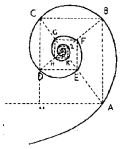
UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Medicina Molecolare



Ciclo XXVII
Anno Accademico 2013/2014

NEXT-GENERATION SEQUENCING APPROACH FOR TRANSCRIPTOME AND EPIGENOME DEFINITION OF HUMAN HEMATOPOIETIC STEM/PROGENITOR CELLS AND THEIR EARLY ERYTHROID AND MYELOID COMMITTED PROGENY

Dottorando: Luca PETITI

Direttore della Scuola di Dottorato: Prof. Mario CLERICI

Tutore: Prof. Cristina BATTAGLIA

Co-tutore: Dr. Clelia PEANO

UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Medicina Molecolare

Ciclo XXVII
Anno Accademico 2013/2014

Settore BIO/10

Next-Generation Sequencing Approach for Transcriptome and Epigenome Definition of Human Hematopoietic Stem/Progenitor Cells and their Early Erythroid and Myeloid Committed Progeny

Dottorando: Luca PETITI

Matricola N° R09618

TUTORE: Prof.ssa Cristina BATTAGLIA
CO-TUTORE: Dott.ssa Clelia PEANO

DIRETTORE DEL DOTTORATO: Prof. Mario CLERICI

ABSTRACT

Somatic stem cells are the basic tools of regenerative medicine and gene therapy, providing unique opportunities for the therapy of genetic and acquired disorders. The molecular mechanisms underlying fundamental characteristics of human somatic stem cells, such as self-renewal, commitment and differentiation, are still poorly understood. A better knowledge of these mechanisms is crucial to the understanding of stem cell biology and to the development of stem cell-based therapies. High-throughput approaches, such as next-generation sequencing technologies (NGS) became fundamental to study the transcriptome, the epigenome and the usage of regulatory elements in the genome. Genome-wide approaches allow investigating the molecular circuitry wiring the genetic and epigenetic programs of human somatic stem cells. Here, we define the transcriptional and epigenetic profile of human hematopoietic stem/progenitor cells (HSPC) and early myeloid and erythroid progenitors through an integrative analysis of Cap Analysis of Gene Expression (CAGE) and chromatin immuno-precipitation (ChIP-seq) data, in order to identify transcription regulatory elements that act in HSPC lineage commitment. CAGE analysis enabled us to define more than 10,000 active promoters in HSPCs and in erythroid (EPP) and myeloid precursors (MPP). The different cell types shared most of the promoters, with only a small fraction (about 4%) being differentially transcribed, suggesting that the transcriptional state is largely maintained in early hematopoietic progenitors and precursors. Interestingly about 30% of the identified of cell-specific promoters was not annotated. These novel transcripts are possibly involved in HSPCs self-renewal, commitment and differentiation. To obtain a genome-wide description of the transcriptional regulatory regions in multipotent and lineage-restricted hematopoietic progenitors, we performed ChIP-seq analysis for histone methylations typical of promoters and enhancers, H3K4me3 and H3K4me1, respectively, and for H3K27ac to mark the active elements. Overall we identified more than 20,000 active enhancers that consistently changed upon erythroid and myeloid commitment: about 80 and 95% of the active enhancers mapped in HSPC disappeared in erytrhoid and myeloid progenitors, respectively, while novel enhancers are acquired during lineage commitment. These data indicate that enhancers are dramatically redefined during lineage commitment, and that differential enhancer usage is responsible for the differential regulation of promoter activity underlying lineage restriction. This study provided an overview of the differential transcriptional programs of HSPCs and committed myeloid and erythroid hematopoietic precursors and represents a unique source of genes and regulatory regions involved in self-renewal, commitment and differentiation of human hematopoietic stem/progenitor cells and their progeny.

SOMMARIO

Le cellule staminali somatiche rappresentano uno strumento chiave della medicina rigenerativa, aprendo opportunità importanti per la terapia di patologie ereditarie e acquisite. I meccanismi molecolari alla base delle cellule staminali somatiche umane, come l'autorinnovamento, il *commitment* e il differenziamento sono ancora oggetto di studio. Una migliore conoscenza di questi meccanismi è fondamentale per la comprensione della biologia delle cellule staminali e per lo sviluppo di terapie basate su di esse. Metodologie ad alta performanza (*high-throughput*), come il sequenziamento di ultima generazione (NGS, *next generation sequencing*) sono alla base degli studi di trascrittomica, epigenomica e per l'annotazione degli elementi regolatori della trascrizione genica. In questo lavoro abbiamo definito il profilo trascrizionale ed epigenetico delle cellule Staminali/Progenitrici Ematopoietiche (HSPC, *hematopoietic/progenitor cells*) e della loro progenie eritroide (EPP) e mieloide (MPP), attraverso analisi integrative di dati da esperimenti di *Cap Analysis of Gene Expression* (CAGE) e immunoprecipitazione della cromatina accoppiata a sequenziamento ultramassivo (ChIP-seq), al fine di identificare gli elementi di regolazione genica responsabili del *commitment* delle HSPC. L'analisi CAGE ha permesso di identificare più di 10,000 promotori attivi in HSPC, EPP e MPP. I tre tipi cellulari condividono più del 90 dei promotori attivi e solo una piccola parte risulta differenzialmente espressa. Il 30% dei promotori identificati con CAGE risulta non annotato, suggerendo che i trascritti originate da essi possano rivestire un ruolo chiave nell'autorinnovamento, commitment e differenziamento delle HSPC. Al fine di ottenere una mappa delle regioni regolatorie, abbiamo messo ap punto analisi ChIP-seq su modificazioni istoniche tipiche di promotori ed *enhancers*, come H3K4me3 e H3K4me1, e note per marcare regioni attive del genoma, come H3K27ac. Complessivamente abbiamo identificato più di 20,000 *enhancers* attivi che cambiano in modo consistente durante il *commitment* eritroide e mieloide, dimostrando che circa il 80 e il 90% degli *enhancers* attivi in HSPC non è più presente nella progenie, mentre nuovi *enhancers* vengono aquisiti da EPP e MPP. Questi dati indicano che gli *enhancers* sono profondamente ridefiniti durante il *commitment* e che l'utilizzo differenziale di questi elementi regolatori è resonsabile della regolazione differenziale dei promotori durante questa fase. Questo studio costituisce una risorsa importante per lo studio dei programmi trascrizionali che coinvolgono le HSPC, in quanto riporta una collezione esaustiva dei geni e delle regioni regolatrici coinvolte nel *commitment* di queste cellule.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABBREVIATIONS

CAGE: Cap Analysis of Gene Expression
ChIP: Chromatin ImmunoPrecipitation
CLP: Common Lymphoid Precursor
CMP: Common Myeloid Precursor
CRE: *Cis* Regulatory Element
DEG: Differentially Expressed Genes
EPP: Erythroid Progenitor
FRiP: Fraction of Reads in Peaks
HM: Histone Modification
HSPC: Hematopoietic Stem/Progenitor Cell
IP: ImmunoPrecipitation
MPP: Myeloid Progenitor
NGS: Next-Generation Sequencing
PTM: Post-Translational Modification
TF: Transcription Factor
TSS: Transcription Start Site

1. INTRODUCTION

The completion of human genome sequence project opened new challenges in understanding how the information encoded in DNA directs the specification, development, and fate of hundreds of different cell types. The accurate realization of biological processes, such as proliferation, apoptosis, aging and differentiation, requires, in fact, a precise and coordinated execution of series of steps that depend on the proper expression of the human genome. The deregulation of these processes results in abnormal phenotypes and eventually in disease.

Currently, in the era of genomics and personalized medicine, a complete and quantitative catalogue of genome features and expression profiles of every cell type is required to understand the fundamental mechanisms that control cellular processes. Despite intensive studies our understanding of the genome is far from being complete. Therefore the complete analysis of regulatory sequences, and of the mechanisms that regulate genome transcription, is essential and is a mandatory requirement to fully understand genome organization, and ultimately cell biology, in both physiological and pathological states.

In the recent years the development of high-throughput approaches, mainly based on next-generation sequencing (NGS) technologies, pushed forward genetic studies to a genome-wide scale. The technological advancement raised the need of proper informatics methodologies to manage, analyse and integrate the huge amount of data produced by genomic experiments. In this framework bioinformatics and computational biology became fields of striking importance in genomic research, for which one of the main goals is the identification of the complete collection of epigenetic modifications of DNA sequences and chromatin (the epigenome), transcripts (the transcriptome) and the functional elements that regulate gene expression of a specific cell type. Each individual aspect offers a unique and complementary view of genome organization and cellular function: integrating them together offers the huge potential to answer many long-standing biological questions.

1.1. The epigenome

An important emerging area of biological research concerns the packaging of DNA into chromatin and, specifically, how cell type-specific chromatin organization enables differential access to and activity of regulatory elements and the manifestation of unique cellular phenotypes.

1.1.1. The structure of the chromatin

The genomic DNA in the eukaryotic cell nucleus is hierarchically packaged with proteins to form chromatin. The eukaryotic chromatin can be viewed as a highly organized dynamical structure in which the primary structure, the DNA-protein polymer, is folded and compacted into three-dimensional chromatin fibres of increasing sizes to form the eukaryotic chromosome. The smallest repeating unit of chromatin is the nucleosome, which is composed by ~147 base pairs of DNA forming a two-turn helix around a compact histone octamer core consisting of two copies of histones H2A, H2B, H3 and H4. The core histone proteins are composed of structured globular domains that mediate the interaction with the DNA, and a flexible relatively unstructured N-terminal tails that extend away from nucleosomal surface. A consequence of the DNA wrapping around the histone core is the sterical occlusion of other DNA-binding proteins thus regulating their access to the DNA. Nucleosomes are connected by short DNA segments (linker DNA, ~10–80 bp in length) into nucleosomal arrays, which undergo short-range interactions with neighbouring nucleosomes to form chromatin fibres [1]. This beads-on-a-string organization of individual nucleosomes is termed chromatin primary structure. Despite the high ordered structure, the nucleosomes are not static but partially unwrap and rewrap spontaneously. This nucleosome property is crucial to regulate occupancy of DNA-binding proteins in a tunable manner through different mechanisms that can modulate the nucleosome stability and dynamics, including DNA modifications, post-translational modifications of histones (PTMs), incorporation of histone variants, ATP-dependent chromatin remodeling and non-coding RNA-mediated pathways. All these changes are epigenetic

modifications, which are heritable changes that do not involve variations in DNA sequence; they regulate chromatin structure and DNA accessibility, and influence how the genome is made manifest across a diverse array of developmental stages, tissue types, and disease states. The epigenome is defined as the combination of all epigenetic modifications in any given cell type, including DNA methylation, post-translational histone modifications and histone variants. Complex organisms such as humans do not have a single epigenome, but instead have multiple epigenome depending on the tissue type and developmental stage. So the epigenome is not static like the genome, but it can be dynamic, influenced by environmental factors and extracellular stimuli, and change in response to these factors. Deregulation of these epigenetic events has been observed in various cancers and human diseases.

1.1.2. Post-translational histone modifications

Histone proteins are subject to a number of covalent modifications, such as methylation, acetylation, phosphorylation, ubiquitylation and ADP-ribosylation, which can occur at many sites. There are over 60 different residues on histones where modifications have been detected either by specific antibodies or by mass spectrometry, but this value may represent an underestimation of the total number of modifications that can take place on histones. This vast array of modifications gives enormous potential for functional responses, but it has to be remembered that not all these modifications will be on the same histone at the same time. The timing of the appearance of a modification will depend on the signalling conditions within the cell [2]. PTMs of histones occur on either within the globular core or on the amino-terminal tails of core histones and can act in two ways: i) directly affecting the chromatin compaction and assembly or ii) serving as binding sites for effector protein [3]. In the first case, histone modifications may affect higher-order chromatin structure by modifying the contact between different histones in adjacent nucleosomes or the interaction of histones with DNA. In the second case, depending on the composition of modifications on a given histone, a set of proteins, such as chromatin remodelers and transcription factors involved in different processes (e.g.

transcription initiation, elongation) are encouraged to bind or are occluded from chromatin. One of the most well studied PTMs is the acetylation of lysine residues of histones, an highly dynamic process regulated by two class of enzymes, histone acetyltransferases (HATs) and histone deacetylases (HDACs). The HATs catalyse the transfer of an acetyl group to the lysine side chains, neutralizing the lysine's positive charge with the consequence of disruption of the electrostatic interactions between histones and DNA, therefore facilitating the DNA access to transcription machinery. This function correlates with their role of transcriptional coactivators. HDACs enzymes reverse the effect of HATs and restore the positive charge of the lysine residues, stabilizing the local chromatin architecture thus acting mainly as chromatin repressors. The other important class of histone modifications is the methylation, which mainly occurs on the side chains of lysines and arginines. Unlike acetylation, this class of PTMs does not alter the charge of histone proteins but it has a higher level of complexity because lysine residues can be mono-, di-, tri-methylated and arginine can be mono-, symmetrically or asymmetrically di-methylated, with different consequences on chromatin dynamics. This characteristic of methylation is supported by the existence of relatively specific classes of enzymes, such as histone lysine methyltransferases (KMTs) that methylate distinct lysine residues (e.g. H3K4, lysine 4 on histone 3) to a specific degree (mono-, di- or tri-methylation) [4]. Methylation of histones is implicated in multiple cellular processes, depending on the residue being methylated and on the degree of methylation. Methylation of five residues within the N-terminal tail (H3K4, H3K9, H3K27, H3K36 and H4K20) of histones H3 and H4, and of two residues in the globular domain (H3K64 and H3K79) of histone H3 have been functionally characterized. In general, H3K9, H3K27, H3K64 and H4K20 methylation have been implicated in transcriptional silencing, whereas H3K4, H3K36 and H3K79 methylation are associated with transcriptionally active regions [5]. However, depending on the methylation states and the genomic location the same modification might have different functional outcomes. H3K9 methylation is involved in euchromatic gene silencing as well as in heterocromatin formation [6, 7] while H3K27

methylation has an important role in the repression of genes during development [8] (*Figure 1*).

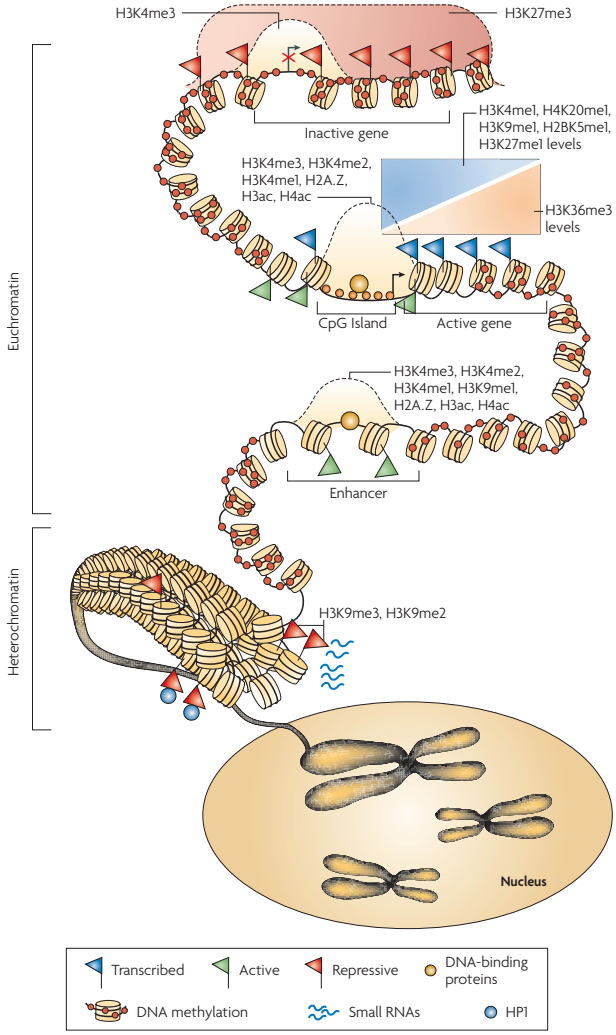


Figure 1. Characteristics of epigenomes. The interaction of DNA methylation, histone modification, nucleosome positioning and other factors such as small RNAs contribute to an overall epigenome that regulates gene expression and allows cells to remember their identity. Histone modifications including H3K4me3, H3K4me2, H3K4me1 as well as histone acetylation and histone variant H2A.Z mark the transcription start site regions of active genes. The H3K27me3 modification is present in broad domains that encompass inactive genes. The monomethylations of H3K4, H3K9, H3K27, H4K20 and H2BK5 mark actively transcribed regions, peaking near the 5′ end of genes. [9]

1.1.3. Chromatin immunoprecipitation sequencing (ChIP-seq)

Genome-wide mapping of protein–DNA interactions and epigenetic marks is essential for a full understanding of transcriptional regulation. The main tool for investigating these mechanisms is chromatin immunoprecipitation (ChIP), which is a technique for assaying protein–DNA binding *in vivo*. In a ChIP-seq experiment, DNA fragments associated with a specific protein are first enriched. The DNA-binding protein is cross-linked to DNA in-vivo by treating cells with formaldehyde and the chromatin is sheared by sonication into small fragments, which are generally in the 200-600 bp in range. An antibody specific to the protein of interest is used to immunoprecipitate the DNA-protein complex. Finally, the crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein. During the construction of a sequencing library, the immunoprecipitated DNA is subject to size selection (typically in the ~150-300 bp range, although there seems to be a bias towards shorter fragments in sequencing). The experimental steps in ChIP involve several potential sources of artefacts. Shearing of DNA, for example, does not result in uniform fragmentation of the genome: open chromatin regions tend to be fragmented more easily than closed regions, which creates an uneven distribution of sequence tags across the genome. Moreover, repetitive sequences might seem to be enriched because of inaccuracies in the number of copies of the repeats in the assembled genome. Also, ChIP-seq presents a bias in fragment selection towards GC-rich content, both in library preparation and in amplification before and during sequencing. For these reasons a peak in the ChIP–seq profile should be compared with the same region in a matched control sample to determine its significance. There are three commonly used types of control sample: input DNA (a portion of the DNA sample removed prior to immunoprecipitation (IP)), mock IP DNA (DNA obtained from IP without antibodies), and DNA from nonspecific IP (IP performed using an antibody, such as immunoglobulin G, against a protein that is not known to be involved in DNA binding or chromatin modification). To obtain accurate estimates throughout the genome, sufficient numbers of tags are needed at each point; otherwise fold enrichment at the peaks will result in large errors due to sampling bias. When an insufficient number of

reads is generated, there is a loss of sensitivity or specificity in detection of enriched regions. Thus effective analysis of ChIP-seq data requires sufficient coverage by sequence reads (sequencing depth). The required depth depends mainly on the size of the genome and the number and size of the binding sites of the proteins. In order to assess the sequencing depth, is has been suggested to progressively downsample the amount of reads of each sample and compute the fraction of peaks that are recovered respect to the total number of peaks called using the whole sample (saturation analysis). In this way it is possible to evaluate if depth e of sequencing has already reached a saturation point (plateau) or if it is necessary to sequence more the sequencing libraries. A typical workflow for the computational analysis of ChIP-seq data consist in the following steps: i) read mapping and quality assessment; ii) identification of statistically enriched regions (peak calling); iii) downstream integrative analysis. In the first step, before mapping the reads to the references genome, a quality cut-off has to be applied to filter out low quality sequences. Sometimes, after this step is necessary trim the end of low-quality reads. The remaining reads should then be mapped using one of the available mappers such as Bowtie [10] or BWA [11], accounting for a proper number of mismatches relative to the read lengths and discarding reads that map in multiple sites in genome, in order to retain as much informative reads as possible and avoiding introduction of noise [12]. One of the first steps in quality controls after read mapping is the check of the library complexity that is the measure of how the library is representative of the binding patterns along the genome. This step is performed viewing the read alignments on an appropriate genome browser (e.g. IGV) and measuring the fraction of nonredundant reads over the total mapped reads. Ideally each fragment of the sequencing library should represent an immunoprecipitation event of the binding protein of interest in a homogenous cell population. In case of low complexity, the library fragments are likely to be clonal, thus the sequencing steps start to resample continuously the same genomic loci, resulting in clonal reads. The low library complexity is linked to many factors such as antibody quality, over-cross-linking, amount of materials, sonication, or over-

amplification by PCR. After sequenced reads are aligned to the genome, the next step is to identify regions that are enriched in the ChIP sample relative to the control with statistical significance. Several *peak-calling* algorithms that scan the genome to identify enriched regions are currently available. These algorithms take the advantage of the directionality of the reads. The fragments are sequenced at the 5' end and the locations of mapped reads should inferred from two distributions, one on the positive strand and the other on the negative strand, with a consistent distance between the peaks of the distributions. In these methods, a smoothed profile of each strand is constructed and the combined profile is calculated either by shifting each distribution towards the center or by extending each mapped position into an appropriately oriented fragment and then adding the fragments together. Given a combined profile, peaks can be scored in several ways. A simple fold ratio between ChIP and Input DNA can provide important information, but it is not adequate because it does not take into account the depth of sequencing of the peak and the corresponding region in Input sample. To this end, statistical models (i.e. Poisson model or binomial model) that can take into account both parameters have been applied in peak calling algorithms A major difficulty in identifying enriched regions is that there are three types: sharp, broad and mixed. Sharp peaks are generally found for protein–DNA binding or histone modifications at regulatory elements, whereas broad regions are often associated with histone modifications that mark domains (e.g. transcribed or repressed regions). Tools such MACS [13] were developed to identify sharp peaks in contrast to SICER [14] that identify broader regions, while spp [12] was developed to identify peaks with mixed profiles. The statistical significance of enriched sites is generally measured by the false discovery rate (FDR) [15], which is the expected proportion of incorrectly identified sites (e.g. peaks identified in input sample) among those that are found to be significant. After obtaining high quality peaks, different approaches can be applied to analyse the biological implications of ChIP-seq data. One of the most common downstream analyses is the discovery of binding sequence motifs [16, 17], which allows the de novo discovery of binding motifs as well as the identification of known binding sites in broader regions, such as

H3K4me1 in enhancers. Another downstream analysis that can be performed using ChIP–seq data is to annotate the location of the peaks on the genome in relation to known genomic features, such as the transcriptional start site, exon–intron boundaries and the 3′ ends of genes. This analysis can be improved adding information about the expression levels of genes, to infer if the gene is a target of an activator, if a chromatin mark is enriched at the promoters of genes with high expression or if ChIP peaks are significantly enriched in a particular molecular function or biological process through Gene Ontology analysis [18, 19]. More advanced analysis includes the discovery of novel elements based on integration of different ChIP–seq data. In the last years, different bioinformatics tools have been designed to analyse ChIP-seq data and to characterize genome regulatory elements in both unsupervised (i.e. without prior information) and supervised ways, as Hidden Markov Models [20], dynamic Bayesian networks [21], profile methods [22] and intersection of enriched island [23]. However no standard method has been developed to detect *cis*-regulatory elements yet, due to different experimental design and to a lack of consensus on the rules to detect *cis*-regulatory modules. Overall, ChIP sequencing together with derived novel methods for downstream analysis, has become a principal tool for understanding transcriptional cascades and deciphering the information encoded in chromatin.

