb from the nearest TSS. On that note, we expect to find methylated CTCF binding sites and enhancers since we know that they exist within intergenic regions and introns respectively Kim et al. (2007), Park et al. (2014). These findings are consistent with a previous study that showed that large methylation shifts in SLE were almost entirely composed of hypomethylation events Absher et al. (2013). Collectively, in mouse F1-L HSPCs, hypomethylation dominates, with the majority of DMRs located on CpG islands, introns, and distal intergenic regions, and less on promoters.

Given that some DMRs are on promoter regions and some were found up to 3kb near to TSS known as core promoters, it was worth investigating the fact that a significant number of DEGs may have close-by DMRs that possibly are also on their core promoters. These results revealed that very few DEGs have near-by DMRs while the majority of the reported nearest gene found for each DMR, might play another role such as regulating other genes. These results also show that *Ccnd1*, a gene involved in proliferation was found significantly upregulated and in consistent with Grigoriou et al. (2020), showing the HSPC activation, has an upstream hypomethylated DMR. Additionally, *Cepbe* a myeloid marker was found to have a hypomethylated near-by exon. Subsequently, we found that the closest DMRs for each DEG actually belonged in a long range distance. Hence, we suspected that CREs-enhancers and distant TFBS-CTCF of DEGs are methylated contributing to their expression. In order to investigate this hypothesis, we first analyzed the relationship of DMRs closest to DEGs within the distances of 50kb, 100kb, 500kb, 1Mb and 2Mb where the last two distances revealed a significance in their inverse relation. Then, we used the distance found to be significant and showed methylated distant CREs-enhancers and TFBS that probably regulate DEGs. These results suggest that distant methylated CREs and TFBS do interact and regulate DEGs.

According to previous studies, haematopoiesis dysregulation occurs in SLE, with a skewing toward the myeloid lineage at the expense of lymphopoiesis and priming of HSPCs with a "trained immunity" signature, which may contribute to inflammation and flare risk Grigoriou et al. (2020). Here, we report the epigenetic profile involved in this particular dysregulation. Consistent with Grigoriou et al. (2020) results, lymphoid markers *Blnk1* and *Cd79a* were found downregulated in F1-L HSPCs, adding the fact that their regulatory features are methylated. Specifically a promoter in a distance of 1Mb, downstream from *Blnk1* was found hypermethylated and the same applies for *Cd79a.* Other genes such as *Cebpe, Vwf, Csf3r, Ccnd1* and *Ciita*, where the first three are myeloid markers, the fourth is a gene involved in proliferation and the last one is an IFN stimulated gene, were all found upregulated. This is also consistent with Grigoriou et al. (2020), with the majority having hypomethylated regulatory features. *Cebpe* has two hypomethylated promoters, one upstream and one downstream, *Vwf* has a downstream hypomethylated TF binding site and *Csf3r* a downstream hypermethylated promoter. *Ciita* has two downstream hypomethylated regulatory features, an enhancer and a promoter together. Finally, we can also detect that in the distance of 2Mb, there are two downstream hypomethylated regulatory features from *Ccnd1*, an enhancer and a CTCF binding site together. These data add that the methylated regions found close to *Ccnd1* and *Cebpe* genes are not regulatory factors but in the distance of 2Mb there were found hypomethylated regulatory features. Taking these results together, confirm the initial finding that HSPCs lean towards myelopoiesis in SLE.

In summary, we have analyzed the association of DNA methylation with DEGs F1-L HSPCs. We provide evidence that CREs specifically enhancers and distant TFBS of DEGs and those involved in pathways of interest in F1-L HSPCs are methylated, contributing to their regulation of expression. Pathways of interest such myelopoiesis and lymphopoiesis have been presented before in recent studies indicating that murine and human lupus SLE HSPC's gene expression program is biased towards myelopoiesis Kokkinopoulos et al. (2021). We demonstrate that regulatory factors of *Blnk1, Cebpe* and *Ciita* that are markers of such pathways, are methylated and possibly contribute to their regulation of expression. Therefore, we suggest that the HSPC reprogramming towards myeloid lineage, which may contribute to increased immunological responses and flares in SLE, is due to the methylation of distant methylated CREs and TFBS of DEGs.

Topologically associating domains (TADs) are fundamental units of three-dimensional (3D) nuclear organization. The regions bordering TADs-TAD boundaries-contribute to the regulation of gene expression by restricting interactions of cis-regulatory sequences to their target genes McArthur and Capra (2021). In the last couple of years it has been shown that TAD and TAD-boundary disruption have been linked to the development of diseases Farooq et al. (2022); McArthur and Capra (2021), while another recent study reported change in chromatin accessibilities at aged HSCs Itokawa et al. (2022). Recent studies have also demonstrated that modulation of gene expression via 3D chromatin structure is important for many physiologic and pathologic cellular functions, including cell-type identity, cellular differentiation, and risk for multiple rare diseases and cancer. Hence, using chromosome-conformation-capture technologies (3C, 4C, 5C, Hi-C), in F1-L HSPCs might reveal and answer many fundamental questions such as if TADs and TAD boundaries are disrupted, hence, linked to the transcription results. Finally, the importance of TFBSs is reflected by the many techniques that have been developed for their identification, including chromatin immunoprecipitation followed by sequencing (ChIP-seq), protein-binding microarray and many others. ChIP-seq, is a method used to analyze protein interactions with DNA and is widely used to study the *in vivo* TFBS, and their regulatory targets. Hence, using ChIP-seq we can further explore the transcription factors, DNA methylation results and transcription analysis in F1-L HSPCs.

4. BIBLIOGRAPHY

D. M. Absher, X. Li, L. L. Waite, A. Gibson, K. Roberts, J. Edberg, W. W. Chatham, and R. P. Kimberly. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ t-cell populations. *PLoS Genet.*, 9(8):e1003678, Aug. 2013.

K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197, Mar. 2000.

S. Andrews. Fastqc. a quality control tool for high throughput sequence data. 2010.

M. T. Baldridge, K. Y. King, and M. A. Goodell. Inflammatory signals regulate hematopoietic stem cells. *Trends Immunol.*, 32(2):57–65, Feb. 2011.

F. Basta, F. Fasola, K. Triantafyllias, and A. Schwarting. Systemic lupus erythematosus (SLE) therapy: The old and the new. *Rheumatol Ther*, 7(3):433–446, Sept. 2020.

T. Baubec and A. Akalin. Genome-wide analysis of DNA methylation patterns by high-throughput sequencing. In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*, pages 197–221. Springer International Publishing, Cham, 2016.

S. Behjati and P. S. Tarpey. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.*, 98(6):236–238, Dec. 2013.

J. Carrelha, Y. Meng, L. M. Kettyle, T. C. Luis, R. Norfo, V. Alcolea, H. Boukarabila, F. Grasso, A. Gambardella, A. Grover, K. Högstrand, A. M. Lord, A. Sanjuan-Pla, P. S. Woll, C. Nerlov, and S. E. W. Jacobsen. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature*, 554(7690):106–111, Feb. 2018.

H. Chen. Venndiagram: Generate high-resolution venn and euler plots. 2022. URL `https://CRAN.R-project.org/package=VennDiagram`. R package version 1.7.3.

H. Cheng, Z. Zheng, and T. Cheng. New paradigms on hematopoietic stem cell differentiation. *Protein Cell*, 11(1):34–44, Jan. 2020.

L. S. Davis and A. M. Reimold. Research and therapeutics-traditional and emerging therapies in systemic lupus erythematosus. *Rheumatology*, 56(suppl_1):i100–i113, Apr. 2017.

N. L. de Silva and S. L. Seneviratne. Haemopoietic stem cell transplantation in systemic lupus erythematosus: a systematic review. *Allergy Asthma Clin. Immunol.*, 15:59, Sept. 2019.

A. S. Doane and O. Elemento. Regulatory elements in molecular networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 9(3), May 2017.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013.

S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, Aug. 2005.

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.*, 4(8):1184–1191, July 2009.