1.2. The transcriptome

A comprehensive understanding of biological networks is strictly dependent on the complete description of the cell transcriptome, that is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition, which represents a key link between information encoded in DNA and phenotype. The classic image of a mammalian transcriptome is composed of a large assembly of spliced [24] mRNAs, each structured with a capped 5' end, a 5' untranslated region (5' UTR), a coding sequence (CDS), a 3' untranslated region (3' UTR) and a poly-A tail. The sequencing project of human genome allowed the identification and the annotation of about 25,000 protein-coding genes [25]. The advances in high-throughput technologies currently challenge this view of mammalian transcriptome, which show an increase in complexity of the transcriptional output of cells. This complexity is reflected by the existence of alternative splice forms that may create functional diversity in protein isoforms presenting different domain combinations, the use of alternative gene starts and thereby potentially different regulatory mechanisms for different forms of the genes. In the last years different studies have shown that the mammalian genome is pervasively transcribed [26], which results in the active transcription of approximately 93% of the human genome [27]. Next generation sequencing technology, such as RNA-seq and CAGE-seq, have allowed the discovery of a broad range of ncRNAs as well as a huge number of genomic loci being transcribed. Recent studies have indeed highlighted the complexity of mammalian genome by revealing tens of thousands sites being transcribed to transcripts with little or no coding potential [28, 29]. The majority of transcripts identified do not encode for proteins and belonging to the class of noncoding RNAs (ncRNAs), like ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). Moreover, the transcriptome seems to double its size when considering the set of transcripts not having a poly-A tail [30]. NcRNAs can be divided in different categories on the basis of their functions: structural ncRNAs, i.e. ribosomal, transfer, small nuclear and small nucleolar RNAs, regulatory ncRNAs, i.e. micro RNAs (miRNAs), small interference RNAs (siRNAs), and long noncoding RNAs (lncRNAs). MiRNAs are 22 nucleotide (nt) RNAs that bind

predominantly to the 3' UTRs of mRNA, causing gene silencing; siRNAs are 21 nt long and also function in the degradation of complementary mRNAs; lncRNAs are transcribed RNA molecules greater than 200 nt in length that are implicated in post-transcriptional gene regulation through controlling protein synthesis, RNA maturation and transport, and also in transcriptional gene silencing through regulating the chromatin structure [31]. LncRNAs can be poly-A+ or poly-A-, are mainly located in the nucleus and can be classified accordingly to their placement respect to close protein coding genes as sense, antisense, bidirectional, intronic and intergenic [24]. Members of intergenic lncRNAs have been demonstrated to regulate epigenetic marks and thus gene expression [32, 33]. In addition, a novel class of RNAs transcribed at enhancers (eRNAs) has been described [34]. The size of eRNAs has been shown to range from 0.1 to 9 kb, with an average size of 800 nt [35], placing most of the eRNAs into a subgroup of lncRNAs. These transcripts elongate both bidirectionally or unidirectionally and can lead to the production of both poly A+ or poly A- transcripts; eRNAs have a short half-life and are positively correlated to levels of nearby mRNA expression, this suggests a possible functional role for eRNAs as transcriptional activators.

1.2.1.The study of the transcriptome through Cap Analysis of Gene Expression (CAGE) sequencing

Various technologies have been developed to deduce and quantify the transcriptome, including hybridization or sequence-based approaches. Hybridization-based approaches on arrays have several limitations, such as the reliance upon existing knowledge about genome sequence, high background levels owing to cross-hybridization, and a limited dynamic range of detection owing to both background and saturation of signals. On the contrary, sequence-based approaches directly determine the cDNA sequence. In recent years several technologies have become available to sequence the DNA at a high throughput levels, and have been applied together with different protocols and methodologies to study the genome and the transcriptome. Among them a new high-throughput approach to gene expression analysis is *Cap Analysis of Gene Expression* (CAGE) [36],

that was introduced as method to determine transcription start sites on a genome-wide scale by isolating and sequencing short sequence tags originating from the 5' end of RNA transcripts. Mapping these tags back to the reference Genome identifies the transcription start sites from which the transcripts originated. CAGE relies on a cap-trapper system to capture full-length RNAs while avoiding rRNA and tRNA transcripts. First, an oligo-dT primer is used to reverse-transcribe poly-A terminated RNAs. Alternatively, a random primer can be used for RNAs without a poly-A tail, which may constitute almost half of the transcriptome. Subsequently, full-length cDNA are select by biotinylated cap-trapped their 5' cap structure, allowing capture by streptavidin-coated magnetic beads. Ligation of a specific linker sequence containing an *Mme1* recognition site to the 5' end of the full-length cDNA creates a restriction site about 20 nucleotides downstream, producing a short CAGE tag starting at the 5' end of eukaryotic mRNAs. Then, 5' ends tags ('CAGE tags') are isolated and after the cleavage, sequences tags are purified and sequenced. At high sequencing depths significantly expressed TSSs are typically sequenced a large number of times. It thus becomes possible to not only map the locations of TSSs but also quantify the expression level of each individual TSS, then CAGE it is a unique tool in the analysis of transcriptional regulatory networks. The first step in the analysis of deep-sequencing expression data is the mapping of the (short) reads to the genome from which they derive. CAGE tags are usually mapped to a reference genome using a novel alignment-algorithm called Kalign2 that maps tags in multiple passes [37]. Once the RNA sequence reads or CAGE tags have been mapped to the genome, the next step is to obtain a collection of positions for which at least one read/tag is present. In multiple samples it could be obtain, for each position, a read-count or tag-count profile that counts the number of reads/tags from each sample, mapping to that position. These tag-count profiles quantify the expression. CAGE-tag based expression measurements directly link the expression to individual TSSs, thereby providing a optimal guidance for analysis of the regulation of transcription initiation. Large-scale sequence analysis, performed in the FANTOM3 project (http://fantom.gsc.riken.jp/), made clear that most genes are transcribed in different isoforms that use

different TSSs. Alternative TSSs not only involve initiation from different areas in the gene locus, but TSSs typically come in local clusters spanning regions ranging from a few to over 100 bps. Hence, these tag clusters are defined by a start and end position, have a count of tags and a distribution of these counts, and they are the representation of the promoters. An analysis of CAGE-defined TSSs in human and mouse accomplished by Carninci et al. [38], illustrate that tag clusters can be divided into different shape classes. The two main types identified are the single peak (SP) class, that is characterized by sharp peak, indicative of a single, well defined TSS, and the broad (BR) shape class, that is characterized by multiple, weakly defined TSSs. Other shape classes can be described as subtypes of the broad class or hybrids between broad and sharp class. Specifically, in single peak (SP) class promoters the majority of tags are concentrated to no more than four consecutive start positions, giving a single dominant TSS. This class of promoters is generally associated with TATA boxes. Importantly, only a minor fraction (<25%) of promoters belong to this class. Instead, the broad (BR) shape class promoters have broad distribution of TSSs generally spread over 100 nt. They are strongly associated with CpG islands and are GC rich. This study showed that more than half of protein coding transcriptional units had two or more alternative promoters, based on the presence of non-overlapping tag clusters.

1.3. Transcriptional regulatory elements

The execution of biological processes such as development, proliferation, apoptosis, aging, and differentiation requires a precise and carefully orchestrated set of steps that depend on the proper spatial and temporal expression of genes. Interpretation of genomic information involves integration of cellular history and extracellular environment, which ultimately occurs at the level of chromatin and is mediated by the functionally diversified *cis*-regulatory elements, such as promoters, enhancers, silencers, and insulators (*Figure 2*).

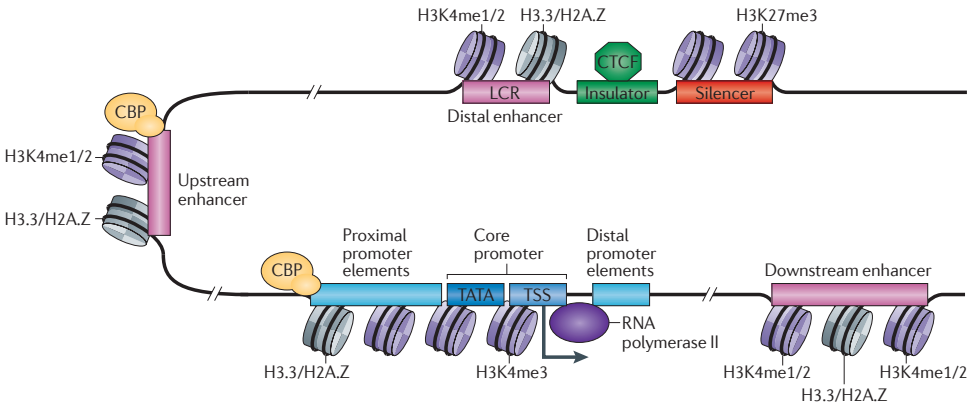


Figure 2. Transcriptional regulatory elements in metazoans. The promoter is typically comprised of proximal, core and downstream elements. Transcription of a gene can be regulated by multiple enhancers that are located distantly and interspersed with silencer and insulator elements. H3K4me1/2, histone H3 mono- or dimethylation at lysine 4; H3K4me3, histone H3 trimethylation at lysine 4; H3K27me3, histone H3 trimethylation at lysine 27; H3.3/H2A.Z, histone variants H3.3 and H2A.Z; LCR, locus control region; TATA, 5′-TATAAAA-3′ core DNA sequences; TSS, transcription start site. [39]

1.3.1. Promoters

Gene core promoters are the regions overlapping transcription start sites (TSSs) of genes that serve as the docking sites for the basic transcriptional machinery, and define the position of the TSSs and the direction of transcription. Core promoters are generally characterized by the presence of multiple elements, which are recognized by different subunits of the preinitiation complex (PIC) [40]. The proximal promoter is defined as the

region immediately upstream (up to a few hundred base pairs) from the core promoter, and typically contains multiple binding sites for activators that act synergistically on transcriptional regulation.

RNA polymerase II (RNAPolII)-transcribed genes are highly heterogeneous with respect to expression level and context specificity and this is reflected by the existence of different architectures, which in turn determine the promoter functions and regulation types. In fact metazoan promoters can be divided in three main classes, depending on particular characteristics. Type I promoters are defined “sharp promoters” because exhibit a precise start site at which most transcription initiates. They have low frequency of CpG islands and possess a TATA box, an AT enriched sequence that is binding site for the TBP subunit of TFIID. These promoters are associated to tissue specific genes. H3K4me3 is generally present downstream of the TSS and there is no RNAPolII binding at these sites when genes are not expressed [41, 42]. Type II promoters are associated with ubiquitously expressed genes and usually possess a unique CpG island overlapping the TSS. These promoters are usually defined “broad promoters”, since they have dispersed TSS [43]. At these genes H3K4me3 histone mark distribution is almost identical with the span of CpG islands, overlapping the 5’ end of the genes [44]. Type III promoters are sharper than type II promoters, are associated to developmental genes, and have multiple features associated with repression, such as large CpG island that extend in the gene bodies, binding of Polycomb group (PcG) proteins, and presence of both H3K4me3 and H3K27me3, associated with activation and repression, respectively [45]. Because of this last characteristic, these promoters are defined “bivalent promoters”. These three main classes of promoters are preferentially associated with different subset of epigenetic states. While tissue-specific genes (type I promoters) seem to depend mostly on *cis*-regulatory modules for regulation, and typically are active or inactive, the housekeeping genes (type II) have few enhancers in neighbouring regions and are generally characterized by active configuration. Developmental genes (type III) are regulated at both promoter and enhancer levels, have high number of enhancers associated, and have heterogeneous promoter states (e.g. poised and repressed

states) [20]. These types of promoters also show different pattern of nucleosome occupancy and positioning, reflecting the different mechanisms of regulation, to which they may be subjected.

1.3.2. Distal regulatory elements

Temporal and tissue-specific gene expression in mammals depends primarily on distal regulatory elements (i.e., enhancers, silencers, insulators, and LCR), which are often located far away from the genes they control .

1.3.2.1. Enhancers

Enhancers play a central role in driving cell-type-specific gene expression and are capable of activating transcription of their target genes at great distances, ranging from several to hundreds, in rare cases even thousands, of kilobases [46]. Recent advances in chromatin profiling methodologies have revealed more than 400,000 of these regulatory elements across several cell types [47], suggesting an enormous combinatorial complexity of expression patterns during human development. These distal elements are viewed as clusters of DNA sequences capable of binding combinations of transcription factors that then interact with components of the Mediator complex (TFIID) to help recruit RNAPolII, and can enhance transcription independent of their location, distance or orientation with respect to the promoters of genes [48]. However, it is not clear whether the enhancer-mediated delivery of factors is predominantly required to initiate transcription and/or to continuously sustain gene expression. Among the enhancers-interacting protein factors, there are also histone modifying enzymes or ATP-dependent chromatin remodeling complexes that alter chromatin structure and increase the accessibility of the DNA to other proteins [49]. Enhancers can be associated with high dynamic nucleosomes with histone variants H3.3 and H2A.Z [50]. Nucleosomes directly flanking TF binding sites are less mobile and usually marked with H3K4me1 and H3K27ac, two modifications that have been extensively used to identify active enhancers in several studies [51]. H3K4me1 is also present at 5' of actively transcribed promoters, but differentially from these, enhancers do not have H3K4me3 signature, because of the lack of CpG

islands at enhancers, which are preferentially recognized by CxxC domain of H3K4me3 methyltransferases complexes. This histone modification premarks enhancers prior to their deployment upon differentiation in different models, such as hematopoietic system; in fact the presence of H3K4me1 often precedes nucleosome depletion, deposition of H3K27ac, and enhancer activation [52, 53]. H3K4me1 is thought to block repressor protein binding, such as DNA methyltransferases or histone deacetylases, or as binding platform for specialized effector/coactivator proteins, such as histone acetyltransferases. Enhancers are indeed bound by p300 and CBP, two highly homologous HATs that has been used for genome-wide enhancer mapping in multiple cell types [54]. These proteins are recruited at enhancers by a broad range of sequence-dependent activator factors, and have H3K27 as main substrate [55]. Acetylated lysine residues are recognized by bromodomains, present in diverse nuclear proteins, including HATs themselves (e.g. p300, CBP, PCAF, and Gcn5), ATP-dependent remodelers (e.g. BRG1), TFIID components, and factors regulating transcriptional pause release. Overall, histone acetylation directly affects enhancer function through attenuating nucleosomal stability, promoting chromatin decompaction and/or regulating enhancer-promoter communication. A large subclass of enhancers lacks H3K27ac end is enriched in H3K27me3 and bound by Polycomb complex PRC2. These regulatory elements are termed “poised enhancers” and are typically found near early developmental genes, which has bivalent promoters [53]. Poised enhancers show similar patterns of active enhancers (p300, BRG1, similar nucleosome depletion levels), except that are unable to drive gene expression until H3K4me3 is removed in favour of the gain of H3K27ac.

1.3.2.2. Silencers, insulators and LCRs.

Other regulatory DNA elements that contribute to the formation and maintenance of active or inactive transcription programs and play an integral part in gene regulation ere silencers, insulators and locus control regions (LCRs).

Silencers are sequence-specific elements that confer a negative (i.e., silencing or repressing) effect on the transcription of a target gene. They

function independently of orientation and distance from the promoter and they can be located as part of a proximal promoter, as part of a distal enhancer, or as an independent distal regulatory module; in this regard, silencers can be located far from their target gene, in its intron, or in its 3'-untranslated region. Silencers bind transcription repressors acting through inhibition of gene transcription. In some cases, repressors appear to function by blocking the binding of a nearby activator, or by directly competing for the same site. A repressor may prevent the general transcriptional machinery from accessing a promoter by establishing a repressive chromatin structure through the recruitment of histone-modifying activities or by inhibiting PIC assembly [56].

Insulators (also known as boundary elements), function in a position-dependent, orientation-independent manner to block genes from being affected by the transcriptional activity of neighbouring genes. They thus create boundaries in chromatin limiting the action of transcriptional regulatory elements into defined domains, and partition the genome into discrete domains of expression. There are two types of insulators: enhancer-blocking insulators, which prevent communication between discrete sequence elements (typically enhancers, or even silencer, and promoters) when positioned between them, and barrier insulators, which prevent the spread of heterochromatin. The core insulator fragment contains binding sites for several transcription factors; in vertebrates, the only known insulator protein is CCCTC-Binding Factor (CTCF) that is necessary for enhancer-blocking activity. It is proposed that CTCF creates distinct loop domains, in which the enhancer in one loop is unable to contact a promoter in a different loop [57, 58].

Finally, developmental and cell lineage-specific regulation of gene expression relies upon Locus control regions (LCRs), that are groups of regulatory elements involved in regulating an entire locus or gene cluster. Locus control regions are operationally defined as elements that enhance the expression of linked genes to physiological levels in a tissue-specific, position-independent and copy-number- dependent manner. LCRs are often marked by a cluster of nearby DNase I hypersensitive sites (HS) and are thought to provide an open-chromatin domain for genes to which they

are linked. LCRs are typically composed of multiple *cis*-acting elements, including enhancers, silencers and insulators. These elements are bound by transcription factors (both tissue-specific and ubiquitous), coactivators, repressors, and/or chromatin modifiers. Each of the components differentially affects gene expression, and it is their collective activity that functionally defines a LCR and confers proper spatial/temporal gene expression. The final most prominent effect of the LCRs is a strong, transcription-enhancing activity. The identification of a large number of LCRs has revealed that, although LCRs are typically located upstream of their target gene(s), they can also be found within an intron of the gene they regulate, downstream of the gene, or even in the intron of a neighbouring gene [59].

1.4. Hematopoiesis

Hematopoiesis is a term used to describe the process of blood cell formation during both the embryonic and adult stages of an organism. Hematopoiesis is also viewed as the process of development, self-renewal and differentiation of hematopoietic stem cells (HSCs), a type of adult stem cell that is the source of all blood cell lineages.

1.4.1. The classical model of hematopoiesis

Hematopoiesis is usually depicted in a hierarchical way, with HSCs giving rise first to progenitors and then to precursors with varying commitments to multiple or single pathways. In this hierarchical schema hematopoietic cells can be broadly classified as transiting through three steps: stem cells, committed progenitor cells, and precursors of mature functional-end cells. Although this representation may oversimplify the hematopoietic process, it can provide a useful framework in which divide and classify the intermediate cells (*Figure 3*).

At the top of this hierarchy, the stem cell compartment is composed of very rare cells with the ability to self-renew and differentiate into multilineage precursors. Lineage markers are absent from these cells, and they normally are found in a quiescent state or are turning over very slowly. Stem cells are also equipped with a regimen of critical transcription factors that are important in the execution of their fundamental cellular functions of cell renewal and multilineage differentiation.

The progenitor cell compartment contains cells that are found at a higher frequency than the stem cell pool and, like the stem cell, are not morphologically distinguishable. Their existence is revealed by their ability to give rise to differentiated progeny *in vitro* in well-defined functional assays. The progenitor cell compartment is derived from stem cells through a process of commitment to different lineage pathways. Transition of stem cells to cells of the committed compartment is achieved not by acquisition of new characteristics or new proteins but by enhancement of certain molecular pathways, already primed in these cells, and abrogation of others. As progenitor cells differentiate, they acquire more distinctive features characteristic of each lineage and move away from shared

primitive progenitor characteristics: they show the enhancement of lineage-specific features, with a diminished or absence of expression of multilineage properties. In this manner, precursors for the various lineages arise.

The precursor cell compartment is defined by morphological criteria and contains cells at different maturation stages; these precursors can be further distinguished by cell-surface markers. Precursor cells for each lineage follow a unique maturation sequence. The morphological characteristics of these cells reflect the accumulation of lineage-specific proteins, and organelles and the decline of nuclear activity, which gives them a unique appearance.

A cellular roadmap that specifies lineage relationships between stem, progenitor, precursor and mature cells is indispensable for a comprehensive view of the transcriptional and epigenetic mechanisms that control normal development. In this regard the first comprehensive ‘‘classical’’ model of hematopoiesis was formulated. The first key postulate of this model is that loss of self-renewal capacity by HSCs during differentiation precedes lineage commitment. At this step multilineage progenitors (MP) remain multipotent, but possess only transient capacity of repopulation. The second postulate states that MPs segregate in the two branches of the hematopoiesis model: lymphoid and myeloid. This earliest myelo-lymphoid split gives rise to common myeloid progenitors (CMPs) and common lymphoid progenitors (CLPs) and each of these undergo further commitment steps. On the lymphoid side CLPs give rise to B cell precursors and the earliest thymic progenitors (ETPs) committed to the T and NK lineages, while on the myeloid one CMPs give rise to granulocyte macrophage progenitors (GMPs), which become committed to the granulocyte (neutrophils, eosinophils, mast cells, and basophils) and monocyte fates, and megakaryocyte-erythrocyte progenitors (MEPs), which will eventually produce erythroid and megakaryocyte cells. This classical model is a simple yet powerful template for understanding blood development and interpreting the function of molecular regulators.

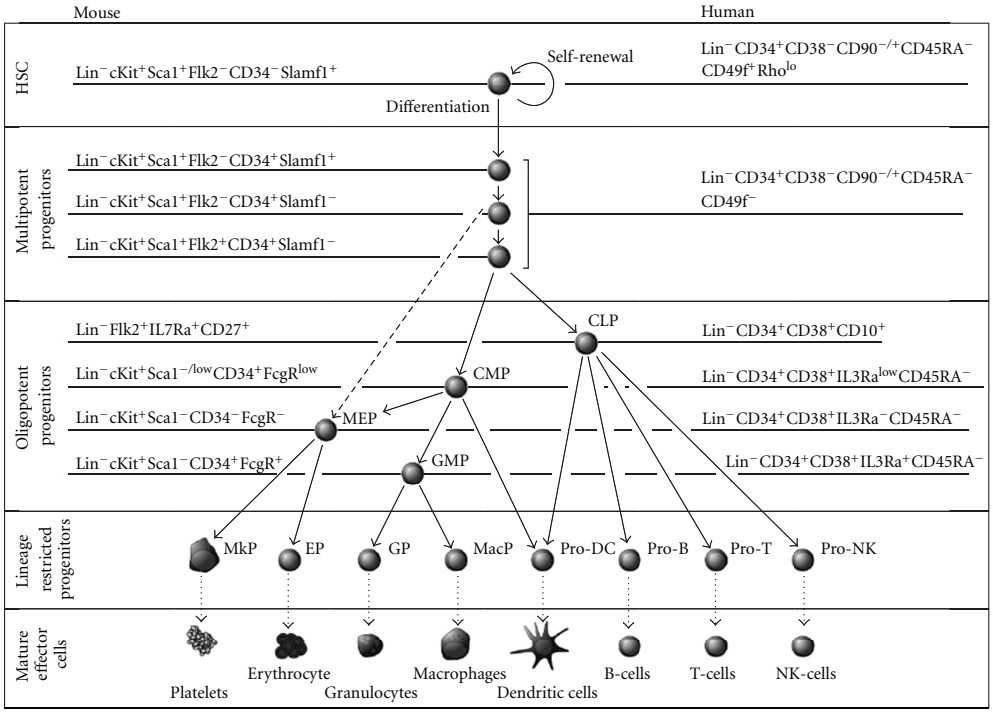


Figure 3. Hierarchy of hematopoiesis. The phenotypic cell surface marker of each population of mouse and human blood system is shown. In the mouse hematopoiesis system, MPPs omit CMPs which directly give rise to MEPs unlink in the human system (dashed line). CLP, common lymphoid progenitor; CMP, common myeloid progenitor; DC, dendritic cell; EP, erythrocyte progenitor; GMP, granulocyte-macrophage progenitor; GP, granulocyte progenitor; HSC, hematopoietic stem cell; MacP, macrophage progenitor; MEP, megakaryocyte-erythrocyte progenitor; MkP, megakaryocyte progenitor; NK, natural killer; Lin, lineage markers. [60]

1.4.2. The hematopoietic stem cell

HSCs are defined in an operational sense as cells competent to reconstitute the entire haematopoietic system of an individual. HSCs are capable of self-renewal and differentiation in all cell types that constitute the blood tissue. These two processes are strongly dependent on the microenvironment in which HSCs reside. The production of these cells is accomplished by the allocation and specification of distinct embryonic cells in a variety of sites that change during development [61]. In mammals, the

sequential sites of hematopoiesis include the yolk sac, the aorta-gonad mesonephros (AGM) region (an area surrounding the dorsal aorta), the fetal liver, and the bone marrow. HSCs arise and migrate between these multiple sites until the bone marrow develops sufficiently to provide the environmental niches necessary for HSC function. The properties of HSCs in each site differ, presumably reflecting diverse niches that support HSC expansion and/or differentiation and intrinsic characteristics of HSCs at each stage. The fetal liver and the bone marrow are major organs for HSC expansion, maintenance and differentiation. Currently, two types of HSC niches have been identified in bone marrow: the endosteal niche located on the surface of trabecular bone, and the vascular niche at the bone marrow sinusoids, which are low-pressure blood vessels with a fenestrated endothelium located in the center of the bone marrow [61]. Osteoblasts are mesenchymal cells that produce the bone matrix to form the bone after mineralization, are found on the endosteal surface lining between the bone and the marrow, and secrete many cytokines that promote the proliferation of hematopoietic cells in culture, and support the in vitro maintenance of HSCs [62]. Osteoblasts probably regulate HSCs through cell-surface adhesion molecules and/or secreted signalling molecules, such as N-cadherin and β1-integrin, important in anchoring HSC to the endosteal niche [63]. Osteoblasts also produce surface signalling ligands, such as Angiopoietin-1, which interact with HSC receptor Tie2. Thus osteoblasts probably regulate HSCs through cell-surface adhesion molecules and/or secreted signalling molecules. Endothelial cells and reticular cells were proposed to act as the niche cells that interact with HSCs within the sinusoidal vascular niche. They express high levels of the chemokine CXCL12 (also known as stromal-derived factor 1, SDF-1) that regulates the migration of HSCs to the vascular niche [64]. Different observations stated that long-term retaining HSCs are highly enriched at the bone surface compared to the center of marrow, raising the hypothesis that in endosteal niche HSCs are probably more quiescent than the HSCs residing in the vascular niche [65]. These characteristics suggest that HSCs reside in various sites within the marrow and that their function might depend on their precise localization.

1.4.2.1. Phenotypic characterization of HSCs

The major obstacle to studying HSC biology is that the cells are extremely rare. Only 1 in 10^6 cells in human bone marrow are HSC [66], requiring purification from the bulk of differentiated cells. Purification of human HSCs requires simultaneous detection of several independent cell surface markers, so human HSCs and progenitors isolation have been characterized by enriching a rare cell population with a combination of monoclonal antibodies [67]. Over past years, various cell surface and metabolic markers have been identified and used to isolate human HSCs and progenitors. CD34 was the first identified marker found to enrich human HSCs as well as more differentiated progenitors, raising the need to search for additional markers for classification. To this end, further studies introduced CD90 (Thy1) as a stem cell marker and CD45RA and CD38 as markers of more differentiated progenitors that negatively enrich for HSCs, thus depicting human HSCs as $CD34^+CD38^-Thy1^+CD45RA^-$ [68, 69]. Since other studies discovered subpopulation of HSCs that do not express detectable levels of CD34, another important marker of HSCs became CD133 (also known as PROM1), which is expressed on both $CD34^+CD38^-$ cells and primitive $CD34^-CD38^-$ subpopulations, therefore providing a more appropriate method to enrich stem cells than CD34 selection alone [70].

1.4.2.2. Molecular regulation of HSC formation and self renewal

One of the major objectives of stem cell biology is to understand the cellular and molecular processes underlying HSCs fate choice. As intrinsic determinants of cellular phenotype, transcription factors provide an entry point for unravelling how HSCs develop during embryogenesis and how lineage-restricted differentiation is programmed. They can act both positively and negatively to regulate the expression of a wide range of genes including growth factors and their receptors, other transcription factors, as well as various molecules important for the function of developing cells. It is the alternative expression of specific combinations of transcription factors that determines the survival, proliferation, commitment, and differentiation responses of hematopoietic progenitors to signals that arise from extrinsic or intrinsic regulatory factors. For discussion purposes it

is possible to divide between factors required for HSC formation or function and those employed in lineage-specific differentiation [71]. Among the transcription factors involved in HSC formation are MLL (for mixed lineage-leukemia gene), Runx1, TEL/ ETV6, SCL/tal1, and LMO2, all genes that are subject to translocations in leukemia patients. These factors have served roles later within differentiation of individual blood lineages, while factors that appear to have more lineage-restricted roles, such as *PU.1, Gfi-1*, and *C/EBP*a, act within HSCs. The basic-helix loop-helix (bHLH) factor SCL/tal-1 and its associated protein partner, LMO2, are individually essential for development of both the primitive and adult hematopoietic systems [72]. The genes encoding the SET-domain containing histone methyltransferase MLL and Runx1 proteins are essential for generation of HSCs within the AGM (and possibly at other sites) [73]. In the absence of Runx1, no hematopoietic clusters (representing presumptive HSCs) form in the dorsal aorta in mice. GATA2 is an indispensable transcription factor for hematopoiesis [74]. In mice, GATA2 is expressed in primitive hematopoietic cells and its level gradually declines during cell maturation into different blood cell lineages. GATA2-deficient embryos are anemic, have reduced numbers of hematopoietic progenitors and die prematurely. This transcription factor has at least two fundamental functions: the production and expansion of HSC in AGM, and the proliferation and survival of HSCs in adult bone marrow. The balance that controls between self-renewal and cell fate decision of HSCs in the bone marrow is mediated by several factors. Most HSCs are in quiescent state (i.e., in G0/G1 phase of the cell cycle), however, when the hematopoietic cells disturbance occurs, hematopoiesis system will respond by regulating several signal transduction pathways, such as SDF-1 (CXCL12)/CXCR4 signaling, BMP signaling, Mpl/Thrombopoietin (TPO) signaling, Tie2/Ang-1 signaling, hedgehog and Notch signaling, as well as Wnt signaling [75], to promote HSC self-renewal. Moreover, several transcription factors seem to be involved in the same process. For instance factors of the HOX and Ikaros families appear to be strong positive regulators of HSC self-renewal [76, 77].

1.4.2.3. Molecular regulation of lineage commitment

Once HSCs divide and generate more differentiated daughter cells, within 10-15 divisions the genetic programs of the descendent cells become fixed toward a single lineage. The first step is represented by the restriction to the ability to generate myeloid versus lymphoid progeny respectively by CLPs, which express specifically GATA-3 and IKZF3 transcription factors, and CMPs, which are specifically regulated by the transcription factors SCL, GATA-2, NF-E2, GATA-1, C/EBPa, c-Myb, and PU.1 [78]. The alternative differentiation potential of these progenitors is not immediately eliminated, but rather repressed in a graded or gradual way in the cells that are committing to a given lineage. Whereas PU.1 and GATA-1 are co-expressed in CMPs, their mutually exclusive expression coincides with further commitment to either granulocytic-monocytic or megakaryocytic-erythroid differentiation. PU.1 was found to inhibit the transcriptional activity of GATA-1 upon its target genes, and vice versa. These effect are mediated by direct physical contact of PU.1 N-terminal part to the GATA-1 C-terminal zinc finger that involves DNA binding, and, in a reverse way, through the interaction of the c-finger of GATA-1 with ETS domain of PU.1, precluding the recruitment of co-activator c-Jun in the context of the promoters of its target genes [79]. Models in which such transcription factors act as part of multimeric complexes best explain the context-dependent action and direct antagonism between key transcriptional regulators. Indeed protein complex in erythroid cells comprised of GATA1 (or its close relative, GATA-2), LMO2 and its partner Ldb1, and SCL/tal1 and its heterodimeric partner E2A, recognizes a consensus GATA-E-box DNA motif. This complex is required for full erythroid and megakaryocytic cell maturation [72]. GATA factors form alternative protein complexes with a specific cofactor known as FOG (for friend of GATA) that is also essential for erythroid and megakaryocytic development. The GATA1 (or GATA-2)/FOG1 complex in both erythroid and megakaryocytic lineages is physically associated with the NuRD chromatin remodelling complex via a NuRD binding motif at the N terminus of FOG1 [80]. In the context of the development of myeloid cells the dosage of PU.1 is important in defining cellular fates in the myeloid lineage. Approximately, low doses of PU.1

promote the development of granulocytes, while higher ones favor the development of monocytes. CCAAT enhancer-binding protein-α (C/EBP-α), a basic region leucine zipper transcription factor, is important for the production of granulocyte-macrophage progenitors from common myeloid progenitors. IFN-γ–responsive factor (IRF)–8 promotes differentiation into monocytes but not granulocytes, whereas GFI-1 is a transcription factor required for granulocyte development. In addition to transcriptional regulators, cell surface cytokine/growth factor receptors and proteins play a role in regulating erythroid and myeloid development [81].

$CD34^+$ cells are a heterogeneous population covering not only stem cells but also earlier multipotent progenitors and later lineage-restricted progenitors. In human hematopoiesis, populations enriched for HSC activity have been identified, but multipotent progenitor activity has not been uniquely isolated. There is reason to believe that cells defined as human HSCs by the cell surface markers previously described are likely to include one or more multipotent progenitor populations. A previous study demonstrated that the $CD34^+$ Lin^- $CD38^-$ fraction of cord blood and bone marrow can be subdivided into three subpopulations: $CD90^+$ $CD45RA^-$, $CD90^-$ $CD45RA^-$, and $CD90^-CD45RA^+$. The $CD90^+$ $CD45RA^-$ subpopulation contains HSCs, while the $CD90^-$ $CD45RA^-$ subpopulation contains candidate multipotent progenitors [82]. This represented the first identification and prospective isolation of a population of candidate human MPPs. Recently it has been reported that CD49f antigen is expressed on about 50% of human $CD90^+$ and about 25% of $CD90^-$ cells. When sorted fractions were assayed in vivo, HSC activity was restricted to the $CD49f^+$ cells in both fractions. By contrast, $CD90^-$ $CD49f^-$ cells mediate transient multilineage repopulation that peaks at 4 weeks and becomes undetectable after 16 weeks [83]. In light of these studies, several markers, such as CD90, CD45RA and CD49f, can be used to distinguish MPPs from HSCs.

The erythropoiesis starts with committed erythroid progenitors, the BFU-E (burst forming unit-erythroid), which expresses the cell surface antigen, CD34, as do all other early hematopoietic progenitors, but is CD36 positive. The next stage, the colony forming unit-erythroid (CFU-E), express also high level of CD71 antigen (the transferrin receptor), which then decreases

slightly during further maturation in pro-erythroblast, characterized by the presence of a euchromatic nucleus and visible nucleoli and the expression of glycophorin A on cell surface. This process is mainly due to the effect of EKLF transcription factor.
On the myeloid side, GMPs differentiate from the CMPs. GMPs are CD34+ cells that are committed to differentiate into myeloid cells and can be commonly recognized by the surface expression of CD13 and CD33 antigens on cell surface. Then these cells give rise to colony forming unit-granulocyte (CFU-G) and colony forming unit- monocyte (CFU-M), mainly through the coordinated action of PU.1 and Gfi-1 transcription factors.

1.4.2.4. Chromatin landscapes in HSC differentiation

In recent years different pathways and genes involved in hematopoiesis have been identified, but it remains unclear how the decision between self-renewal and differentiation is controlled and how differentiation is specified at molecular levels. Recently the epigenetic profile of HSCs has been described, and some hypotheses about differentiation mechanisms were outlined [84]. In particular it has been shown that differentiation of $CD133^+$ cells into $CD36^+$ cells is accompanied by dramatic changes of histone modifications at critical genetic regions. For instance H3K4me3 and H3K27me3 counteracting modifications provide a mechanism of either gene activation or repression through a shift in their balance. These bivalent modifications maintain the activation or silencing potential of critical differentiation genes, in fact some of these “bivalent genes” can lose H3K4me3 while others can lose H3K27me3 during differentiation. In particular 20% of bivalent genes lose H3K27me3 becoming activated in transition to CD36, are associated to H3K4me1 and RNA PolII in HSCs, thus are likely to be poised for activation. These characteristics are consistent with a model in which the fate of bivalent genes (i.e. the induction of genes involved in one particular lineage and the silencing of genes required for other lineages) during differentiation is controlled by epigenetic modifications that occur in HSC commitment to hematopoietic progenitor cells (HPCs). Another chromatin-related striking characteristics of HSC commitment is the enhancer “priming” by specific histone

modifications, such as H3K4me1, H3K27me1, and H3K9me1 that are associated with critical regulatory elements in the cell stage before their target genes are expressed. In fact these marks are found to be present in LCR upstream human β-globin locus even if the globin genes are not expressed. Specific signatures at enhancers are also removed after differentiation, as in the case of neutrophil-specific myeloperoxidase gene (Mpo) that is primed for expression in HSCs through H3K4me1 and H3K9me1 marks in a region upstream the TSS, but does not show the same signature in $CD36^{+}$ cells, causing a loss of activation potential after differentiation. Although the model of gradual transition from an open chromatin state in multipotent stem cells to a compacted state in more differential cell, progressively closing the regulatory potential of the genome, is widely accepted, it cannot explain some characteristics of the regulatory regions during HSCs differentiation. Recently, in fact, it has been proposed that enhancers can be *de-novo* specified during establishment of lineage progenitors through the action of pioneer factors that recognize specific sites in the genome, deposing H3K4me1 and eventually modify the chromatin landscape [85]. Together with H3K4me1 deposition, active enhancers are also marked by H3K27ac, but while the monomethylation usually appear first in the root lineage progenitors, the acetylation of H3K27 is acquired later, together with activation of regulated genes.

1.5. Aim of the work

The goals of this work are:

1. Characterization of the transcriptional and epigenetic profiles of HSPCs and their early committed erythroid and myeloid progeny;

2. Identification of active regulatory elements, responsible of the early commitment of HSPCs to erythroid and myeloid lineage-restricted progenitors;

3. Development of a bioinformatic pipeline to analyse and integrate quantitative genomics data from epigenetic and transcriptomic experiments involving the application of NGS technologies in order to reach these goals.

2. MATERIALS AND METHODS

2.1. Cell types

Thanks to the collaboration with San Raffalele Hospital (Milan, Italy), we obtained human umbilical cord blood from informed healthy donors in accordance with the Declaration of Helsinki. San Raffaele Scientific Institute Ethical Committee approved the study. Mononuclear cell were isolated using gradient separation (by Ficoll-Hypaque, Lymphoprep; Sentinel Diagnostics) and $CD34^+$ cells were purified by immunomagnetic sorting (EasySep Human CD34 Positive Selection kit, StemCell Technologies Inc.).

2.1.1. Hematopoietic stem/progenitor cells (HSPCs)

We seeded $CD34^+$ cells at 0.5-1x10^6 cells/ml and cultured them for 36h in IMDM medium (Lonza) containing 20% fetal bovine serum (FBS) (Hyclone) and supplemented with 100 ng/ml human stem cell factor (hSCF), 100 ng/ml human Flt3-ligand (hFlt3-l), 20 ng/ml human trombopoietin (hTPO) and 20 ng/ml human Interleukin-6 (hIL-6) (all PeproTech). After 36h, we labeled the cells with fluorescein isothiocyanate (FITC)-conjugated anti-CD34, phycoerythrin (PE)-conjugated anti-CD133 and tri-color (TC) anti-CD38 antibodies. We sorted $CD34^+$/$CD133^+$ multipotent progenitors using a MoFlo cell sorter (Beckman Coulter).

2.1.2. Erythroid progenitors (EPPs)

We cultured $CD34^+$ cells for 5 days as described by Roselli *et al*. [86]. We seeded the cells at 10^5 cells/ml in StemSpan medium (Stem Cell Technologies) containing 20% FBS (Hyclone) and supplemented with 50 ng/ml hSCF (Peprotech), 1 U/ml human erythropoietin (EPO) (Janssen), 1 ng/ml hIL-3 (Peprotech), 10^{-6} M dexamethasone (Sigma), and 10^{-6} M ß-estradiol (Sigma). At day 5, we labeled the cells with FITC-conjugated anti-CD36 antibody and sorted CD36+ erythroid progenitors using a MoFlo cell sorter (Beckman Coulter). We also stained $CD36^+$ cells with PE-conjugated anti-CD34, Allophycocyanin (APC)-conjugated anti-glycophorin A and Peridinin chlorophyll- conjugated (PerCP) anti-CD71 antibodies.

2.1.3. Myeloid progenitors (MPPs)

We cultured $CD34^+$ cells at 10^5 cells/ml in IMDM medium (Lonza) containing 10% FBS (Hyclone) and supplemented with 100 ng/ml hSCF (Peprotech), 20 ng/ml hIL-3 (Peprotech), 100 ng/ml G-CSF (ITALFARMACO SpA). At day 5, we stained the cells with FITC-conjugated anti-CD34 and PE-conjugated anti-CD13 antibodies and purified $CD13^+/CD34^-$ MPP using a MoFlo cell sorter (Beckman Coulter). We also labelled MPP with APC-conjugated anti-CD33 or APC-conjugated anti-CD11b antibodies.

We performed analyses by fluorescence-activated cell-sorter (FACS) using FACSCanto flow cytometer (BD Biosciences). Appendix table 6.1 reports the complete list of antibodies used for FACS.

2.1.4. CFU assay

We plated multipotent and lineage–committed progenitors cells at 10^3 cells/ml in methylcellulose medium (GFH4434, Stem Cell Technologies). We scored burst forming unit-erythroid (BFU-E), unit-granulocyte/ macrophage (CFU-GM) and unit-granulocyte/erythroid/macrophage/ megakaryocyte (CFU-GEMM) colonies after 14 days.

2.1.5. Gene expression profiling

We determined the transcriptional profiles of multipotent and lineage–committed progenitors using Affymetrix HG-U133 Plus 2.0 GeneChip arrays (3 samples for each population) (Affymetrix, Santa Clara, CA). We preprocessed and normalized the data in R ("Affy" package). We annotated the arrays using a custom chip definition file, the Gene Annot CDF for Human Gene U133 Plus 2.0 arrays, and the corresponding Bioconductor libraries [87]. We used DNA-Chip Analyzer (dChip) (www.dchip.org) software [88] to identify differentially expressed genes and perform hierarchical clustering.

2.2. ChIP-seq

2.2.1. ChIP assay

We prepared chromatin from EPP and MPP after cross-linking for 10' at RT with 1% formaldehyde- containing medium, using *tru*ChIPTM High Cell Chromatin Shearing Kit with SDS Shearing Buffer (Covaris). We sonicated nuclear extracts to obtain DNA fragments averaging 200 bp in length and immunoprecipitated the equivalent of 10^7 cells overnight with 10 µg of rabbit antibodies against H3K4me1 (ab8895, Abcam), H3K4me3 (ab8580, Abcam), and H3K27ac (ab4729, Abcam), as previously described [84, 89]. We used real-time SYBR Green PCR to validate genomic regions enriched in H3K4me1, H3K4me3 and H3K27ac.

2.2.2. ChIP-seq library preparation and sequencing

We prepared Illumina libraries, for EPP and MPP, starting from 10 ng of immunoprecipitated DNA (IP) and control DNA (INPUT: nuclear extracts sonicated but non-immunoprecipitated) following the Illumina ChIP-seq DNA sample preparation kit. We checked the libraries by capillary electrophoresis by Agilent Bioanalyzer 2100 with the High sensitivity DNA assay and quantified them with Quant-iTTM PicoGreen® dsDNA Kits (Invitrogen) by Nanodrop Fluorometer. We sequenced each library in one lane of a single strand 51 bp Illumina GAIIx run.

2.2.3. Bioinformatic data analysis

We mapped raw reads against the human reference genome (build hg19) using Bowtie [10] allowing up to 2 or 3 mismatches. We then processed each BAM file by using SAMtools [90], and converted each into a bed file using BEDTools [91]. We checked the quality of each sequenced sample using cross-correlation analysis implemented in spp R package (version 1.11) [12], shifting strands by a window of -50/+500 bp. All the quality control steps were performed using custom scripts in R programming language (http://cran.r-project.org/). For each sample the normalised strand coefficient (NSC) and the relative strand correlation (RSC) were calculated dividing the highest values in cross-correlation by the minimum value in cross-correlation in strand shift windows and by the cross-correlation value

relative to read-length strand shift [92]. Saturation analysis was performed downsampling each sample in 10 steps. At each step 5% of the total amount of reads was randomly removed, broad regions were called using spp R package [12] and compared to the regions called using the total set. Plots of histone modification signal at genomic features were performed using ngs.plot functions [93]. We performed ChIP-seq peak calling using SICER default parameters [14] and using each INPUT data to model the background noise. We downloaded HSPC raw H3K4me3, H3K4me1 and H3K27ac ChIP-seq data from the NIH Roadmap Epigenomics Mapping Consortium database (http://www.roadmapepigenomics.org/; GSM773041, GSM773043, GSM772894) and analysed them as described above for EPP and MPP.

2.2.4. Identification of *cis*-regulatory elements

We developed a custom R-workflow to identify promoters and enhancers. The pipeline analyses the histone modification islands generated by SICER and includes three steps. In the first step, the R script invokes BEDtools [91] to identify regions where H3K4me1 overlaps or does not overlap with H3K4me3. $H3K4me3^{+}/H3K4me1^{-}$ and $H3K4me3^{-}/H3K4me1^{+}$ regions are classified as putative promoters and enhancers, respectively. In the second step, the R script first normalizes the tag counts of H3K4me3 and H3K4me1 using the sequencing depths of both libraries and then calculates the log-ratios between H3K4me3 and H3K4me1 tag counts for $H3K4me3^{+}/H3K4me1^{+}$ regions. If the H3K4me3/H3K4me1 ratio is greater than 0, the region is defined as putative promoter, otherwise as putative enhancer. Regions identified with this method were then merged with $H3K4me3^{+}/H3K4me1^{-}$ and $H3K4me3^{-}/H3K4me1^{+}$ to have the complete sets of putative promoters and putative enhancers, respectively. Finally, we intersected putative regulatory elements with H3K27ac+ regions to identify active chromatin regions.

2.3. CAGE

2.3.1. Library preparation, sequencing and mapping

We extracted RNA from multipotent and lineage-committed progenitors using RNeasy Plus Mini kit (QIAGEN). DNAFORM Inc. at RIKEN Omics Science Center (Yokohama, Japan) performed DeepCAGE library preparation. Briefly, the cDNA synthesis was performed with 5 μg of total RNA, the random (N15) reverse-transcription primers and PrimeScript Reverse Transcriptase (TAKARA). Capped RNA was biotinylated and treated with RnaseOne. Then, hybrid cDNA with biotinylated Capped RNA was selected with cap-trapper method [94]. cDNA was released from streptavidin beads. A sample-specific linker, containing a recognition site for the sample-specific barcode sequence (3 bp) and the type III restriction-modification enzyme EcoP15I, was ligated to the single-strand cDNA. After ligation, the 2nd strand synthesis was performed and the resulting double-stranded cDNA was cleaved with EcoP15I. After, a second linker was ligated to the CAGE tag. The CAGE tags were separated from unmodified DNA with streptavidin beads. Then, the DNA fragments were PCR-amplified by using linker-specific primers. Each library was sequenced in one lane of a single strand 38 bp Illumina GAIIx run. 13 to 15 million reads were obtained per lane. DNAFORM at RIKEN GeNAS performed CAGE data analysis. Briefly, tags were extracted and mapped to human genome version hg19 (NCBI build 37), with a minimum match length of 21 bases and a maximum of one error; tags mapping the human ribosomal DNA sequence were eliminated. For CAGE tags mapping to multiple genome locations, a weighting strategy, based on the number of CAGE tags within a 200bp neighbourhood around each candidate mapping location, was applied. Equal weights were used if no unique tags were found within the 200 bp region for all candidate mapping locations [95]. The average mapping rate was 40%.

2.3.2. Promoter construction

We defined level-1 promoters ("transcription start sites") by summing the weighted number of CAGE tags at each genome position. Then we clustered level-1 promoters into level-2 promoters ("promoters") if they

were within 20 bp of each other on the same chromosomal strand. We calculated the expression level for each level-1 and level-2 promoter by dividing the number of CAGE tags of each promoter in each experimental condition by the total number of mapped CAGE tags in that condition, and multiplying by 10^6 (tags-per-million, TPM). The expression of each level-2 promoter of at least 10 TPM in at least one experimental condition was imposed.