A. Farooq, G. Trøen, J. Delabie, and J. Wang. Integrating whole genome sequencing, methylation, gene expression, topological associated domain information in regulatory mutation prediction: A study of follicular lymphoma. *Comput. Struct. Biotechnol. J.*, 20:1726–1742, Mar. 2022.

M. Grigoriou, A. Banos, A. Filia, P. Pavlidis, S. Giannouli, V. Karali, D. Nikolopoulos, A. Pieta, G. Bertsias, P. Verginis, I. Mitroulis, and D. T. Boumpas. Transcriptome reprogramming and myeloid skewing in haematopoietic stem and progenitor cells in systemic lupus erythematosus. *Ann. Rheum. Dis.*, 79(2):242–253, Feb. 2020.

E. Gunsilius, G. Gastl, and A. L. Petzer. Hematopoietic stem cells. *Biomed. Pharmacother.*, 55(4):186–194, May 2001.

C. M. Hedrich, K. Mäbert, T. Rauen, and G. C. Tsokos. DNA methylation in systemic lupus erythematosus. *Epigenomics*, 9(4):505–525, Apr. 2017.

A. A. Herrada, N. Escobedo, M. Iruretagoyena, R. A. Valenzuela, P. I. Burgos, L. Cuitino, and C. Llanos. Innate immune cells' contribution to systemic lupus erythematosus. *Front. Immunol.*, 10:772, Apr. 2019.

S. A. Huber, Michael Love. DESeq2, 2017.

N. Itokawa, M. Oshima, S. Koide, N. Takayama, W. Kuribayashi, Y. Nakajima-Takagi, K. Aoyama, S. Yamazaki, K. Yamaguchi, Y. Furukawa, K. Eto, and A. Iwama. Epigenetic traits inscribed in chromatin accessibility in aged hematopoietic stem cells. *Nat. Commun.*, 13(1):2691, May 2022.

S. Kim, N.-K. Yu, and B.-K. Kaang. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.*, 47:e166, June 2015a.

T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov, and B. Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245, Mar. 2007.

Y. W. Kim, S. Lee, J. Yun, and A. Kim. Chromatin looping and eRNA transcription precede the transcriptional activation of gene in the β-globin locus. *Biosci. Rep.*, 35 (2), Mar. 2015b.

K. Y. King and M. A. Goodell. Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nat. Rev. Immunol.*, 11(10):685–692, Sept. 2011.

I. Kokkinopoulos, A. Banos, M. Grigoriou, A. Filia, T. Manolakou, T. Alissafi, N. Malissovas, I. Mitroulis, P. Verginis, and D. T. Boumpas. Patrolling human SLE haematopoietic progenitors demonstrate enhanced extramedullary colonisation; implications for peripheral tissue injury. *Sci. Rep.*, 11(1):15759, Aug. 2021.

M. Kondo, I. L. Weissman, and K. Akashi. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*, 91(5):661–672, Nov. 1997.

S. Kurdyukov and M. Bullock. DNA methylation analysis: Choosing the right method. *Biology*, 5(1), Jan. 2016.

X. Li, D. Liu, L. Zhang, H. Wang, Y. Li, Z. Li, A. He, B. Liu, J. Zhou, F. Tang, and Y. Lan. The comprehensive DNA methylation landscape of hematopoietic stem cell development. *Cell Discov*, 7(1):86, Sept. 2021.

Y. Liu, Y. Han, L. Zhou, X. Pan, X. Sun, Y. Liu, M. Liang, J. Qin, Y. Lu, and P. Liu. A comprehensive evaluation of computational tools to identify differential methylation regions using RRBS data. *Genomics*, 112(6):4567–4576, Nov. 2020.

M. G. Manz, T. Miyamoto, K. Akashi, and I. L. Weissman. Prospective isolation of human clonogenic common myeloid progenitors. *Proc. Natl. Acad. Sci. U. S. A.*, 99 (18):11872–11877, Sept. 2002.