2.3.3. Promoter annotation

We annotated transcription start sites using gene tracks from UCSC Genome Browser (https://genome.ucsc.edu). We annotated CAGE promoters on the base of their proximity to RefSeq genes, ENSEMBL ncRNA, ncRNA included in publicly available data sets [29, 96, 97], Vertebrate Genome Annotation (Vega) pseudogenes [98] and Yale Gerstein Group pseudogenes [99]. For each dataset, we found the annotation with the smallest distance to the CAGE-defined promoter on the same chromosome strand. We defined the distance as follows: 1. If the 3' end of the promoter was upstream of the 5' end of the annotation, then we used the distance between the 3' end of the promoter and the 5' end of the annotation. 2. If the 5' end of the promoter was downstream of the 5' end of the annotation, then we used the distance between the 5' of the annotation and the 5' end of the promoter. 3. Otherwise, the promoter overlapped the 5' end of the annotation. In this case, we considered the distance to be zero. If this distance was less than 400 bp, we associated this promoter to the gene or transcript.

2.3.4. Statistical analysis

We performed statistical differential analysis on level-2 promoters using edgeR [100] and raw count data (i.e. the number of cage tags within each promoter), manually setting dispersion value to 0.09.

3. RESULTS AND DISCUSSION

3.1. Purification and characterization of multipotent and lineage-restricted hematopoietic progenitors

Hematopoietic stem/progenitor cells (HSPCs) were obtained from cord blood of fully informed donors. HSPCs were enriched as $CD34^{+}$ $CD133^{+}$ populations by FACS sorting and pooled from different donors. Committed erythroid and myeloid progenitors were FACS-enriched as pools of $CD34^{low}/CD36^{high}$ and $CD34^{-}/CD13^{+}$ populations, respectively, after induction to differentiation in the appropriate conditions [86]. In order to characterize HSPCs and their erythroid and myeloid progeny, we investigated the presence of several surface markers by staining the cells with specific antibodies and performing flow cytometric analysis. HSPCs showed high levels of CD38, indicating that the majority of the cells was early hematopoietic progenitors [101]. The 95% of erythroid $CD34^{low}/CD36^{high}$ cells were CD71-positive and expressed low or undetectable levels of glycophorin A (GYPA), indicating that they are mainly composed by CFU-E and BFU-E progenitors [102], while MPP expressed the myeloid differentiation markers CD33 (99%) and CD11b (95%) [103] (*Figure 4*).

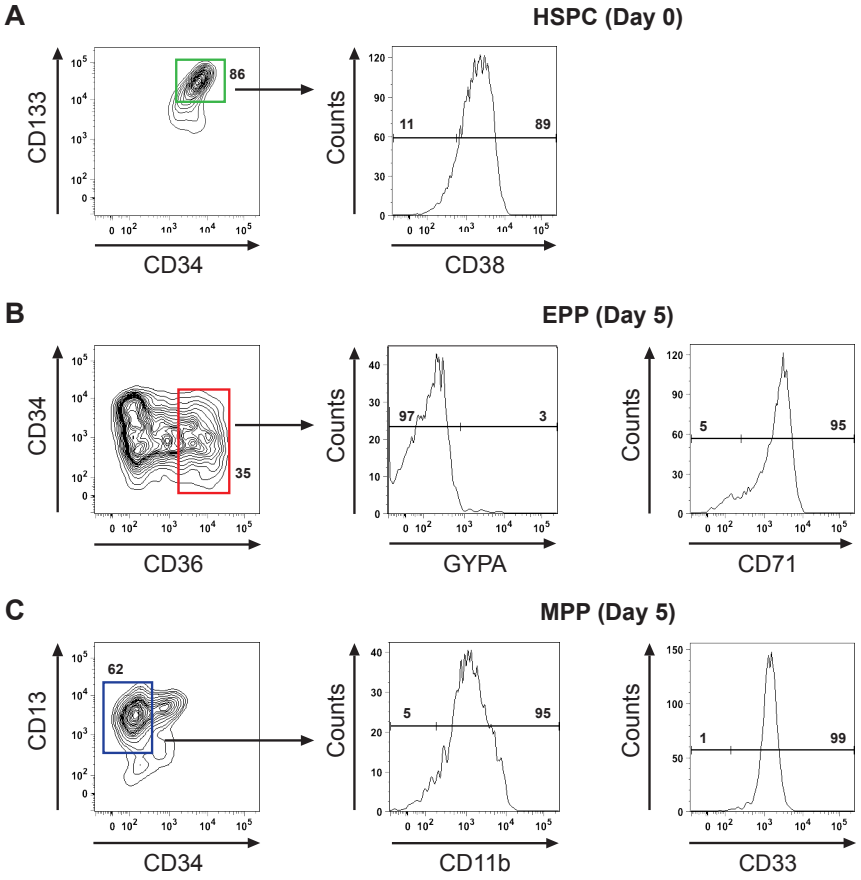


Figure 4. Purification and characterization of multipotent and lineage-restricted hematopoietic progenitors. (A-C) FACS analysis and CFC assay of HSPC, EPP and MPP.

To better characterize the composition of the cell populations, we performed a colony forming cell (CFC) assay, which assessed *in vitro* primitive human hematopoietic progenitors. A defined number of hematopoietic cells was seeded within a semi-solid (methylcellulose) medium that supports the growth of human progenitors and the development of erythroid and myeloid lineages, and resulted in the formation of discrete colonies of various morphologies. The colonies were classified into six different types: colony forming unit-erythroid (CFU-E), burst forming unit-erythroid (BFU-E), colony forming unit-granulocyte/macrophage (CFU-GM), colony forming unit-granulocyte (CFU-

G), colony forming unit-macrophage (CFU-M), colony forming unit-granulocyte/erythroid/macrophage/megakaryocyte (CFU-GEMM).
After 14 days, HSPCs gave rise to mixed cell colonies (CFU-GEMM), and both myeloid (CFU-GM, CFU-G, CFU-M) and erythroid (BFU-E and CFU-E) colonies, thus confirming their multilineage potential. In contrast, EPP and MPP populations generated >90% erythroid (BFU-E and CFU-E) and myeloid (CFU-GM, CFU-G, CFU-M) colonies respectively, confirming their lineage-restricted potential (*Figure 5*).

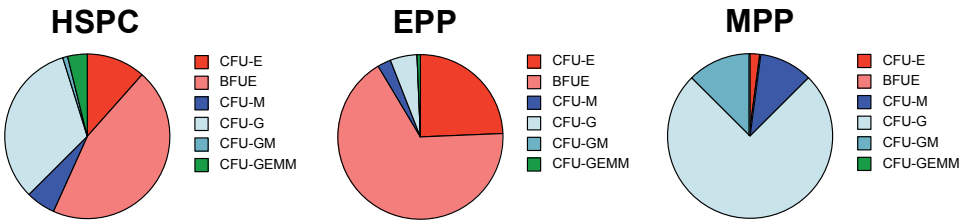


Figure 5. CFC assay. CFU-E, colony forming unit-erythroid; BFUE, burst forming unit-erythroid; CFU-M, colony forming unit-macrophage; CFU-G, colony forming unit-granulocyte; CFU-GM, colony forming unit-granulocyte/macrophage; CFU-GEMM, colony forming unit-granulocyte/erythroid/macrophage/megakaryocyte.

Global microarray gene expression profiling was used in order to identify genes enriched in each individual population and firmly establish their distinct identity. Microarray analysis was performed using mRNA isolated from $CD34^{+}/CD133^{+}$ HSPCs, erythroid $CD34^{low}/CD36^{high}$ and myeloid $CD34^{-}/CD13^{+}$ cells using Affymetrix Human Genome U133 Plus 2.0 array. Supervised analysis was performed using a fold-change equal to of greater than 2 and a p-value threshold of 0.05, to obtain differentially-expressed genes (DEGs). This analysis identified 415 DEGs between $CD34^{+}/CD133^{+}$ HSPCs and erythroid $CD34^{low}/CD36^{high}$ cells, of which 177 down-regulated and 238 up-regulated as a result of erythroid cells differentiation, and 126 DEGs between $CD34^{+}/CD133^{+}$ HSPCs and myeloid $CD34^{-}/CD13^{+}$ cells, of which 41 down-regulated and 85 up-regulated upon myeloid differentiation. Genes down-regulated in both conditions belong to TGFβ/BMP signaling pathways implicated in HSC self-renewal (e.g. *NOG*, *CHRDL1*, *TGFB1I1*), or to the tumour necrosis factor superfamily, involved in T- and B-cell

functions (e.g. *TNFSF13B*, *TNFSF4*, *LTB*), while genes involved in leukocyte activation and immune response (e.g. *MPO*, *CTSG*, *GZMA*) were down-regulated specifically upon erythroid commitment. Genes up-regulated in EPP are mainly involved in erythrocyte differentiation, maturation and homeostasis (e.g., *HBB*, *AHSP*, *ANK1*, *PKLR*, *CD36*, *GYPA*, *DNTM*), and include the master transcriptional regulators *GATA1* and *KLF1*. Instead, genes up-regulated in MPP are mainly involved in inflammatory response and immune defence function of neutrophils and macrophages (e.g., *ELANE*, *AZU1*, *CTSG*). For the complete list of DEGs, refer to appendix.

Taken together, these results confirmed the identity of the hematopoietic populations analysed in this study. $CD34^{+}/CD133^{+}$ cells well represent an enriched population of multilineage stem/progenitor hematopoietic cells, $CD34^{low}/CD36^{high}$ cells are representative of committed erythroid progenitors, while $CD34^{-}/CD13^{+}$ cells have a distinct myeloid phenotype.

3.2. Characterization of regulatory elements usage in HSPC and lineage-restricted progenitors

In order to obtain a genome-wide description of chromatin changes occurring upon HSPC commitment, we performed ChIP sequencing experiments in EPPs and MPPs, and used publicly available data sets for HSPCs. We designed ChIP-seq analyses on histone methylations typical of promoters and enhancers, H3K4me3 and H3K4me1 respectively, and on histone acetylation H3K27ac known to mark active regulatory elements. We then employed a bioinformatic framework for data analysis and integration in order to define and characterize regulatory elements in HSPC and lineage-restricted progenitors EPP and MPP.

3.2.1. Genome-wide histone modification profiling by ChIP sequencing

3.2.1.1. Quality assessment of ChIP-seq data

We designed ChIP-seq experiments on H3K4me1, H3K4me3 and H3K27ac in order to define regulatory elements in EPPs and MPPs, and reprocessed public datasets for HSPCs (GSE17312). For each cell type an Input sample was generated, in order to control for background. Chromatin was prepared and sonicated in order to obtain DNA fragments averaging 200 bp in length; each library was sequenced using Illumina platform, generating a total of more than 370 million reads of 51 bp in length. These reads represent the 5' of library fragments from either DNA strand. The positions of the raw reads were then determined mapping them against human reference genome (build hg19), allowing up to two mismatches in order to maximize the number of aligned reads and discarding ambiguous alignments (i.e. reads mapped to multiple genomic position). In this step the alignment parameters are kept stringent in order to avoid the mapping of spurious tags that can increase the noise of a ChIP-seq experiment. In fact a previous study [12] showed that reads at least 25 bp long significantly increase the cross-correlation profiles (expression of signal-to-noise ratio) of IP data when mapped with two mismatches. Using these criteria, we mapped a total of 344,985,809 reads, samples ranging from about 12 millions to 31 million reads (*Table 1*).

sample	mapped reads	NRF	FRiP	NSC	RSC
HSPC_H3K4me3*	31,131,313	0.78	0.82	2.08	1.07
HSPC_H3K4me1*	31,276,729	0.85	0.74	1.25	1.10
HSCP_H3K27ac*	31,602,356	0.40	0.40	1.46	1.49
HSCP_input*	39,436,811	0.97	-	1.01	0.64
EPP_H3K4me3	28,128,391	0.75	0.85	1.30	1.02
EPP_H3K4me1	26,185,846	0.94	0.32	1.27	1.19
EPP_H3K27ac	25,763,530	0.72	0.07	1.15	2.03
EPP_input	20,272,745	0.98	-	1.04	0.73
MPP_H3K4me3	28,732,471	0.43	0.81	3.02	1.05
MPP_H3K4me1	27,091,353	0.71	0.19	1.19	1.21
MPP_H3K27ac	37,720,127	0.30	0.12	1.81	2.90
MPP_input	23,370,210	0.54	-	1.04	1.38

Table 1. Sequencing data and QC statistics. NRF: Non-Redundant Fraction of mapped reads; FRiP: Fraction of Reads inside Peaks; NSC: normalized strand coefficient; RSC: relative strand correlation; *data from GSE17312 were reprocessed as EPP and MPP samples.

An automatic pipeline was set up for data quality control and to facilitate the reproducibility of the analyses.

We first checked if the depth of sequencing was high enough to robustly call the enriched regions in all samples. In order to do so, we performed a saturation analysis [92], progressively downsampling each sample and finding broad enriched regions using spp [12]. We found that H3K4me3 and H3K4me1 samples reached the saturation point at 90% of mapped reads, while H3K27ac samples were close to saturation point (*Figure 6*). Moreover a recent study demonstrated that these levels of mapped reads were acceptable for most applications [104]. We concluded that the amount of mapped reads of each sample is sufficient for exhaustively calling the enriched regions of histone modifications under study.

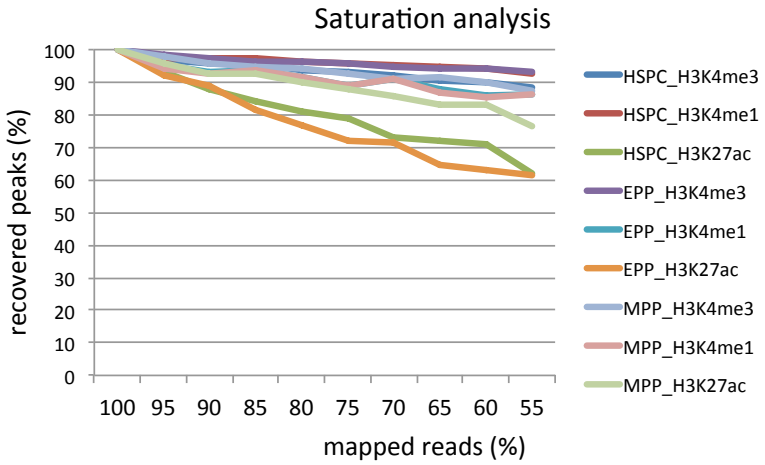


Figure 6. Check for sufficient sequencing depth.

In order to verify the good performance of our ChIP-seq experiments, we computed the fraction of non-redundant mapped reads for all the data sets and for IP samples we obtained values between 0.3 and 0.85, that are considered sufficiently high for histone modification ChIP-seq data, while, as expected, for input samples NRF values were close to 1, indicating that they are representative of the background signal along the whole genome (*Table 1*). The alignment step produced typical read distributions of a ChIP-seq experiment, with significant clustering of enriched DNA sequences at locations bound by the protein of interest, with mapped reads accumulated on forward and reverse strands roughly centred around the binding site. Because of this binding characteristic, we inspected the quality of the IP samples using cross-correlation analysis. The two DNA strands were shifted relative to each other by increasing distances in 500 steps, measuring the correlation between the enrichment at positive and negative strand at each step. A typical cross-correlation profile shows a maximum in correlation correspondent to the strand shift value of the library fragment and a second one that is a local maximum in cross correlation when the strands are shifted for the read length (read-length peak). We obtained cross-correlation profiles resembling typical situations of good ChIP-seq experiments (*Figure 7*).

Figure 7. Cross-correlation profiles. For each sample the cross-correlation at specific strand shift values is plotted. The magnitude of the peak reflects the fraction of tags in the data set that appears in accordance with the expected binding tag pattern.

A deeper analysis of cross-correlation profiles has been set up by ENCODE consortium, which proposed two metrics to estimate the signal-to-noise ratios in ChIP-seq experiments, measuring the ratio between fragment-length peak and the minimum value of cross correlation (normalized strand coefficient, NSC) and the ratio between the fragment-length peak and the read-length peak (relative strand correlation, RSC). For all the IP samples, we obtained values of NSC and RSC scores above minimum thresholds of 1.05 and 0.8, respectively (*Table 1*).

Good ChIP-seq experiments should show good enrichments relatively to annotated genomic elements, such as TSSs, TESs, and gene bodies. In order to verify enrichment at precise locations relative to genes, we plotted the normalized signals of assayed histone modification. For all the cell types we noted typical profiles of H3K4me3 enrichment at 5' of Ensembl genes, with peaks surrounding the TSS and a decrease of the signal upon it due to the physical occupancy of the transcriptional machinery, while H3K4me1 and H3K27ac signals remained lower, accordingly with the

nature of these histone modifications that do not typically mark promoters (*Figure 8*). These results suggest that H3K4me3 and H3K4me1 could be used to epigenetically discriminate transcription initiation at promoters from distal regions of open chromatin, typically associated to CREs, such as enhancers.

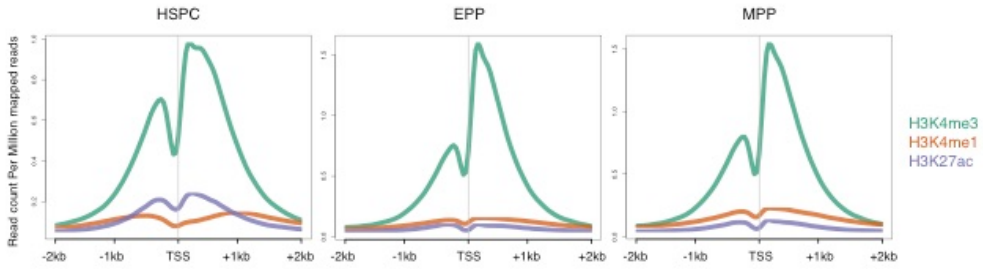


Figure 8. Histone mark enrichment at transcription start sites.

Finally we measured the global ChIP-enrichment, calculating the fraction of reads in enriched regions (FRiP) for each experiment. Although this analysis is typically performed for transcription factor ChIP-seq experiments, it can also help to evaluate the performance of IP experiments involving histone modifications. As expected, all the samples showed FRiP values above the minimum threshold of 0.01 [92] (*Table 1*).
In summary we performed a comprehensive evaluation of the quality of ChIP-seq experiments, considering depth of sequencing, library complexity, visual assessment of ChIP-seq profiles, global ChIP enrichment and cross-correlation analysis. Overall these results reflect the good quality of ChIP-seq data, thus representing a positive starting point for the downstream analyses.

3.2.1.1. Identification of histone modifications enriched regions

In order to identify the significantly enriched regions for each mark, we applied a peak calling procedure. Since H3K4me3, H3K4me1 and H3K27ac marks are broad, lack of well-defined peaks and span from several nucleosomes to very large domains, we decided to use SICER [14], a specific algorithm designed to deal with this kind of data. For each sample, we tuned the gap size (*g*), a SICER parameter that defines the

distance between two consecutive significant islands, to maximize their aggregate score.

Using this approach, we identified numbers of enriched regions ranging from 27,172 to 48,799 for H3K4me3, from 56,808 to 96,569 H3K4me1 and between 28,259 and 31,094 for H3K27ac mark (*Table 2*).

HMs	cell	Tot	median (bp)	Q_1 (bp)	Q_3 (bp)	G.C. (%)
H3K4me3	HSPC	48,799	1,600	1,000	2,600	3.2
	EPP	27,172	2,000	1,200	3,400	2.3
	MPP	30,346	1,600	1,000	2,600	2.0
H3K4me1	HSPC	96,569	1,600	1,000	2,800	7.8
	EPP	56,808	3,200	2,000	5,800	9.3
	MPP	86,741	2,400	1,600	4,000	9.7
H3K27ac	HSPC	31,030	3,800	2,400	6,000	5.2
	EPP	28,259	4,000	2,600	6,000	4.7
	MPP	31,094	2,800	1,800	4,000	3.4

Table 2. Statistics of histone modifications enriched regions. Tot: number of enriched regions identified; median, Q1 and Q3: median, first and third quantile of enriched regions length distributions; G.C.: genome coverage.

These results were consistent with the nature of the histone modifications involved in the study. H3K4me1 and H3K27ac showed larger enriched regions and higher genomic coverage, compared to H3K4me3, as expected from histone modification known to mark large intergenic regions instead of 5' end of genes.

3.2.2. Identification of regulatory elements

The identification of *cis*-regulatory elements is of striking importance for the study of molecular mechanisms at the basis of stem cell self–renewal, commitment and differentiation. In order to define putative promoters and putative enhancers in HSPC, EPP and MPP cells, we developed a custom bioinformatics workflow using ChIP-seq signals of H3K4me1 and H3K4me3 histone marks. Then we used H3K27ac histone mark to distinguish active ($H3K27ac^+$) promoters and active enhancers.

3.2.2.1. Definition of promoters and enhancers by ChIP-seq data integration

In recent years several bioinformatics tools have been designed to characterize genomic regulatory elements, using different approaches depending on the number and the type of histone modification involved. As general rule, high levels of H3K4me1 and low levels of H3K4me3 characterize strong-enhancers, but other histone marks, like H3K27ac, can be informative to detect enhancers that are active in one defined tissue and at a particular differentiation step. In our case the experimental model is composed by H3K4me1, H3K4me3 and H3K27ac, thus the identification of promoters and enhancers can be realized using computational methods that integrate the information of the enrichment of the first two marks [23] with the presence of the latter one to discriminate a subset of cell-specific active enhancers that may drive the HSPC commitment. We developed a custom R pipeline that takes as inputs the enriched regions identified using SICER from our ChIP-seq data sets and classifies the regions in three categories: regions characterized by the co-localization of both methylation marks ($H3K4me3^{+}/H3K4me1^{+}$), and regions having only one of the two methylation mark peaks ($H3K4me3^{+}/H3K4me1^{-}$ and $H3K4me3^{-}/H3K4me1^{+}$). In order to assign $H3K4me3^{+}/H3K4me1^{+}$ regions to one *cis*-regulatory element category, we proceeded with the following steps: i) identification of overlaps between H3K4me1 and H3K4me3 regions; ii) normalization of raw read counts by the sequencing depths for both marks; iii) calculation of log-ratio between the two signals; iv) labelling of promoters and enhancers if log-ratio is lower or greater than zero, respectively. Promoters classified in this way are then merged with $H3K4me3^{+}/H3K4me1^{-}$ regions and enhancers with $H3K4me3^{-}/H3K4me1^{+}$ in order to have complete sets of putative promoters and putative enhancers.

With these annotations, in total we identified more than 30,000 promoter regions in HSPC, EPP and MPP, with a similar average size, and we defined from 40,000 to 70,000 putative enhancers in both multipotent and committed progenitors with similar length distributions (*Table 3*).

cell	Tot. promoters	median (bp)	Q_1 (bp)	Q_3 (bp)
HPSC	17,810	3,000	1,600	4,800
EPP	12,045	5,000	2,800	8,400
MPP	13,185	4,000	2,400	6,200
cell	**Tot. enhancers**	**median (bp)**	**Q_1 (bp)**	**Q_3 (bp)**
HPSC	77,064	3,800	2,400	6,000
EPP	45,349	4,000	2,600	6,000
MPP	74,264	2,800	1,800	4,000

Table 3. Statistics of all promoters and enhancers identified by ChIP-seq data integration. Tot: number of regulatory elements identified; median, Q1 and Q3: median, first and third quantile of enriched regions length distributions.

In order to define active regulatory elements, we integrated promoters and enhancers with enriched regions of H3K27ac, a marker of active regions. We applied a very stringent method, defining as "active promoters" and "active enhancers" those putative regulatory elements that were covered with H3K27ac signal for at least 50% of their length. We found that more than 50% of putative promoters and about 20% of putative enhancers carried H3K27ac marker in HSPC, while lower percentages of *cis*-regulatory elements associated with activation were found for EPP and MPP, consistently with the hypothesis that large portions of HSPC genome are silenced during commitment (*Table 4*).

cell	Tot. promoters	median (bp)	Q_1 (bp)	Q_3 (bp)
HPSC	9,671 (54%*)	3,200	2,000	4,800
EPP	4,996 (41%*)	3,400	2,200	5,200
MPP	3,672 (27%*)	2,400	1,600	3,600
cell	**Tot. enhancers**	**median (bp)**	**Q_1 (bp)**	**Q_3 (bp)**
HPSC	15,338 (19%*)	2,400	1,400	4,200
EPP	7,221 (15%*)	2,800	1,800	5,000
MPP	3,336 (4%*)	1,800	1,200	2,800

Table 4. Statistics of H3K27ac$^+$ promoters and enhancers identified by ChIP-seq data integration. Tot: number of active regulatory elements identified; median, Q1 and Q3: median, first and third quantile of enriched regions length distributions; *percentage of the total number of regulatory elements reported in **Table 3**.

3.2.2.2. Comparative analysis of active cis-regulatory elements in HSPC, EPP and MPP

We evaluated the dynamics of promoters and enhancers upon HSPC commitment in order to identify changes in *cis*-regulatory elements usage that are responsible of early lineage-restricted progenitor definition. We found that the great majority of EPP and MPP (92% and 93%, respectively) active promoter regions are shared with HSPC. Conversely a much lower proportion of active enhancers are shared upon lineage commitment while the majority are cell-specific (*Figure 9*), suggesting that enhancers play a major role in HSPC commitment.

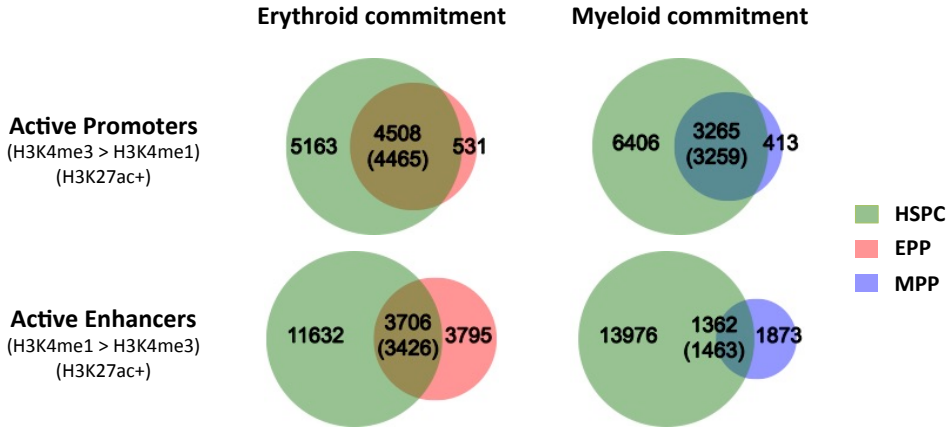


Figure 9. Dynamics of promoter and enhancer chromatin signatures upon HSPC commitment. Venn diagrams show the overlap of strong (H3K27ac+) promoters (H3K4me3>H3K4me1) and enhancers (H3K4me1>H3K4me3) identified in HSPC, EPP and MPP. Values in intersections refers to HSPC, those in brackets to EPP and MPP.

Figure 10 shows an example of active enhancers identified at LCR of β-globin locus only in HSPC and EPP (*Figure 10*). In HSPC enhancers are already active, prior to erythroid commitment. These regulatory elements are then switched off in myeloid commitment, as can be seen evaluating H3K4me1 and H3K27ac signals at LCR.

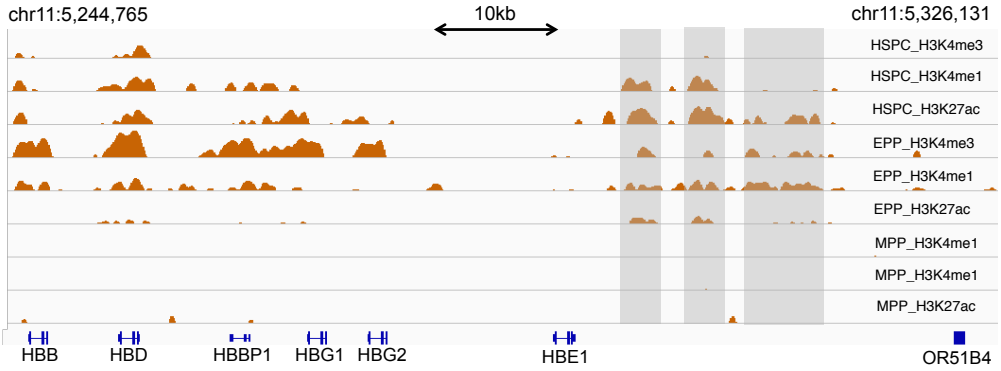


Figure 10. Example of identified enhancer regions at human beta globin locus. The shaded regions showed the enhancers identified in HSPC and EPP at locus control region. MPP do not show enrichments in active enhancer-specific histone marks.

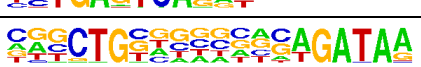
We next explored whether enhancer maps might reveal trans-regulatory factors involved in the HSPC commitment gene expression regulation. Motif analysis performed with HOMER [17] showed that ChIP-defined cell-specific active enhancers are enriched in binding sites for general and hematopoietic TFs (*Table 5*). Among the most represented motifs spanning all the three cell types, there are members of ETS TF family, that are the core of the gene regulatory networks controlling the key aspects of hematopoietic specification during embryogenesis as well as the maintenance and differentiation of HSCs [105]. Along with these factors, members of AP-1 complex (e.g. Jun, Fos, Fra1) are also represented HSPC, EPP and MPP, consistent with their roles in functional development of hematopoietic precursor cells into mature blood cells along most of the hematopoietic cell lineages [106]. Similarly, BATF, Bach1 and Bach2 motifs are enriched in active enhancers of all three cell types. BATF1 has been shown to limit the self-renewal of HSCs thus promoting the commitment and differentiation [107], while Bach1 and Bach2 have been reported to suppress the myeloid lineage and promote the lymphoid lineage [108], suggesting that at the stages of HSPCs, EPP and MPP they may act with different mechanisms. Interestingly, HSPC active enhancers were exclusively enriched in RUNX1 and IRF2 motifs, which are TFs involved in regulating the development and maintenance of HSCs [109] and in

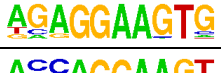
protecting the quiescent HSCs from differentiation [110], respectively. Moreover EPP and MPP active enhancers were specifically enriched in GATA factors and PU.1 motifs, respectively, indicating that their regulatory activities of megakaryocytic-erythroid and granulocytic-monocytic cell fates are already present at this stage of commitment. MPP were also specifically enriched in the motif of SpiB TF, which is required in myeloid specification [111].

a) HSPC-active enhancers				
Motif	Name	P-value	Target	BG
	ETS1(ETS)	1e-242	71.61%	56.71%
	Fli1(ETS)	1e-233	70.85%	56.19%
	ETV1(ETS)	1e-219	80.63%	67.53%
	PU.1(ETS)	1e-210	47.11%	33.25%
	GABPA(ETS)	1e-205	63.11%	49.05%
	ERG(ETS)	1e-201	87.07%	75.85%
	ELF5(ETS)	1e-155	60.51%	48.24%
	EHF(ETS)	1e-139	81.27%	71.15%
	SpiB(ETS)	1e-111	27.37%	18.84%
	Elk1(ETS)	1e-94	35.00%	26.32%
	Elk4(ETS)	1e-93	34.46%	25.89%
	SPDEF(ETS)	1e-81	66.35%	57.71%
	ELF1(ETS)	1e-79	33.70%	25.83%
	RUNX1(Runt)	1e-59	69.43%	62.21%
	NF-E2(bZIP)	1e-46	6.57%	3.78%
	RUNX2(Runt)	1e-43	61.48%	55.13%
	Bach2(bZIP)	1e-43	16.95%	12.51%
	Bach1(bZIP)	1e-42	6.30%	3.68%
	Nrf2(bZIP)	1e-41	5.63%	3.19%
	Fosl2(bZIP)	1e-40	26.95%	21.73%

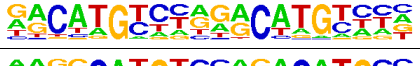
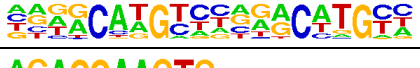
	Jun-AP1(bZIP)	1e-39	20.36%	15.72%
	IRF1(IRF)	1e-38	17.67%	13.38%
	Gata2(Zf)	1e-36	53.72%	47.82%
	MyoD(bHLH)	1e-34	45.82%	40.22%
	BMYB(HTH)	1e-33	79.23%	74.42%
	Gata1(Zf)	1e-32	48.81%	43.30%
	AMYB(HTH)	1e-31	79.11%	74.44%
	Fra1(bZIP)	1e-31	43.71%	38.40%
	Atf3(bZIP)	1e-31	49.59%	44.21%
	IRF2(IRF)	1e-29	12.77%	9.53%
	MYB(HTH)	1e-29	83.08%	78.94%
	BATF(bZIP)	1e-27	48.68%	43.66%
	ISRE(IRF)	1e-26	8.62%	6.09%
	Bcl6(Zf)	1e-26	78.42%	74.13%

b) EPP-active enhancers

Motif	Name	P-value	Target	BG
	Gata1(Zf)	1e-84	59.59%	46.28%
	Gata2(Zf)	1e-78	64.05%	51.33%
	Gata4(Zf0	1e-60	79.89%	70.00%
	Fosl2(bZIP)	1e-46	29.35%	21.01%
	Jun(bZIP)	1e-45	22.71%	15.30%
	GATA:SCL	1e-45	15.26%	9.17%
	Bach1(bZIP)	1e-40	7.51%	3.62%
	GATA3(Zf)	1e-36	91.97%	86.42%
	NF-E2(bZIP)	1e-34	7.21%	3.65%
	GABPA(ETS)	1e-33	53.85%	45.52%
	Fra1(bZIP)	1e-30	47.10%	39.39%
	BATF(bZIP)	1e-29	52.66%	44.92%

	Bach2(bZIP)	1e-28	17.23%	11.98%
	Nrf2(bZIP)	1e-27	6.08%	3.13%
	Elk1(ETS)	1e-27	29.83%	23.38%
	ETV1(ETS)	1e-26	71.38%	64.47%
	Elk4(ETS)	1e-25	29.16%	22.98%
	ELF5(ETS)	1e-25	54.18%	47.05%
	MYB(HTH)	1e-24	85.15%	79.66%
	ETS1(ETS)	1e-24	60.91%	54.00%
	Atf3(bZIP)	1e-23	52.34%	45.42%
	ERG(ETS)	1e-23	79.31%	73.38%
	ELF1(ETS)	1e-22	28.55%	22.74%
	BMYB(HTH)	1e-22	81.49%	75.88%
	AMYB(HTH)	1e-22	80.72%	75.09%
	PU.1(ETS)	1e-21	38.20%	31.91%
	EHF(ETS)	1e-20	75.46%	69.62%
	Fli1(ETS)	1e-20	59.57%	53.26%

c) MPP-active enhancers

Motif	Name	P-value	Targets	BG
	Fosl2(bZIP)	1e-99	35.76%	21.31%
	Fra1(bZIP)	1e-96	53.01%	36.97%
	Jun-AP1(bZIP)	1e-86	27.70%	15.57%
	Atf3(bZIP)	1e-85	58.03%	42.80%
	BATF(bZIP)	1e-81	56.96%	42.04%
	AP-1(bZIP)	1e-72	60.70%	46.68%
	Bach2(bZIP)	1e-50	20.65%	12.30%
	p53(p53)	1e-48	12.95%	6.59%
	p63(p53)	1e-43	35.25%	25.41%
	PU.1(ETS)	1e-34	42.55%	33.30%

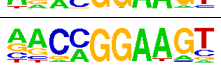
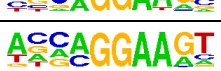
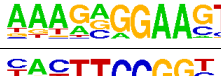
	GABPA(ETS)	1e-32	59.24%	49.93%
	ETS1(ETS)	1e-31	66.21%	57.19%
	Fli1(ETS)	1e-31	66.62%	57.63%
	ELF5(ETS)	1e-26	56.19%	47.77%
	ETV1(ETS)	1e-25	75.84%	68.44%
	EHF(ETS)	1e-24	77.24%	70.01%
	Bach1(bZIP)	1e-24	6.77%	3.47%
	SpiB(ETS)	1e-23	25.04%	18.67%
	Elk4(ETS)	1e-22	35.54%	28.38%
	NF-E2(bZIP)	1e-21	6.79%	3.62%
	ERG(ETS)	1e-21	82.17%	76.03%
	ELF1(ETS)	1e-20	34.62%	27.92%
	Elk1(ETS)	1e-20	35.54%	28.81%

Table 5. Analysis of TFBS in epigenetically defined enhancers. Putative TFBS in cell-specific promoters and enhancers were identified using HOMER. Targets, percentage of query sequences that are motif enriched; BG, percentage of background sequences that are motif enriched.

3.3. Transcriptomic analysis of HSPC during lineage commitment

In order to define the promoter usage in HSPC and their committed progeny, we used CAGE *sequencing*, a technique that identifies active TSSs at single-base-pair resolution and measures the expression level of each transcript, and applied a bioinformatics framework to identify clusters of transcription initiation events and classify them on the basis of annotated features and expression levels.

3.3.1. Definition of whole genome high resolution map of transcription initiation events by CAGE sequencing

3.3.1.1 CAGE sequencing mapping statistics and single nucleotide resolution annotation

We extracted total RNA from each cell population and generated CAGE tag libraries from the 5' ends of capped RNA PolII transcripts that were sequenced in separate Illumina lanes, generating a total of 33 millions reads of 29 bp in length that were then mapped to the human genome (hg19) using the rescue strategy of multimapped tag to improve the detection of TSSs [95]. This step resulted in an average mapping rate of 40% producing a total of 13,439,976, of which 4,328,146, 4,695,128, and 4,416,702 for HSPCs, EPPs, and MPPs respectively.

We mapped more than 70% of the TSSs to promoters and 5' UTRs of coding and non-coding genes in all cell types. Interestingly more than 20% of TSSs mapped to intergenic regions, introns, exons and 3'UTRs, supporting the hypothesis of the existence of alternative or novel promoters in hematopoietic progenitors. About 2% of TSSs mapped to the antisense strand of known genes, mostly in promoters and introns of coding transcripts, suggesting the presence of regulatory mechanisms involving antisense transcription (*Figure 11*).

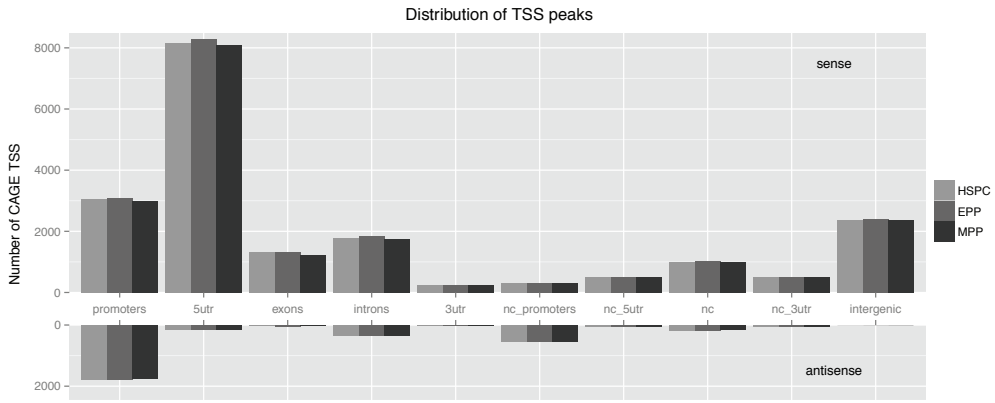


Figure 11. Genomic distribution of CAGE TSSs in HSPC, EPP and MPP. TSSs were mapped to regions annotated as promoters, 5' UTR, exon, intron and 3' UTR of coding and noncoding genes (in sense or antisense orientation) or as intergenic regions.

3.3.1.2. Promoter identification and annotation from CAGE-seq data

In order to define transcriptional initiation regions, we clustered CAGE TSSs in promoters if they are within 20 bp of each other on the same chromosomal strand and have similar expression levels. Thus promoters are defined as local clusters of nearby and co-expressed TSSs. We identified similar numbers of promoters with comparable length distributions in the three cell types (*Table 6*).

cell	Tot	median	Q_1 (bp)	Q_3 (bp)
HSPC	13.582	143	100	200
EPP	14.041	143	99	201
MPP	13.609	144	95	197

Table 6. Statistics of promoters identified by CAGE TSSs clustering. Tot: number of promoters identified by CAGE; median, Q1 and Q3: median, first and third quantile of enriched regions length distributions.

We then assigned CAGE promoters to the closest coding or non-coding transcript (including lincRNAs, miRNAs, rRNAss and snRNAs) using publicly available data sets. The majority (~80%) of them were annotated to known genes or transcripts in HSPCs (11,074 out of 13,852), EPPs (11,227 out of 14,041) and MPPs (10,934 out of 13,609), and particularly to protein-coding transcripts (69% in HSPCs and 70% in MPPs and EPPs). Interestingly, more than 2,600 promoters (~20%) were not associated with any known gene or transcript in all three cell types, and may drive transcription of yet unknown coding and non-coding transcripts (*Figure 12*).

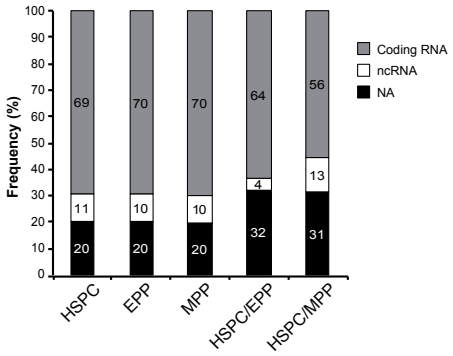


Figure 12. Annotation of total and differentially used CAGE promoters. The graphs show the proportions of total and differentially used CAGE promoters associated to coding RNA and ncRNA (miRNA, rRNA, snoRNA and snRNAand lincRNA).

In order to identify the possible involvement of transposable elements in gene regulation, we assessed if CAGE promoters overlapped with annotated repeat masker categories from UCSC Genome Browser. We found that about 6% of all CAGE promoters overlapped with repetitive elements such as long interspersed elements (LINE), short interspersed elements (SINE) and long terminal repeat elements (LTR) (*Figure 13*).

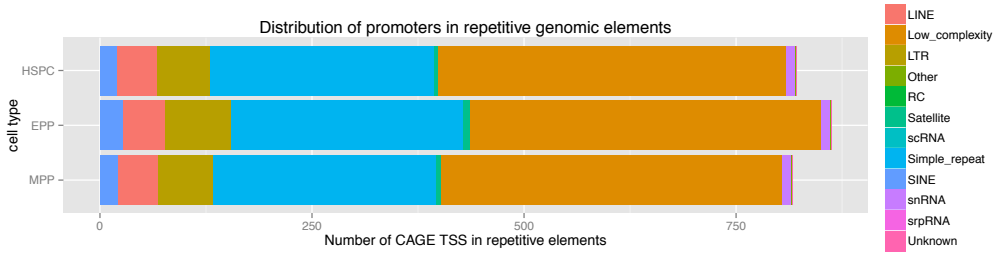


Figure 13. Genomic distribution of CAGE promoters in repetitive elements. CAGE promoters of HSPC, EPP and MPP were annotated with RepeatMasker regions.

These results support the hypothesis that many of genes involved in the specification of HSPC cell fate may be regulated by transposable elements.

3.3.2. Quantitative transcriptomic analysis of HSPC and lineage-restricted progenitors

3.3.2.1. Differential expression analysis of CAGE promoters

In order to determine the differential promoter usage during erythroid and myeloid commitment, we applied a statistical method (edgeR) using read counts of CAGE promoters, adjusting the resulting p-values with Benjamini-Hochberg correction. This analysis identified 725 differentially used promoters between HSPC and EPP, (265 down-regulated and 459 up-regulated) and 1050 between HSPC and MPP (599 down-regulated and 450 up-regulated). We looked at promoters that were specifically expressed in HSPC, and we found that only 77 promoters were expressed exclusively in HSPC and down-regulated in both EPP and MPP ("HSPC-specific" promoters), while the remaining down-regulated promoters remained expressed in one of the two lineages (188 in EPP and 522 in

MPP). Of the up-regulated genes, 421 out of 459 were specific for EPP, and 412 out of 450 were up-regulated in MPP only.

We then analysed the pathways involved in differentially expressed genes using DAVID [112]. Genes transcribed by HSPC-specific promoters were significantly enriched in functional categories related to multicellular organismal development, system development and immune response (modified Fisher's Test EASE score ≤0.05 after Benjamini correction for multiple testing). Genes associated to EPP-specific promoters were enriched in hemoglobin pathway and erythrocyte development categories, while those associated to MPP-specific promoters were mainly involved in leukocyte biology and immune response. Conversely, promoters down-regulated in EPP were associated with genes involved in immune response and leukocyte biology, while those down-regulated in MPP were associated to translation, structural constituent of ribosome, and macromolecular complex organization (*Figure 14*).

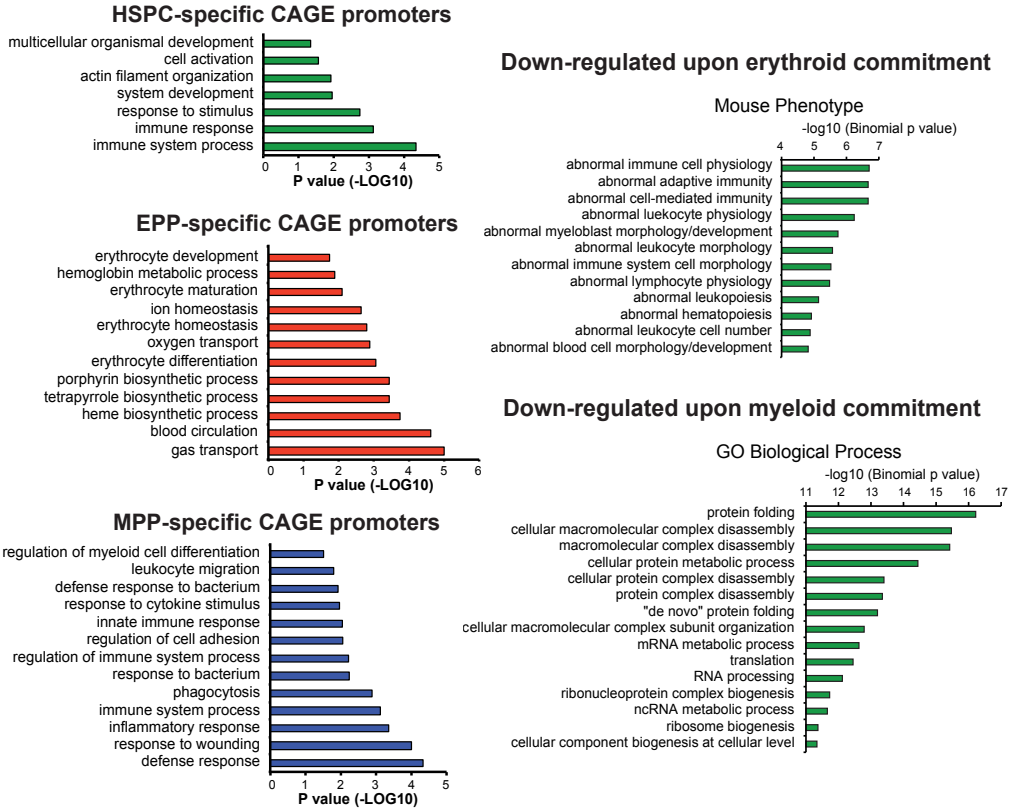


Figure 14. Functional annotation of genes associated to cell-specific promoters. Gene ontology analysis (DAVID 6.7) of genes associated with HSPC-, EPP- and MPP-specific CAGE promoters.

We also found that a significant number of genes associated with CAGE promoters were functionally linked in cell-specific molecular networks, e.g., the DNA methyltransferase DNMT3A and CD34 pathways for HSPC-specific and down-regulated promoters, and the GATA1 and CSF3R (G-CSF receptor) pathways for EPP- and MPP-specific promoters, respectively.