E. McArthur and J. A. Capra. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.*, 108(2):269–283, Feb. 2021.

D. Metcalf. Hematopoietic cytokines. *Blood*, 111(2):485–491, Jan. 2008.

L. D. Moore, T. Le, and G. Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, Jan. 2013.

S. J. Morrison, A. M. Wandycz, H. D. Hemmati, D. E. Wright, and I. L. Weissman. Identification of a lineage of multipotent hematopoietic progenitors. *Development*, 124(10):1929–1939, May 1997.

L. Pan, M.-P. Lu, J.-H. Wang, M. Xu, and S.-R. Yang. Immunological pathogenesis and treatment of systemic lupus erythematosus. *World J. Pediatr.*, 16(1):19–30, Feb. 2020.

S. G. Park, S. Hannenhalli, and S. S. Choi. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. *BMC Genomics*, 15:526, June 2014.

S. Pathak and C. Mohan. Cellular and molecular pathogenesis of systemic lupus erythematosus: lessons from animal models. *Arthritis Res. Ther.*, 13(5):241, Sept. 2011.

G. H. Putri, S. Anders, P. T. Pyl, J. E. Pimanda, and F. Zanini. Analysing high-throughput sequencing data in python with HTSeq 2.0. *Bioinformatics*, 38(10): 2943–2945, May 2022.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar. 2010.

R Core Team. R: A language and environment for statistical computing. 2021. URL `https://www.R-project.org/`.

L. Robb. Cytokine receptors and hematopoietic differentiation. *Oncogene*, 26(47): 6715–6723, Oct. 2007.

N. C. Roy, E. Altermann, Z. A. Park, and W. C. McNabb. A comparison of analog and Next-Generation transcriptomic tools for mammalian studies. *Brief. Funct. Genomics*, 10(3):135–150, May 2011.

C. Schulz, U. H. von Andrian, and S. Massberg. Hematopoietic stem and progenitor cells: their mobilization and homing to bone marrow and peripheral tissue. *Immunol. Res.*, 44(1-3):160–168, 2009.

J. Seita and I. L. Weissman. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, 2(6):640–653, Nov. 2010.

J. Shendure. The beginning of the end for microarrays? *Nat. Methods*, 5(7):585–587, July 2008.

T. Suda, K. Takubo, and G. L. Semenza. Metabolic regulation of hematopoietic stem cells in the hypoxic niche. *Cell Stem Cell*, 9(4):298–310, Oct. 2011.

A. Trumpp, M. Essers, and A. Wilson. Awakening dormant haematopoietic stem cells. *Nat. Rev. Immunol.*, 10(3):201–209, Mar. 2010.

V. van Heyningen and W. Bickmore. Regulation from a distance: long-range control of gene expression in development and disease. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1620):20120372, May 2013.

Q. Wang, Y. Jia, Y. Wang, Z. Jiang, X. Zhou, Z. Zhang, C. Nie, J. Li, N. Yang, and L. Qu. Evolution of cis- and trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. *BMC Genomics*, 20(1):933, Dec. 2019.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, Jan. 2009.

H. Wickham. *Ggplot2: Elegant graphics for data analysis.* Use R! Springer, New York, NY, 1 edition, Dec. 2009.

Wikipedia contributors. Transcription factor — Wikipedia, the free encyclopedia. 2022. [Online; accessed 4-July-2022].

L. Yang, D. Bryder, J. Adolfsson, J. Nygren, R. Månsson, M. Sigvardsson, and S. E. W. Jacobsen. Identification of Lin(-)Sca1(+)kit(+)CD34(+)Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. *Blood*, 105(7):2717–2723, Apr. 2005.

G. Yu, L.-G. Wang, and Q.-Y. He. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14):2382–2383, July 2015.

C. C. Zhang and H. F. Lodish. Cytokines regulating hematopoietic stem cell function. *Curr. Opin. Hematol.*, 15(4):307–311, July 2008.

J. L. Zhao and D. Baltimore. Regulation of stress-induced hematopoiesis. *Curr. Opin. Hematol.*, 22(4):286–292, July 2015.

J. Zhu and S. G. Emerson. Hematopoietic cytokines, transcription factors and lineage commitment. *Oncogene*, 21(21):3295–3313, May 2002.