We next looked specifically at the set of differentially used promoters driving the expression of transcription factors (TFs), co-factors and chromatin remodelers. A few factors were over-expressed in HSPCs, such as the HSC regulators MYCN and DNMT3A, HOXA7, an essential TF in hematopoietic progenitors, SLA2, implicated in lymphocyte biology and NAP1L3 (nucleosome assembly protein 1-like 3), whose function in

hematopoiesis is yet unknown. EPP- and MPP-specific promoters drove the expression of known erythroid and myeloid transcriptional regulators, such as TAL1, GATA1, KLF1, NFE2, ZFPM1, LDB1, GFI1B, STAT5, MXI1 and KLF3 in EPP, and NFIA, MNDA, LRRFIP1, ZBTB20, KLF4, NCOA4, STAT6, MEF2A, SREBF1, JARID2, MLL3 and MLLT10 in MPP. Along with these results we found that more than 50% of EPP- and MPP-specific TFs and co-factors were not previously associated to erythropoiesis or myelopoiesis, such as CREB3L3, EGR1, FBXO7, FHL2, HES6, SEC14L2, TEAD1, TSC22D1, TSC22D3, ZBTB16 in erythroid progenitors and ARID4A, ASH1L, BAZ2B, C21orf66, CNOT6, CREB5, DENND4A, FOS, GTF2A1, ID3, MXD4, NFIL3, NR2C1, REL, RREB1, SFRS14, SHPRH, TBC1D22A, ZBTB7B, ZNF587 and ZNF70 in myeloid progenitors.

In order to assess if the different transcriptional programs could be possibly due to alternative promoter usage of genes we assessed if different CAGE promoters were assigned to the same gene (e.g. to different transcripts of the same gene), with different levels of expression between cell types. We observed lineage-specific alternative promoter usage only for six genes. As an example, the *LMO2* gene, coding for a developmentally regulated transcription factor with a crucial role in haematopoietic development, is transcribed from three different promoters, of which Promoter 3 was active only in HSPCs, Promoter 1 mainly in HSPCs and MPPs, and Promoter 2 essentially used by EPPs (*Figure 15*).

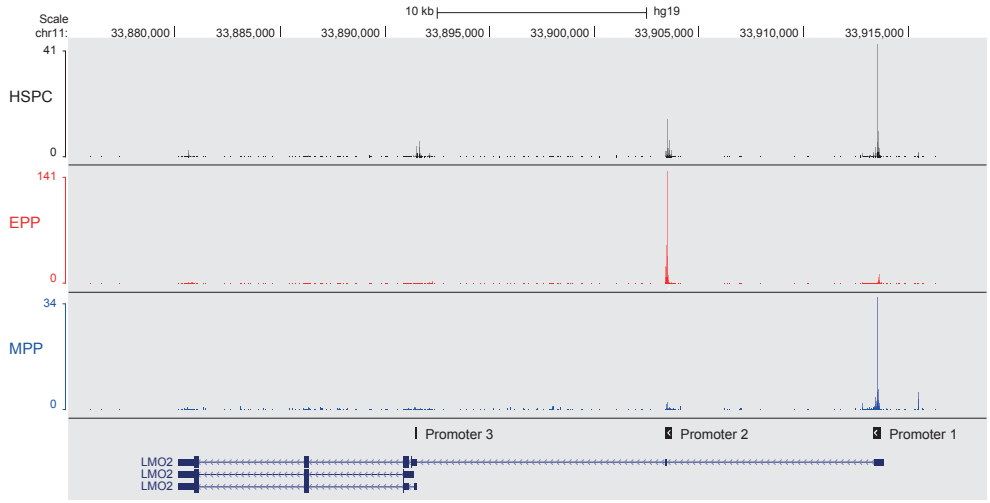


Figure 15. Example of alternative promoter usage. Three promoters of LMO2 gene were differentially used in HSPC, EPP and MPP (Promoter 1, 2 and 3). The different promoters and their expression level (i.e. CAGE tag raw counts of each TSS) are shown.

To better understand the regulatory circuitry operating on lineage-specific CAGE promoters, we analysed putative transcription factor binding sites (TFBS) within the proximal regions (-300 to +100 bp from TSSs) of differentially used promoters by HOMER tool [17]. HSPC-specific promoters were enriched for binding motifs of the ETS family of TFs, which regulate development and maintenance of HSCs and their differentiation along multiple lineages [105]. EPP- and MPP-specific promoters were instead enriched for motifs of ubiquitous promoter-associated TFs, like SP1, TBP, MAZ, NFY, NRF1, and lineage-specific TFs, such as GATA1, GATA2, TAL1, KLF1 and KLF4 for erythroid progenitors, and GABPA, FLI1, ETS1, ELK1, ELF1 and PU.1 for myeloid progenitors (*Figure 16*).

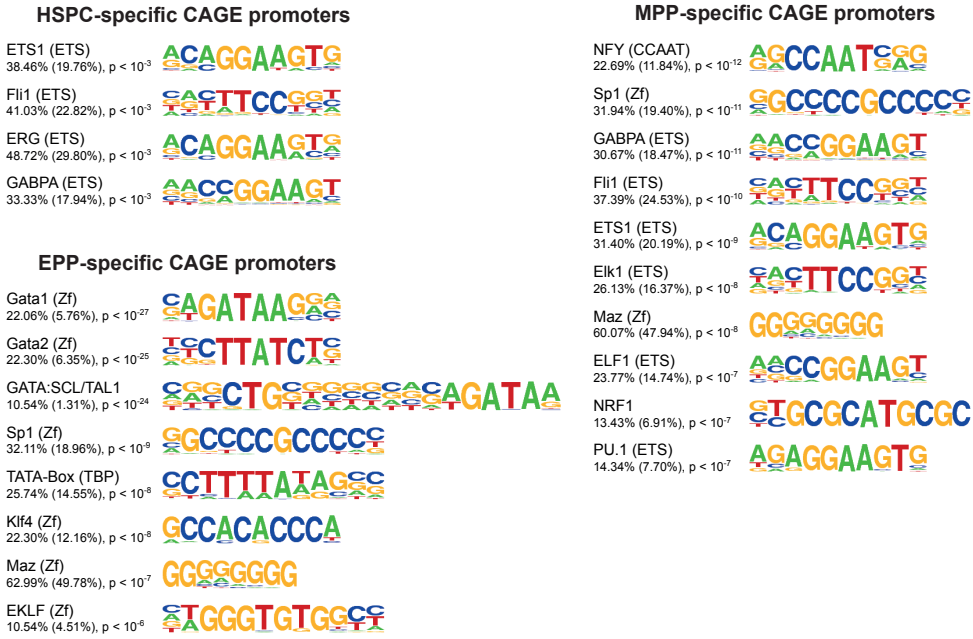


Figure 16. Analysis of putative TFBS within CAGE promoters. Transcription factor motif finding in cell-specific promoters was performed using HOMER software.

Next, we analysed CAGE promoters not assigned to any known genes or transcripts. Around 22% of these unannotated promoters harboured an epigenetic enhancer signature, a value that increased up to 45% for DU promoters (*Figure 17*) suggesting that they may represent enhancer-derived RNA acting in *cis* on adjacent target genes [113, 114].

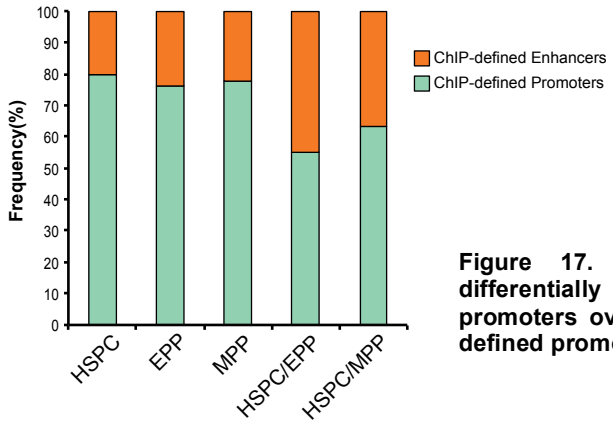


Figure 17. Distribution of total and differentially used unannotated CAGE promoters overlapping with epigenetically defined promoters and enhancers.

To analyse the influence of H3K27ac+ cell-specific enhancers on nearby genes, we analysed the expression level of the closest CAGE promoters taking the nearest neighbour gene to each enhancer. We observed a significant increase in gene expression compared to the whole set of promoters for all cell types (*Figure 18*), confirming that identified active

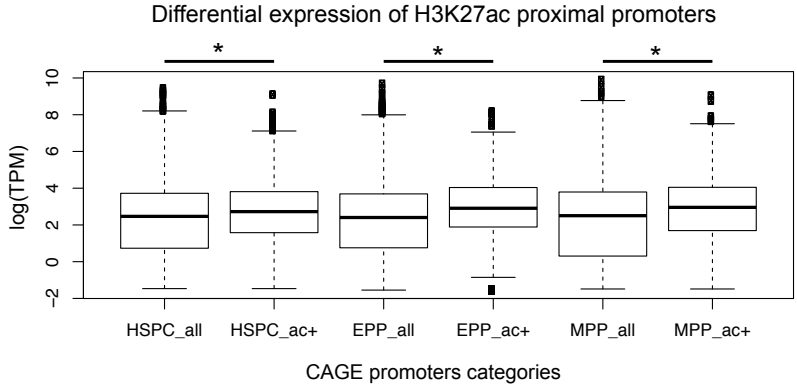


Figure 18. Effect of H3K27ac on nearest neighbor genes. TPM: tags per million mapped reads; *P<1e-10, Kolmogorov-Smirnov test.

enhancers are of striking importance in regulating transcriptional programs of HSPC in erythroid and myeloid commitment.

3.4. Chromatin dynamics at *cis*-regulatory elements

We then examined the chromatin dynamics of HSPCs early commitment, by analysing histone marks enrichments at identified regulatory elements, such as developmental specific enhancers and differentially expressed promoters. Differentially expressed promoters showed between two condition showed non-specific enrichment in H3K4me3 and variable enrichment in H3K27ac that is dependent on the number of promoters being up-regulated in the examined conditions (*Figure 19*).

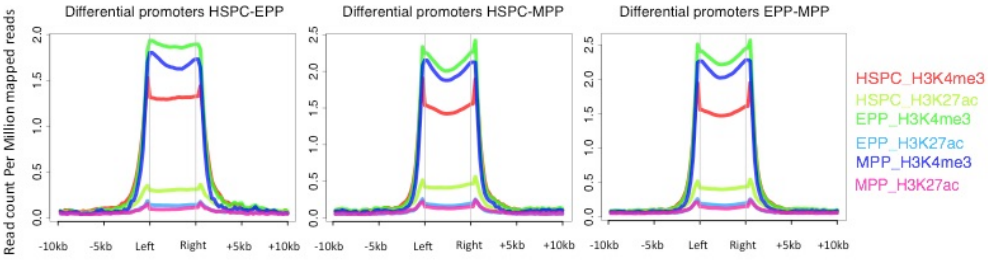


Figure 19. Histone modification average enrichments at differentially expressed promoters.

Conversely cell-specific active enhancers showed differential enrichment in histone marks during commitment. HSPC active enhancers showed high enrichment in H3K4me1 and H3K27ac at lower degree, while signal relative to EPP and MPP enhancers is absent. In erythroid and myeloid progenitors the situation was slightly different, showing specificity for H3K4me1 in EPP, while for MPP–specific active enhancers we noticed an enrichment of the same mark also in HSPC and EPP, but the presence of the H3K27ac mark was specific for myeloid commitment (*Figure 20*).

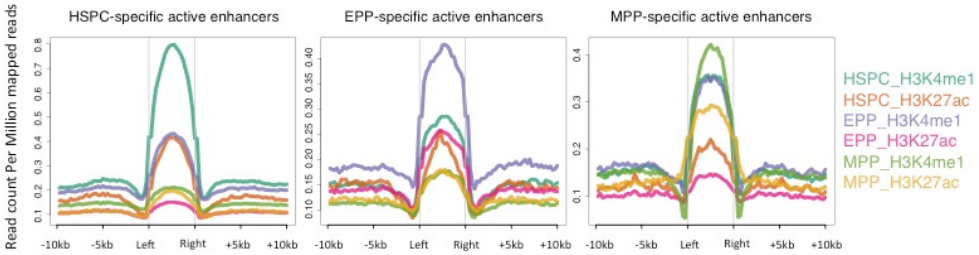


Figure 20. Histone modification average profile at lineage-specific active enhancers.

These results showed that promoters were ubiquitously occupied by the same histone marks in different steps of HSPC early commitment, while enhancers are associated with active chromatin marks in a cell-type specific manner, thus acting a main role in erythroid and myeloid commitment.

4. CONCLUSIONS AND PERSPECTIVES

As the last part of the dissertation, this section summarizes the results obtained in this study, suggesting future perspectives.

Human hematopoietic stem cell-based therapy represents a real and potentially long−term cure for patients affected by hematological and oncological diseases. A better knowledge of HSC self-renewal, commitment and differentiation mechanisms is crucial to understand HSC biology and to develop HSC-based therapies.

New high-throughput approaches based on the next-generation sequencing (NGS) technologies are essential to study the transcriptome, the epigenome and the usage of regulatory elements in the genome. In this study we used different high-throughput genome analysis tools, such as CAGE and ChIP-Seq, to define the transcriptional and epigenetic profile of cord blood-derived human hematopoietic stem/progenitor cells (HSPCs) and their committed erythroid and myeloid progeny, to characterize genes and regulatory regions fundamental in HPSC self-renewal, commitment and differentiation.

We comprehensively analysed the transcriptome and the epigenome of $CD34^{low}/CD36^{high}$ EPP and $CD34^{low}/CD13^{+}$ MPP cell populations, that represent earlier stages of erythroid and myeloid differentiation compared to the $CD34^{-}/CD36^{+}/GYPA^{+}$ erythroid precursors and the $CD14^{+}$ mature monocyte-macrophage elements analysed in previous studies [84, 115-118].

We used CAGE-seq analysis to define more than 13,000 transcriptionally active promoters in HSPC, EPP and MPP, the majority of which harboured an active (acetylated) epigenetic promoter signature. Our results showed that the three cell types shared most of the promoters and transcripts, suggesting that transcriptional states are largely maintained in early hematopoietic differentiation and progenitor identity during commitment is determined by a relatively small number of differentially used promoters. Moreover we showed that the differentially regulated fraction of promoters is significantly enriched in binding sites for transcription factors essential for hematopoietic development. CAGE and epigenetically defined strong

promoters overlapped in most cases, providing a robust signature for defining promoter usage. Interestingly, we identified only 77 promoters expressed exclusively in HSPCs and down-regulated in both EPPs and MPPs, while the remaining down-regulated promoters remained expressed in one of the two lineages. On the contrary, >95% of the 459 promoters up-regulated in EPPs and 450 up-regulated in MPPs were lineage-specific. These data indicate the existence of very few strictly HSPC-specific promoters and factors maintaining multilineage potential, and that lineage commitment is essentially exerted by up-regulation of a few hundred promoters, including those driving the expression of lineage-specific master TFs, such as TAL1, GATA1, KLF1, NFE2 and STAT5 in EPP, and NFIA, KLF4 and STAT6 in MPP.

Analysis of histone modification signatures by ChIP-seq allowed the identification of more than 45,000 putative enhancers in each cell population, indicating that most promoters likely interact with multiple enhancers. Differently from promoters, enhancers consistently changed upon erythroid and myeloid commitment: about a half of the active, acetylated enhancers mapped in EPPs and MPPs were not shared with HSPC, while 75 and 90% of the active enhancers mapped in HSPCs disappeared in erythroid and myeloid commitment respectively. These data indicate that enhancers are dramatically redefined during lineage commitment, and that differential enhancer usage is responsible for the differential regulation of promoter activity underlying lineage restriction. Thus activation of the set of lineage-specific enhancers is most likely responsible for both activation of lineage-specific promoters and fine tuning of the non-specific ones.

In conclusion, our data indicate that hematopoietic commitment and differentiation involve small changes of the transcriptional profile and "classical" promoter usage (CAGE and ChIP-Seq H3K4me3 data), and are mainly regulated by enhancers, which are differentially used in the specific lineages (as observed from CAGE and ChIP-Seq H3K4me1 data).

Overall, we provided an overview of the differential transcriptional programs of HSPCs and committed myeloid and erythroid hematopoietic precursors.

Moreover, this study represents a unique source of genes and regulatory regions involved in HPSC self-renewal, commitment and differentiation.
In the next future we will profile the action of lineage-specific transcription factors that were enriched in active enhancers and showed significant up-regulation in erythroid or myeloid progenitors. Furthermore, chromatin conformation analyses, such as 5C and Hi-C, will be required to link active enhancers with their interacting promoters, in order to unveil the fine regulatory circuitry at the basis of HSPC erythroid and myeloid commitment.

5. REFERENCES

1. Felsenfeld G, Groudine M (2003) Controlling the double helix. Nature 421:448–453. doi: 10.1038/nature01411

2. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705. doi: 10.1016/j.cell.2007.02.005

3. Izzo A, Schneider R (2010) Chatting histone modifications in mammals. Briefings in Functional Genomics 9:429–443. doi: 10.1093/bfgp/elq024

4. Dillon SC, Zhang X, Trievel RC, Cheng X (2005) The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol 6:227. doi: 10.1186/gb-2005-6-8-227

5. Hublitz P, Albert M, Peters AHFM (2009) Mechanisms of transcriptional repression by histone lysine methylation. Int J Dev Biol 53:335–354. doi: 10.1387/ijdb.082717ph

6. Nielsen SJ, Schneider R, Bauer UM, et al. (2001) Rb targets histone H3 methylation and HP1 to promoters. Nature 412:561–565. doi: 10.1038/35087620

7. Peters AHFM, Mermoud JE, O'Carroll D, et al. (2002) Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. Nat Genet 30:77–80. doi: 10.1038/ng789

8. Cao R, Wang H, He J, et al. (2008) Role of hPHF1 in H3K27 methylation and Hox gene silencing. Mol Cell Biol 28:1862–1872. doi: 10.1128/MCB.01589-07

9. Schones DE, Zhao K (2008) Genome-wide approaches to studying chromatin modifications. Nature Publishing Group 9:179–191. doi: 10.1038/nrg2270

10. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi: 10.1186/gb-2009-10-3-r25

11. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi: 10.1093/bioinformatics/btp324

12. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology 26:1351–1359. doi: 10.1038/nbt.1508

13. Zhang Y, Liu T, Meyer CA, et al. (2008) Model-based analysis of ChIP-Seq

(MACS). Genome Biol 9:R137. doi: 10.1186/gb-2008-9-9-r137

14. Zang C, Schones DE, Zeng C, et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25:1952–1958. doi: 10.1093/bioinformatics/btp340

15. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440–9445. doi: 10.1073/pnas.1530509100

16. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369–73. doi: 10.1093/nar/gkl198

17. Heinz S, Benner C, Spann N, et al. (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Molecular Cell 38:576–589. doi: 10.1016/j.molcel.2010.05.004

18. Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. doi: 10.1038/75556

19. McLean CY, Bristor D, Hiller M, et al. (2010) GREAT improves functional interpretation of *cis*-regulatory regions. Nature Biotechnology 28:nbt.1630–9. doi: 10.1038/nbt.1630

20. Ernst J, Kheradpour P, Mikkelsen TS, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473:43–49. doi: 10.1038/nature09906

21. Hoffman MM, Buske OJ, Wang J, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9:473–476. doi: 10.1038/nmeth.1937

22. Heintzman ND, Stuart RK, Hon G, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311–318. doi: 10.1038/ng1966

23. Barski A, Cuddapah S, Cui K, et al. (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129:823–837. doi: 10.1016/j.cell.2007.05.009

24. Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641. doi: 10.1016/j.cell.2009.02.006

25. Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. doi: 10.1038/35057062

26. Kapranov P, Cawley SE, Drenkow J, et al. (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919. doi: 10.1126/science.1068597

27. Birney E, Stamatoyannopoulos JA, Dutta A, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816. doi: 10.1038/nature05874

28. Guttman M, Amit I, Garber M, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227. doi: 10.1038/nature07672

29. Cabili MN, Trapnell C, Goff L, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927. doi: 10.1101/gad.17446611

30. Cheng J, Kapranov P, Drenkow J, et al. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154. doi: 10.1126/science.1108625

31. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Molecular Cell 43:904–914. doi: 10.1016/j.molcel.2011.08.018

32. Nagano T, Mitchell JA, Sanz LA, et al. (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720. doi: 10.1126/science.1163802

33. Zhao J, Ohsumi TK, Kung JT, et al. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. Molecular Cell 40:939–953. doi: 10.1016/j.molcel.2010.12.011

34. Wang D, Garcia-Bassets I, Benner C, et al. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474:390–394. doi: 10.1038/nature10006

35. Ørom UA, Derrien T, Beringer M, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58. doi: 10.1016/j.cell.2010.09.001

36. Shiraki T, Kondo S, Katayama S, et al. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 100:15776–15781. doi: 10.1073/pnas.2136655100

37. Lassmann T, Frings O, Sonnhammer ELL (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37:858–865. doi: 10.1093/nar/gkn1006

38. Carninci P, Sandelin A, Lenhard B, et al. (2006) Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635. doi: 10.1038/ng1789

39. Ong C-T, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Publishing Group 12:284–293. doi: 10.1038/nrg2957

40. Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annu Rev Genom Human Genet 7:29–59. doi: 10.1146/annurev.genom.7.080505.115623

41. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotechnology 28:817–825. doi: 10.1038/nbt.1662

42. Zeitlinger J, Stark A, Kellis M, et al. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet 39:1512–1516. doi: 10.1038/ng.2007.26

43. Akalin A, Fredman D, Arner E, et al. (2009) Transcriptional features of genomic regulatory blocks. Genome Biol 10:R38. doi: 10.1186/gb-2009-10-4-r38

44. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. Genes Dev 25:1010–1022. doi: 10.1101/gad.2037511

45. Bernstein BE, Mikkelsen TS, Xie X, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125:315–326. doi: 10.1016/j.cell.2006.02.041

46. Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. Cell 144:327–339. doi: 10.1016/j.cell.2011.01.024

47. Buecker C, Wysocka J (2012) Enhancers as information integration hubs in development: lessons from genomics. Trends Genet 28:276–284. doi: 10.1016/j.tig.2012.02.008

48. Amano T, Sagai T, Tanabe H, et al. (2009) Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Developmental Cell 16:47–57. doi: 10.1016/j.devcel.2008.11.011

49. Li Z, Gadue P, Chen K, et al. (2012) Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. Cell 151:1608–1616. doi: 10.1016/j.cell.2012.11.018

50. Goldberg AD, Banaszynski LA, Noh K-M, et al. (2010) Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell 140:678–691. doi: 10.1016/j.cell.2010.01.003

51. Zentner GE, Tesar PJ, Scacheri PC (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Research 21:1273–1283. doi: 10.1101/gr.122382.111

52. Creyghton MP, Cheng AW, Welstead GG, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 107:21931–21936. doi: 10.1073/pnas.1016071107

53. Rada-Iglesias A, Bajpai R, Swigut T, et al. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279–283. doi: 10.1038/nature09692

54. Ghisletti S, Barozzi I, Mietton F, et al. (2010) Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. Immunity 32:317–328. doi: 10.1016/j.immuni.2010.02.008

55. Jin Q, Yu L-R, Wang L, et al. (2011) Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. EMBO J 30:249–262. doi: 10.1038/emboj.2010.318

56. Ogbourne S, Antalis TM (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem J 331 (Pt 1):1–14.

57. Valenzuela L, Kamakaka RT (2006) Chromatin insulators. Annu Rev Genet 40:107–138. doi: 10.1146/annurev.genet.39.073003.113546

58. Bushey AM, Dorman ER, Corces VG (2008) Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Molecular Cell 32:1–9. doi: 10.1016/j.molcel.2008.08.017

59. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G (2002) Locus control regions. Blood 100:3077–3086. doi: 10.1182/blood-2002-04-1104

60. Chotinantakul K, Leeanansaksiri W (2012) Hematopoietic Stem Cell Development, Niches, and Signaling Pathways. Bone Marrow Research 2012:1–16. doi: 10.1155/2012/270425

61. Huang X, Cho S, Spangrude GJ (2007) Hematopoietic stem cells: generation and self-renewal. Cell Death Differ 14:1851–1859. doi: 10.1038/sj.cdd.4402225

62. Askenasy N, Zorina T, Farkas DL, Shalit I (2002) Transplanted hematopoietic cells seed in clusters in recipient bone marrow in vivo. Stem Cells 20:301–310. doi: 10.1634/stemcells.20-4-301

63. Potocnik AJ, Brakebusch C, Fässler R (2000) Fetal and adult hematopoietic stem cells require beta1 integrin function for colonizing fetal liver, spleen, and bone marrow. Immunity 12:653–663.

64. Sugiyama T, Kohara H, Noda M, Nagasawa T (2006) Maintenance of the hematopoietic stem cell pool by CXCL12-CXCR4 chemokine signaling in bone marrow stromal cell niches. Immunity 25:977–988. doi: 10.1016/j.immuni.2006.10.016

65. Zhang J, Niu C, Ye L, et al. (2003) Identification of the haematopoietic stem cell niche and control of the niche size. Nature 425:836–841. doi: 10.1038/nature02041

66. Wang JC, Doedens M, Dick JE (1997) Primitive human hematopoietic cells are enriched in cord blood compared with adult bone marrow or mobilized peripheral blood as measured by the quantitative in vivo SCID-repopulating cell assay. Blood 89:3919–3924.

67. Reitsma MJ, Lee BR, Uchida N (2002) Method for purification of human hematopoietic stem cells by flow cytometry. Methods Mol Med 63:59–77. doi: 10.1385/1-59259-140-X:059

68. Baum CM, Weissman IL, Tsukamoto AS, et al. (1992) Isolation of a candidate human hematopoietic stem-cell population. Proc Natl Acad Sci USA 89:2804–2808.

69. Bhatia M, Wang JC, Kapp U, et al. (1997) Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. Proc Natl Acad Sci USA 94:5320–5325.

70. Gallacher L, Murdoch B, Wu DM, et al. (2000) Isolation and characterization of human CD34(-)Lin(-) and CD34(+)Lin(-) hematopoietic stem cells using cell surface markers AC133 and CD7. Blood 95:2813–2820.

71. Orkin SH, Zon LI (2008) Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132:631–644. doi: 10.1016/j.cell.2008.01.025

72. Kim S-I, Bresnick EH (2007) Transcriptional control of erythropoiesis: emerging mechanisms and principles. Oncogene 26:6777–6794. doi: 10.1038/sj.onc.1210761

73. Orkin SH (2000) Diversification of haematopoietic stem cells to specific lineages. Nat Rev Genet 1:57–64. doi: 10.1038/35049577

74. Vicente C, Conchillo A, García-Sánchez MA, Odero MD (2012) The role of the GATA2 transcription factor in normal and malignant hematopoiesis. Crit Rev Oncol Hematol 82:1–17. doi: 10.1016/j.critrevonc.2011.04.007

75. Bryder D, Rossi DJ, Weissman IL (2006) Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. Am J Pathol 169:338–346. doi: 10.2353/ajpath.2006.060312

76. Antonchuk J, Sauvageau G, Humphries RK (2002) HOXB4-induced expansion of adult hematopoietic stem cells ex vivo. Cell 109:39–45.

77. Papathanasiou P, Attema JL, Karsunky H, et al. (2009) Self-renewal of the long-term reconstituting subset of hematopoietic stem cells is regulated by Ikaros. Stem Cells 27:3082–3092. doi: 10.1002/stem.232

78. Akashi K, Traver D, Miyamoto T, Weissman IL (2000) A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature 404:193–197. doi: 10.1038/35004599

79. Nerlov C, Querfurth E, Kulessa H, Graf T (2000) GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. Blood 95:2543–2551.

80. Gao Z, Huang Z, Olivey HE, et al. (2010) FOG-1-mediated recruitment of NuRD is required for cell lineage re-enforcement during haematopoiesis. EMBO J 29:457–468. doi: 10.1038/emboj.2009.368

81. O'Connell RM, Zhao JL, Rao DS (2011) MicroRNA function in myeloid biology. Blood 118:2960–2969. doi: 10.1182/blood-2011-03-291971

82. Majeti R, Park CY, Weissman IL (2007) Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. Cell Stem Cell 1:635–645. doi: 10.1016/j.stem.2007.10.001

83. Notta F, Doulatov S, Laurenti E, et al. (2011) Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. Science 333:218–221. doi: 10.1126/science.1201219

84. Cui K, Zang C, Roh T-Y, et al. (2009) Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation. Stem Cell 4:80–93. doi: 10.1016/j.stem.2008.11.011

85. Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, et al. (2014) Immunogenetics. Chromatin state dynamics during blood formation. Science 345:943–949. doi: 10.1126/science.1256271

86. Roselli EA, Mezzadra R, Frittoli MC, et al. (2010) Correction of beta-thalassemia major by gene transfer in haematopoietic progenitors of pediatric patients. EMBO Mol Med 2:315–328. doi: 10.1002/emmm.201000083

87. Ferrari F, Bortoluzzi S, Coppe A, et al. (2007) Novel definition files for human GeneChips based on GeneAnnot. BMC Bioinformatics 8:446. doi: 10.1186/1471-2105-8-446

88. Li J, Wong L (2001) Emerging patterns and gene expression data. Genome Inform 12:3–13.

89. Cattoglio C, Maruggi G, Bartholomae C, et al. (2010) High-Definition Mapping of Retroviral Integration Sites Defines the Fate of Allogeneic T Cells After Donor Lymphocyte Infusion. PLoS ONE 5:e15688. doi: 10.1371/journal.pone.0015688

90. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi: 10.1093/bioinformatics/btp352

91. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi: 10.1093/bioinformatics/btq033

92. Landt SG, Marinov GK, Kundaje A, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22:1813–1831. doi: 10.1101/gr.136184.111

93. Shen L, Shao N, Liu X, Nestler E (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics 15:284. doi: 10.1186/1471-2164-15-284

94. Carninci P, Westover A, Nishiyama Y, et al. (1997) High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res 4:61–66.

95. Faulkner GJ, Forrest ARR, Chalk AM, et al. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 91:281–288. doi: 10.1016/j.ygeno.2007.11.003

96. Khalil AM, Guttman M, Huarte M, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106:11667–11672. doi: 10.1073/pnas.0904715106

97. Kretz M, Webster DE, Flockhart RJ, et al. (2012) Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev 26:338–343. doi: 10.1101/gad.182121.111

98. Wilming LG, Gilbert JGR, Howe K, et al. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36:D753–60. doi: 10.1093/nar/gkm987

99. Zhang Y, Liu XS, Liu Q-R, Wei L (2006) Genome-wide in silico identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. Nucleic Acids Res 34:3465–3475. doi: 10.1093/nar/gkl473

100. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. doi: 10.1093/bioinformatics/btp616

101. Szilvassy SJ (2003) The biology of hematopoietic stem cells. Arch Med Res 34:446–460. doi: 10.1016/j.arcmed.2003.06.004

102. Li J, Hale J, Bhagia P, et al. (2014) Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. Blood 124:3636–3645. doi: 10.1182/blood-2014-07-588806

103. Gaines P, Berliner N (2005) Differentiation and characterization of myeloid cells. Curr Protoc Immunol Chapter 22:Unit 22F.5–22F.5.14. doi: 10.1002/0471142735.im22f05s67

104. Jung YL, Luquette LJ, Ho JWK, et al. (2014) Impact of sequencing depth in ChIP-seq experiments. Nucleic Acids Res 42:e74–e74. doi: 10.1093/nar/gku178

105. Ciau-Uitz A, Wang L, Patient R, Liu F (2013) ETS transcription factors in hematopoietic stem cell development. Blood Cells, Molecules, and Diseases 51:248–255. doi: 10.1016/j.bcmd.2013.07.010

106. Liebermann DA, Gregory B, Hoffman B (1998) AP-1 (Fos/Jun) transcription factors in hematopoietic differentiation and apoptosis. Int J Oncol 12:685–700.

107. Wang J, Sun Q, Morita Y, et al. (2012) A Differentiation Checkpoint Limits Hematopoietic Stem Cell Self-Renewal in Response to DNA Damage. Cell 148:1001–1014. doi: 10.1016/j.cell.2012.01.040

108. Itoh-Nakadai A, Hikota R, Muto A, et al. (2014) The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. Nat Immunol 15:1171–1180. doi: 10.1038/ni.3024

109. Ichikawa M, Yoshimi A, Nakagawa M, et al. (2013) A role for RUNX1 in hematopoiesis and myeloid leukemia. Int J Hematol 97:726–734. doi: 10.1007/s12185-013-1347-3

110. Sato T, Onai N, Yoshihara H, et al. (2009) Interferon regulatory factor-2 protects quiescent hematopoietic stem cells from type I interferon–dependent exhaustion. Nat Med 15:696–700. doi: 10.1038/nm.1973

111. Costa RMB, Soto X, Chen Y, et al. (2008) spib is required for primitive myeloid development in Xenopus. Blood 112:2287–2296. doi: 10.1182/blood-2008-04-150268

112. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. doi: 10.1038/nprot.2008.211

113. Redmond AM, Carroll JS (2013) Enhancer-derived RNAs: “spicing up” transcription programs. EMBO J 32:2096–2098. doi: 10.1038/emboj.2013.151

114. Li W, Notani D, Ma Q, et al. (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498:516–520. doi: 10.1038/nature12210

115. Hu G, Schones DE, Cui K, et al. (2011) Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. Genome Research 21:1650–1658. doi: 10.1101/gr.121145.111

116. Pham T-H, Benner C, Lichtinger M, et al. (2012) Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. Blood 119:e161–71. doi: 10.1182/blood-2012-01-402453

117. Xu J, Shao Z, Glass K, et al. (2012) Combinatorial Assembly of Developmental Stage-Specific Enhancers Controls Gene Expression Programs during Human Erythropoiesis. Developmental Cell 23:796–811. doi: 10.1016/j.devcel.2012.09.003

118. Saeed S, Quintin J, Kerstens HHD, et al. (2014) Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. Science 345:1251086–1251086. doi: 10.1126/science.1251086

6. APPENDIX

6.1. Antibodies used for FACS analysis.

Antibody	Catalog #	Company
GpA-PE	R7078	Dako
GpA-APC	551336	BD Pharmingen
CD11b-APC	553312	BD Pharmingen
CD11b-FITC	553310	BD Pharmingen
CD11b-pe	553311	BD Pharmingen
CD33- APC	551378	BD Pharmingen
CD13-PE	MHCD1304	CALTAG
CD34-FITC	345801	BD Pharmingen
CD34-PE	345802	BD Pharmingen
CD36-FITC	555454	BD Pharmingen
CD38-APC	555462	BD Pharmingen
CD71-PE	IM2001U	Beckman Coulter
CD71 PerCP	551374	BD Pharmingen
CD133/2(293C3)-PE	130-090-853	Miltenyi Biotech
CD14-FITC	555393	BD Pharmingen

6.2. Differentially expressed genes in erythroid commitment detected by microarray analysis.

Gene	Gene Name	FC	P value
HBBP1	hemoglobin, beta pseudogene 1	8.03	0.000001
HBB	hemoglobin, beta	7.42	0.000046
HBE1	hemoglobin, epsilon 1	7.26	0.000506
CD36	CD36 molecule (thrombospondin receptor)	6.73	0.000005
RHAG	Rh-associated glycoprotein	6.32	0.000011
CNRIP1	cannabinoid receptor interacting protein 1	6.28	0.005413
TUBB2A	tubulin, beta 2A	6.20	0.000044
NMU	neuromedin U	6.17	0.001302
XK	X-linked Kx blood group (McLeod syndrome)	6.13	0.000027
ERAF	erythroid associated factor	5.85	0.000640
KLF1	Kruppel-like factor 1 (erythroid)	5.40	0.000265
PF4	platelet factor 4	5.28	0.003868
APOC1	apolipoprotein C-I	5.03	0.000018
EPB42	erythrocyte membrane protein band 4.2	5.00	0.001048
KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2	4.94	0.000124
S100A8	S100 calcium binding protein A8	4.90	0.010992
SPTA1	spectrin, alpha, erythrocytic 1 (elliptocytosis 2)	4.85	0.000068
TMEM56	transmembrane protein 56	4.85	0.002134
HBD	hemoglobin, delta	4.75	0.010117
CA2	carbonic anhydrase II	4.72	0.004686
ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	4.49	0.000020
TIMP3	TIMP metallopeptidase inhibitor 3	4.46	0.000087
SLC6A8	solute carrier family 6 (neurotransmitter transporter, creatine), member 8	4.45	0.006580
IRF6	interferon regulatory factor 6	4.37	0.002134
KEL	Kell blood group, metallo-endopeptidase	4.36	0.000011
TSC22D3	TSC22 domain family, member 3	4.35	0.001438

Gene	Gene Name	FC	P value
PRG2	proteoglycan 2, bone marrow (natural killer cell activator, eosinophil granule major basic protein)	4.32	0.004047
ANK1	ankyrin 1, erythrocytic	4.28	0.000031
ALAS2	aminolevulinate, delta-, synthase 2	4.27	0.002038
PKLR	pyruvate kinase, liver and RBC	4.24	0.001943
RFESD	Rieske (Fe-S) domain containing	4.18	0.000068
ABCC4	ATP-binding cassette, sub-family C (CFTR/MRP), member 4	4.16	0.000007
PNMT	phenylethanolamine N-methyltransferase	4.14	0.002940
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	4.05	0.000001
DHRS3	dehydrogenase/reductase (SDR family) member 3	4.03	0.004707
GATA1	GATA binding protein 1 (globin transcription factor 1)	4.03	0.000503
BLVRB	biliverdin reductase B (flavin reductase (NADPH))	4.03	0.000495
PROS1	protein S (alpha)	4.00	0.000575
FAM132B	family with sequence similarity 132, member B	3.99	0.000600
CHST2	carbohydrate (N-acetylglucosamine-6-O) sulfotransferase 2	3.98	0.000200
DDIT4	DNA-damage-inducible transcript 4	3.95	0.000967
YPEL4	yippee-like 4 (Drosophila)	3.94	0.000002
GAD1	glutamate decarboxylase 1 (brain, 67kDa)	3.90	0.003131
ALOX5	arachidonate 5-lipoxygenase	3.87	0.002718
CLC	Charcot-Leyden crystal protein	3.79	0.043636
ICAM4	intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)	3.78	0.000346
MYO1D	myosin ID	3.77	0.006220
SEPT10	septin 10	3.75	0.000025
TRIB2	tribbles homolog 2 (Drosophila)	3.73	0.000017
CD24	CD24 molecule	3.73	0.046025
TSPAN6	tetraspanin 6	3.62	0.000025
PMP22	peripheral myelin protein 22	3.49	0.001247
GYPA	glycophorin A (MNS blood group)	3.47	0.009992
ALAS1	aminolevulinate, delta-, synthase 1	3.46	0.000010
HBZ	hemoglobin, zeta	3.45	0.001038
FAM171A1	family with sequence similarity 171, member A1	3.44	0.002002
ATP7B	ATPase, Cu++ transporting, beta polypeptide	3.44	0.000590
APOE	apolipoprotein E	3.44	0.019903
OSBPL6	oxysterol binding protein-like 6	3.37	0.002182
TUBB1	tubulin, beta 1	3.37	0.007354
JAZF1	JAZF zinc finger 1	3.36	0.000144
PPEF1	protein phosphatase, EF-hand calcium binding domain 1	3.36	0.017919
FAM178B	family with sequence similarity 178, member B	3.35	0.019388
NDFIP2	Nedd4 family interacting protein 2	3.34	0.000012
GNAQ	guanine nucleotide binding protein (G protein), q polypeptide	3.33	0.000058
TMOD1	tropomodulin 1	3.30	0.000143
PKM2	pyruvate kinase, muscle	3.30	0.001942
HES6	hairy and enhancer of split 6 (Drosophila)	3.28	0.001757
ZFPM1	zinc finger protein, multitype 1	3.28	0.000246
LXN	latexin	3.27	0.000968
CLCN4	chloride channel 4	3.27	0.000001
CMTM5	CKLF-like MARVEL transmembrane domain containing 5	3.26	0.005450
LEPR	leptin receptor	3.22	0.002566
HBM	hemoglobin, mu	3.18	0.002604
C2orf88	chromosome 2 open reading frame 88	3.12	0.001122
CBS	cystathionine-beta-synthase	3.07	0.005707
PLK1	polo-like kinase 1 (Drosophila)	3.04	0.001060
ADFP	adipose differentiation-related protein	3.03	0.000006
LOC388588	hypothetical LOC388588	3.01	0.000884
EPB49	erythrocyte membrane protein band 4.9 (dematin)	3.01	0.002203

Gene	Gene Name	FC	P value
GFI1B	growth factor independent 1B transcription repressor	3.01	0.000122
G0S2	G0/G1switch 2	3.00	0.000113
AQP1	aquaporin 1 (Colton blood group)	2.99	0.002404
IL2RG	interleukin 2 receptor, gamma (severe combined immunodeficiency)	2.98	0.008690
NUCB1	nucleobindin 1	2.98	0.000366
IL2RA	interleukin 2 receptor, alpha	2.97	0.000043
CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	2.94	0.000083
RTN1	reticulon 1	2.92	0.038727
MUC1	mucin 1, cell surface associated	2.91	0.000299
HBQ1	hemoglobin, theta 1	2.90	0.000007
FKBP5	FK506 binding protein 5	2.88	0.000086
MYL4	myosin, light chain 4, alkali; atrial, embryonic	2.87	0.005637
LMNA	lamin A/C	2.86	0.000179
WFDC1	WAP four-disulfide core domain 1	2.85	0.014183
KLHDC8B	kelch domain containing 8B	2.84	0.002207
TUBB2B	tubulin, beta 2B	2.82	0.010122
SLC16A9	solute carrier family 16, member 9 (monocarboxylic acid transporter 9)	2.82	0.041530
LMAN1	lectin, mannose-binding, 1	2.82	0.001399
PBK	PDZ binding kinase	2.80	0.000088
PDLIM1	PDZ and LIM domain 1	2.80	0.000875
PPAP2A	phosphatidic acid phosphatase type 2A	2.79	0.001531
TFR2	transferrin receptor 2	2.79	0.003792
DNAJA4	DnaJ (Hsp40) homolog, subfamily A, member 4	2.79	0.006862
TGM2	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)	2.79	0.011459
FCGR2B	Fc fragment of IgG, low affinity IIb, receptor (CD32)	2.78	0.003071
P4HA2	prolyl 4-hydroxylase, alpha polypeptide II	2.78	0.000507
ITGAM	integrin, alpha M (complement component 3 receptor 3 subunit)	2.78	0.020347
ADD2	adducin 2 (beta)	2.77	0.001392
SH2D2A	SH2 domain protein 2A	2.76	0.031300
RGS16	regulator of G-protein signaling 16	2.75	0.002007
EPS8	epidermal growth factor receptor pathway substrate 8	2.75	0.005551
LOX	lysyl oxidase	2.74	0.030823
GTF2I	general transcription factor II, i	2.71	0.000070
REPS2	RALBP1 associated Eps domain containing 2	2.71	0.000129
BNIP3	BCL2/adenovirus E1B 19kDa interacting protein 3	2.71	0.020271
ABCB6	ATP-binding cassette, sub-family B (MDR/TAP), member 6	2.70	0.000164
FHL2	four and a half LIM domains 2	2.69	0.000124
CAST	calpastatin	2.68	0.000004
FZD3	frizzled homolog 3 (Drosophila)	2.68	0.001500
GSTM3	glutathione S-transferase mu 3 (brain)	2.67	0.022365
NCKAP1	NCK-associated protein 1	2.67	0.000688
DARC	Duffy blood group, chemokine receptor	2.66	0.000131
TRAPPC1	trafficking protein particle complex 1	2.66	0.000743
PLD3	phospholipase D family, member 3	2.66	0.001067
COL18A1	collagen, type XVIII, alpha 1	2.65	0.000027
BCL2L11	BCL2-like 11 (apoptosis facilitator)	2.63	0.000600
ARL4A	ADP-ribosylation factor-like 4A	2.63	0.000777
TLE1	transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)	2.62	0.001484
C20orf108	chromosome 20 open reading frame 108	2.62	0.000031
PRKAR2B	protein kinase, cAMP-dependent, regulatory, type II, beta	2.59	0.000123
HDHD3	haloacid dehalogenase-like hydrolase domain containing 3	2.57	0.002685

Gene	Gene Name	FC	P value
PARVB	parvin, beta	2.57	0.000260
ADORA2B	adenosine A2b receptor	2.56	0.006139
THBS1	thrombospondin 1	2.54	0.010131
TPM1	tropomyosin 1 (alpha)	2.52	0.001261
C17orf99	chromosome 17 open reading frame 99	2.52	0.009234
SLAMF1	signaling lymphocytic activation molecule family member 1	2.51	0.038275
KLF9	Kruppel-like factor 9	2.50	0.001700
MPP1	membrane protein, palmitoylated 1, 55kDa	2.50	0.000033
ANKRD57	ankyrin repeat domain 57	2.49	0.001585
ALOX5AP	arachidonate 5-lipoxygenase-activating protein	2.49	0.002462
AMMECR1	Alport syndrome, mental retardation, midface hypoplasia and elliptocytosis chromosomal region gene 1	2.48	0.000003
G6PD	glucose-6-phosphate dehydrogenase	2.48	0.000878
EMP3	epithelial membrane protein 3	2.47	0.000081
C1orf150	chromosome 1 open reading frame 150	2.46	0.018192
ST6GALNAC1	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 1	2.46	0.007024
ERRFI1	ERBB receptor feedback inhibitor 1	2.43	0.039253
CD46	CD46 molecule, complement regulatory protein	2.43	0.002322
S100A10	S100 calcium binding protein A10	2.41	0.000162
HIPK3	homeodomain interacting protein kinase 3	2.41	0.000970
TNIK	TRAF2 and NCK interacting kinase	2.40	0.000191
IL9R	interleukin 9 receptor	2.40	0.000143
ALDH6A1	aldehyde dehydrogenase 6 family, member A1	2.40	0.006546
CTSH	cathepsin H	2.40	0.018772
NTRK1	neurotrophic tyrosine kinase, receptor, type 1	2.39	0.026085
STON1	stonin 1	2.39	0.006653
NCOA2	nuclear receptor coactivator 2	2.39	0.000103
ALAD	aminolevulinate, delta-, dehydratase	2.38	0.017645
FREQ	frequenin homolog (Drosophila)	2.38	0.002354
ALDOC	aldolase C, fructose-bisphosphate	2.38	0.040266
MAN1A1	mannosidase, alpha, class 1A, member 1	2.35	0.000176
TXNIP	thioredoxin interacting protein	2.34	0.001908
CCNA1	cyclin A1	2.33	0.007367
C19orf59	chromosome 19 open reading frame 59	2.33	0.036549
GAS2L1	growth arrest-specific 2 like 1	2.33	0.000389
EPAS1	endothelial PAS domain protein 1	2.32	0.000629
LRP11	low density lipoprotein receptor-related protein 11	2.32	0.000364
PRKAB1	protein kinase, AMP-activated, beta 1 non-catalytic subunit	2.32	0.005656
ERMAP	erythroblast membrane-associated protein (Scianna blood group)	2.30	0.000591
REEP3	receptor accessory protein 3	2.29	0.001283
ATP8B1	ATPase, class I, type 8B, member 1	2.28	0.049081
PRKCSH	protein kinase C substrate 80K-H	2.28	0.000032
IGFBP2	insulin-like growth factor binding protein 2, 36kDa	2.27	0.002099
TPSAB1	tryptase alpha/beta 1	2.26	0.017475
VEGFA	vascular endothelial growth factor A	2.25	0.007244
GEM	GTP binding protein overexpressed in skeletal muscle	2.25	0.004995
H1F0	H1 histone family, member 0	2.25	0.002134
IL1RL1	interleukin 1 receptor-like 1	2.24	0.016679
UBXN10	UBX domain protein 10	2.22	0.001364
MTSS1	metastasis suppressor 1	2.22	0.000029
S100A6	S100 calcium binding protein A6	2.21	0.000172
EPCAM	epithelial cell adhesion molecule	2.21	0.009886
PPP1R15A	protein phosphatase 1, regulatory (inhibitor) subunit 15A	2.21	0.000168
ADAMTS3	ADAM metallopeptidase with thrombospondin type 1 motif, 3	2.20	0.000160

Gene	Gene Name	FC	P value
PIP5K1B	phosphatidylinositol-4-phosphate 5-kinase, type I, beta	2.20	0.000685
WBP2	WW domain binding protein 2	2.20	0.000066
SLC24A3	solute carrier family 24 (sodium/potassium/calcium exchanger), member 3	2.20	0.018555
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	2.19	0.021642
SNAP23	synaptosomal-associated protein, 23kDa	2.19	0.004031
RARRES1	retinoic acid receptor responder (tazarotene induced) 1	2.18	0.010656
C13orf15	chromosome 13 open reading frame 15	2.18	0.009746
RTN2	reticulon 2	2.18	0.000990
CD37	CD37 molecule	2.18	0.001087
FYN	FYN oncogene related to SRC, FGR, YES	2.17	0.011288
ZNF192	zinc finger protein 192	2.17	0.002480
WNK3	WNK lysine deficient protein kinase 3	2.17	0.041102
ANKRD37	ankyrin repeat domain 37	2.16	0.001488
AKAP12	A kinase (PRKA) anchor protein 12	2.16	0.021380
PTGS1	prostaglandin-endoperoxide synthase 1 (prostaglandin G/H synthase and cyclooxygenase)	2.16	0.000362
SLC25A37	solute carrier family 25, member 37	2.15	0.014699
P2RX5	purinergic receptor P2X, ligand-gated ion channel, 5	2.15	0.008397
ATP6V0A1	ATPase, H+ transporting, lysosomal V0 subunit a1	2.15	0.000093
ABHD14B	abhydrolase domain containing 14B	2.14	0.000417
BHLHE40	basic helix-loop-helix family, member e40	2.14	0.001208
AGPAT1	1-acylglycerol-3-phosphate O-acyltransferase 1 (lysophosphatidic acid acyltransferase, alpha)	2.14	0.000011
SLC44A2	solute carrier family 44, member 2	2.14	0.002525
LCN2	lipocalin 2	2.13	0.006845
TPST2	tyrosylprotein sulfotransferase 2	2.13	0.000207
CAMK1	calcium/calmodulin-dependent protein kinase I	2.11	0.001968
SOCS1	suppressor of cytokine signaling 1	2.11	0.000003
MAZ	MYC-associated zinc finger protein (purine-binding transcription factor)	2.11	0.007220
APOBEC3C	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3C	2.10	0.000087
EREG	epiregulin	2.09	0.037605
PIR	pirin (iron-binding nuclear protein)	2.09	0.002241
TIMP1	TIMP metallopeptidase inhibitor 1	2.09	0.008648
CDC42EP4	CDC42 effector protein (Rho GTPase binding) 4	2.09	0.000373
FECH	ferrochelatase (protoporphyria)	2.09	0.001234
MYRIP	myosin VIIA and Rab interacting protein	2.08	0.021268
PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	2.08	0.006428
LOC541471	hypothetical LOC541471	2.07	0.000435
STXBP6	syntaxin binding protein 6 (amisyn)	2.07	0.004146
NAPA	N-ethylmaleimide-sensitive factor attachment protein, alpha	2.06	0.000412
TMEM214	transmembrane protein 214	2.05	0.012853
CPA3	carboxypeptidase A3 (mast cell)	2.05	0.001252
GALC	galactosylceramidase	2.05	0.000110
RAB6B	RAB6B, member RAS oncogene family	2.04	0.019713
PDZD8	PDZ domain containing 8	2.04	0.003795
FRAT1	frequently rearranged in advanced T-cell lymphomas	2.04	0.002574
ENSG00000183700	NA	2.04	0.002810
ADAM10	ADAM metallopeptidase domain 10	2.04	0.002900
BSG	basigin (Ok blood group)	2.04	0.000181
ELL2	elongation factor, RNA polymerase II, 2	2.03	0.000372
MOSPD3	motile sperm domain containing 3	2.03	0.000089
IL10RA	interleukin 10 receptor, alpha	2.03	0.014880
FHDC1	FH2 domain containing 1	2.02	0.009671

Gene	Gene Name	FC	P value
LPAR5	lysophosphatidic acid receptor 5	2.02	0.002706
SLC4A2	solute carrier family 4, anion exchanger, member 2 (erythrocyte membrane protein band 3-like 1)	2.01	0.001793
TAL1	T-cell acute lymphocytic leukemia 1	2.00	0.000340
TARDBP	TAR DNA binding protein	-2.00	0.000083
TMEM163	transmembrane protein 163	-2.01	0.017745
FSTL1	follistatin-like 1	-2.01	0.002255
SLC22A4	solute carrier family 22 (organic cation/ergothioneine transporter), member 4	-2.01	0.001119
SGK3	serum/glucocorticoid regulated kinase family, member 3	-2.01	0.002213
CORO1A	coronin, actin binding protein, 1A	-2.01	0.000244
CD34	CD34 molecule	-2.02	0.013742
RSPH10B2	radial spoke head 10 homolog B2 (Chlamydomonas)	-2.02	0.000832
BEX5	brain expressed, X-linked 5	-2.02	0.001639
RHOBTB1	Rho-related BTB domain containing 1	-2.03	0.021154
IFI16	interferon, gamma-inducible protein 16	-2.04	0.014353
P2RY8	purinergic receptor P2Y, G-protein coupled, 8	-2.04	0.003957
STYK1	serine/threonine/tyrosine kinase 1	-2.04	0.010581
ZNF573	zinc finger protein 573	-2.04	0.047831
BAX	BCL2-associated X protein	-2.04	0.002302
SIDT1	SID1 transmembrane family, member 1	-2.05	0.000065
PABPC4L	poly(A) binding protein, cytoplasmic 4-like	-2.05	0.000015
SNX10	sorting nexin 10	-2.06	0.000125
NSUN6	NOL1/NOP2/Sun domain family, member 6	-2.06	0.000316
DNAJC17	DnaJ (Hsp40) homolog, subfamily C, member 17	-2.06	0.000093
SH3TC1	SH3 domain and tetratricopeptide repeats 1	-2.07	0.001603
NRIP3	nuclear receptor interacting protein 3	-2.07	0.001599
PSTPIP1	proline-serine-threonine phosphatase interacting protein 1	-2.07	0.002279
ALCAM	activated leukocyte cell adhesion molecule	-2.08	0.015837
S100Z	S100 calcium binding protein Z	-2.08	0.000753
ZC3H12D	zinc finger CCCH-type containing 12D	-2.08	0.009060
SOX4	SRY (sex determining region Y)-box 4	-2.08	0.000171
ATL1	atlastin GTPase 1	-2.08	0.000343
BEX2	brain expressed X-linked 2	-2.09	0.017168
NUTF2	nuclear transport factor 2	-2.09	0.014930
ASGR1	asialoglycoprotein receptor 1	-2.11	0.000220
CERK	ceramide kinase	-2.12	0.000077
FAM43A	family with sequence similarity 43, member A	-2.13	0.001233
RGS19	regulator of G-protein signaling 19	-2.13	0.002537
KIAA1841	KIAA1841	-2.15	0.001256
PRAGMIN	homolog of rat pragma of Rnd2	-2.15	0.005921
GARNL4	GTPase activating Rap/RanGAP domain-like 4	-2.15	0.003128
CLEC5A	C-type lectin domain family 5, member A	-2.16	0.045442
DSE	dermatan sulfate epimerase	-2.18	0.007097
BAT2D1	BAT2 domain containing 1	-2.20	0.002105
GALNT3	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 3 (GalNAc-T3)	-2.20	0.000256
HERC5	hect domain and RLD 5	-2.20	0.003103
PHGDH	phosphoglycerate dehydrogenase	-2.21	0.001139
CASP1	caspase 1, apoptosis-related cysteine peptidase (interleukin 1, beta, convertase)	-2.21	0.043319
ANPEP	alanyl (membrane) aminopeptidase	-2.21	0.001708
ATP9A	ATPase, class II, type 9A	-2.22	0.007067
PRRT3	proline-rich transmembrane protein 3	-2.24	0.000155
FAM111B	family with sequence similarity 111, member B	-2.26	0.000053
MGC29506	hypothetical protein MGC29506	-2.27	0.002300

Gene	Gene Name	FC	P value
GPR183	G protein-coupled receptor 183	-2.27	0.020849
ZBTB8A	zinc finger and BTB domain containing 8A	-2.28	0.008525
TRBV27	T cell receptor beta variable 27	-2.28	0.014474
AMICA1	adhesion molecule, interacts with CXADR antigen 1	-2.28	0.022934
RAB27A	RAB27A, member RAS oncogene family	-2.28	0.000178
CACHD1	cache domain containing 1	-2.29	0.000312
MALAT1	metastasis associated lung adenocarcinoma transcript 1 (non-protein coding)	-2.30	0.000390
CCDC58	coiled-coil domain containing 58	-2.31	0.006058
CHST13	carbohydrate (chondroitin 4) sulfotransferase 13	-2.31	0.004531
C9orf91	chromosome 9 open reading frame 91	-2.31	0.002074
IMPA2	inositol(myo)-1(or 4)-monophosphatase 2	-2.32	0.001674
FNBP1	formin binding protein 1	-2.34	0.000009
ZNF738	zinc finger protein 738	-2.35	0.000664
RTP4	receptor (chemosensory) transporter protein 4	-2.36	0.000586
ZNF662	zinc finger protein 662	-2.37	0.000716
GAPT	GRB2-binding adaptor protein, transmembrane	-2.37	0.001412
RNASE3	ribonuclease, RNase A family, 3 (eosinophil cationic protein)	-2.38	0.000573
ZNF257	zinc finger protein 257	-2.38	0.000433
NKG7	natural killer cell group 7 sequence	-2.39	0.011124
DLK1	delta-like 1 homolog (Drosophila)	-2.40	0.009290
C12orf75	chromosome 12 open reading frame 75	-2.43	0.005260
GPX7	glutathione peroxidase 7	-2.44	0.004387
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)	-2.44	0.004600
ZNF492	zinc finger protein 492	-2.46	0.005513
IL17D	interleukin 17D	-2.48	0.002180
CPNE2	copine II	-2.48	0.007568
C1orf228	chromosome 1 open reading frame 228	-2.49	0.001539
GPR124	G protein-coupled receptor 124	-2.49	0.001546
C9orf43	chromosome 9 open reading frame 43	-2.50	0.000034
TMSB15A	thymosin beta 15a	-2.50	0.010245
EMP1	epithelial membrane protein 1	-2.52	0.001110
MDFIC	MyoD family inhibitor domain containing	-2.53	0.000534
NPDC1	neural proliferation, differentiation and control, 1	-2.53	0.001252
C1orf59	chromosome 1 open reading frame 59	-2.54	0.000001
ANKRD36B	ankyrin repeat domain 36B	-2.54	0.008665
IGHM	immunoglobulin heavy constant mu	-2.54	0.001262
ERG	v-ets erythroblastosis virus E26 oncogene homolog (avian)	-2.57	0.001374
LOXL1	lysyl oxidase-like 1	-2.58	0.000676
NLRC3	NLR family, CARD domain containing 3	-2.58	0.002137
CYYR1	cysteine/tyrosine-rich 1	-2.58	0.014357
IL12RB2	interleukin 12 receptor, beta 2	-2.59	0.000866
FAM26F	family with sequence similarity 26, member F	-2.59	0.016239
HOXA5	homeobox A5	-2.59	0.004721
DUSP4	dual specificity phosphatase 4	-2.59	0.000923
CD93	CD93 molecule	-2.59	0.004464
TIMP2	TIMP metallopeptidase inhibitor 2	-2.62	0.034073
HSF5	heat shock transcription factor family member 5	-2.63	0.000349
ZNF439	zinc finger protein 439	-2.63	0.004860
HOXA7	homeobox A7	-2.64	0.000106
TMEM71	transmembrane protein 71	-2.64	0.005164
LOC729680	hypothetical protein LOC729680	-2.64	0.001055
GPR84	G protein-coupled receptor 84	-2.65	0.000016
LOC643332	similar to Nonsecretory ribonuclease precursor (Ribonuclease US) (Eosinophil-derived neurotoxin) (RNase Upl-2) (Ribonuclease 2) (RNase 2)	-2.65	0.000513

Gene	Gene Name	FC	P value
RASGRP1	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	-2.65	0.001957
RASGRP2	RAS guanyl releasing protein 2 (calcium and DAG-regulated)	-2.71	0.000106
BZRAP1	benzodiazapine receptor (peripheral) associated protein 1	-2.72	0.000001
UCHL1	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	-2.73	0.004048
CRHBP	corticotropin releasing hormone binding protein	-2.73	0.000679
LMNB1	lamin B1	-2.73	0.000000
PION	pigeon homolog (Drosophila)	-2.74	0.000029
CALN1	calneuron 1	-2.76	0.000036
ARMCX1	armadillo repeat containing, X-linked 1	-2.77	0.005759
GBP1	guanylate binding protein 1, interferon-inducible, 67kDa	-2.79	0.017261
HOXA9	homeobox A9	-2.80	0.012864
IFITM1	interferon induced transmembrane protein 1 (9-27)	-2.80	0.034237
SPNS3	spinster homolog 3 (Drosophila)	-2.81	0.010428
TCTEX1D1	Tctex1 domain containing 1	-2.82	0.003789
FAM169A	family with sequence similarity 169, member A	-2.86	0.000835
MSRB3	methionine sulfoxide reductase B3	-2.87	0.004449
PXDN	peroxidasin homolog (Drosophila)	-2.90	0.000035
SH3BP4	SH3-domain binding protein 4	-2.92	0.000822
TRIM22	tripartite motif-containing 22	-2.92	0.001888
GZMA	granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	-2.93	0.001155
LTB	lymphotoxin beta (TNF superfamily, member 3)	-2.94	0.000258
TNFSF4	tumor necrosis factor (ligand) superfamily, member 4	-2.95	0.013789
DNLZ	DNL-type zinc finger	-2.96	0.000696
GIMAP7	GTPase, IMAP family member 7	-2.97	0.002201
ATP8B4	ATPase, class I, type 8B, member 4	-3.02	0.001027
SUCNR1	succinate receptor 1	-3.08	0.000456
PRAM1	PML-RARA regulated adaptor molecule 1	-3.08	0.019985
EPB41L3	erythrocyte membrane protein band 4.1-like 3	-3.10	0.011381
CD200	CD200 molecule	-3.14	0.000074
NEURL1B	neuralized homolog 1B (Drosophila)	-3.15	0.000005
TAF15	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa	-3.17	0.001031
MN1	meningioma (disrupted in balanced translocation) 1	-3.17	0.002195
SLC22A16	solute carrier family 22 (organic cation/carnitine transporter), member 16	-3.18	0.000327
GOLGA9P	golgi autoantigen, golgin subfamily a, 9 pseudogene	-3.18	0.001387
SLC16A14	solute carrier family 16, member 14 (monocarboxylic acid transporter 14)	-3.22	0.000915
KBTBD11	kelch repeat and BTB (POZ) domain containing 11	-3.27	0.000051
PLEKHO1	pleckstrin homology domain containing, family O member 1	-3.30	0.000034
CXCR7	chemokine (C-X-C motif) receptor 7	-3.31	0.000253
C5orf23	chromosome 5 open reading frame 23	-3.31	0.007896
KIAA1274	KIAA1274	-3.31	0.000006
NPR3	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)	-3.33	0.000049
C12orf5	chromosome 12 open reading frame 5	-3.37	0.004099
GLIPR1	GLI pathogenesis-related 1	-3.38	0.002403
CXorf21	chromosome X open reading frame 21	-3.39	0.000638
CTHRC1	collagen triple helix repeat containing 1	-3.41	0.000054
ENSG00000167912	NA	-3.43	0.000122
TMEM200A	transmembrane protein 200A	-3.46	0.000652
DUSP6	dual specificity phosphatase 6	-3.47	0.001800

Gene	Gene Name	FC	P value
IFI44	interferon-induced protein 44	-3.48	0.004774
CTSG	cathepsin G	-3.48	0.002935
MIRHG2	microRNA host gene 2 (non-protein coding)	-3.52	0.000059
SCHIP1	schwannomin interacting protein 1	-3.58	0.000015
HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	-3.59	0.000648
TFEC	transcription factor EC	-3.63	0.000013
SORL1	sortilin-related receptor, L(DLR class) A repeats-containing	-3.72	0.004780
LOC284422	similar to HSPC323	-3.78	0.000569
MPO	myeloperoxidase	-3.84	0.024036
MT-ND6	mitochondrially encoded NADH dehydrogenase 6	-3.87	0.000014
HLF	hepatic leukemia factor	-3.88	0.002206
MYO5C	myosin VC	-3.89	0.000000
LY75	lymphocyte antigen 75	-3.94	0.000002
SELL	selectin L	-3.97	0.000034
ARMCX2	armadillo repeat containing, X-linked 2	-3.97	0.000015
TNFSF13B	tumor necrosis factor (ligand) superfamily, member 13b	-3.98	0.000784
AIM1	absent in melanoma 1	-4.02	0.014656
PTPRCAP	protein tyrosine phosphatase, receptor type, C-associated protein	-4.04	0.001413
RGL4	ral guanine nucleotide dissociation stimulator-like 4	-4.12	0.000677
CTSZ	cathepsin Z	-4.18	0.000173
BAALC	brain and acute leukemia, cytoplasmic	-4.49	0.000148
CHRDL1	chordin-like 1	-4.58	0.000011
UBTD2	ubiquitin domain containing 2	-4.59	0.000104
PROM1	prominin 1	-4.82	0.010952
C1QTNF4	C1q and tumor necrosis factor related protein 4	-4.96	0.001387
NOG	noggin	-5.06	0.001137
SPINK2	serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)	-6.73	0.000456

6.3. Differentially expressed genes in myeloid commitment.

Gene	Gene Name	FC	P value
S100A8	S100 calcium binding protein A8	6.91	0.000007
PRTN3	proteinase 3	5.78	0.000368
ELANE	elastase, neutrophil expressed	5.41	0.006886
CLC	Charcot-Leyden crystal protein	4.67	0.018955
AZU1	azurocidin 1	4.57	0.000465
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	3.94	0.003005
CD24	CD24 molecule	3.91	0.000881
S100A9	S100 calcium binding protein A9	3.88	0.001064
ALOX5	arachidonate 5-lipoxygenase	3.85	0.011731
VCAN	versican	3.77	0.001689
SLPI	secretory leukocyte peptidase inhibitor	3.71	0.001067
HBB	hemoglobin, beta	3.68	0.005643
C19orf59	chromosome 19 open reading frame 59	3.58	0.000217
PRG2	proteoglycan 2, bone marrow (natural killer cell activator, eosinophil granule major basic protein)	3.54	0.015258
MNDA	myeloid cell nuclear differentiation antigen	3.49	0.000992
MS4A6A	membrane-spanning 4-domains, subfamily A, member 6A	3.43	0.003291
SERPINB10	serpin peptidase inhibitor, clade B (ovalbumin), member 10	3.43	0.000376
CSTA	cystatin A (stefin A)	3.36	0.000171
CST7	cystatin F (leukocystatin)	3.35	0.000170
STON1	stonin 1	3.34	0.000823
HP	haptoglobin	3.34	0.000556

Gene	Gene Name	FC	P value
C5orf20	chromosome 5 open reading frame 20	3.30	0.000751
MS4A3	membrane-spanning 4-domains, subfamily A, member 3 (hematopoietic cell-specific)	3.20	0.007733
CFD	complement factor D (adipsin)	3.07	0.004227
S100P	S100 calcium binding protein P	3.03	0.006748
NCF2	neutrophil cytosolic factor 2	3.00	0.002454
RETN	resistin	2.94	0.001154
HCK	hemopoietic cell kinase	2.89	0.000786
CD1D	CD1d molecule	2.81	0.005186
RAB20	RAB20, member RAS oncogene family	2.80	0.000007
PIWIL4	piwi-like 4 (Drosophila)	2.79	0.002774
ENPP2	ectonucleotide pyrophosphatase/phosphodiesterase 2	2.72	0.002343
P2RY2	purinergic receptor P2Y, G-protein coupled, 2	2.69	0.002087
LOC283663	hypothetical LOC283663	2.69	0.007697
LYZ	lysozyme (renal amyloidosis)	2.62	0.002217
ADAMDEC1	ADAM-like, decysin 1	2.61	0.011886
CNRIP1	cannabinoid receptor interacting protein 1	2.60	0.020620
ANXA3	annexin A3	2.60	0.000589
PARP8	poly (ADP-ribose) polymerase family, member 8	2.54	0.000548
HAL	histidine ammonia-lyase	2.53	0.000282
LIN7A	lin-7 homolog A (C. elegans)	2.50	0.003810
EGID-79948	plasticity-related gene 2	2.50	0.000003
CTSG	cathepsin G	2.47	0.000223
ELOVL3	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 3	2.46	0.009838
S100A12	S100 calcium binding protein A12	2.44	0.012070
SERPINB2	serpin peptidase inhibitor, clade B (ovalbumin), member 2	2.44	0.009607
P2RY13	purinergic receptor P2Y, G-protein coupled, 13	2.43	0.000658
CEACAM8	carcinoembryonic antigen-related cell adhesion molecule 8	2.43	0.012363
PLBD1	phospholipase B domain containing 1	2.39	0.029536
OLR1	oxidized low density lipoprotein (lectin-like) receptor 1	2.33	0.048440
CLEC12A	C-type lectin domain family 12, member A	2.33	0.003719
SLC22A15	solute carrier family 22, member 15	2.32	0.000013
CD14	CD14 molecule	2.31	0.003414
DYSF	dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)	2.28	0.001532
TGFBI	transforming growth factor, beta-induced, 68kDa	2.28	0.009135
ASGR2	asialoglycoprotein receptor 2	2.27	0.000117
NAPSB	napsin B aspartic peptidase pseudogene	2.27	0.000762
MPEG1	macrophage expressed 1	2.26	0.005274
LRG1	leucine-rich alpha-2-glycoprotein 1	2.24	0.004884
FCER1G	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide	2.22	0.002326
CD36	CD36 molecule (thrombospondin receptor)	2.22	0.000078
BEX1	brain expressed, X-linked 1	2.22	0.044141
IPCEF1	interaction protein for cytohesin exchange factors 1	2.21	0.007888
FZD2	frizzled homolog 2 (Drosophila)	2.21	0.021064
TCN1	transcobalamin I (vitamin B12 binding protein, R binder family)	2.20	0.049048
CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)	2.20	0.022493
C12orf59	chromosome 12 open reading frame 59	2.18	0.004136
RNASE2	ribonuclease, RNase A family, 2 (liver, eosinophil-derived neurotoxin)	2.16	0.000946
FGR	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	2.13	0.000647
CLU	clusterin	2.13	0.005949

Gene	Gene Name	FC	P value
GPR160	G protein-coupled receptor 160	2.12	0.002806
KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2	2.12	0.008066
ALOX5AP	arachidonate 5-lipoxygenase-activating protein	2.12	0.002995
LY86	lymphocyte antigen 86	2.11	0.003033
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	2.11	0.002477
TYROBP	TYRO protein tyrosine kinase binding protein	2.10	0.000550
ACPP	acid phosphatase, prostate	2.09	0.000879
ALAS1	aminolevulinate, delta-, synthase 1	2.06	0.000194
RNASE6	ribonuclease, RNase A family, k6	2.06	0.000505
FCN1	ficolin (collagen/fibrinogen domain containing) 1	2.02	0.010237
SEPP1	selenoprotein P, plasma, 1	2.01	0.000268
SLC47A1	solute carrier family 47, member 1	2.01	0.002740
PCOLCE2	procollagen C-endopeptidase enhancer 2	2.00	0.001528
ALOX12	arachidonate 12-lipoxygenase	2.00	0.004849
CILP2	cartilage intermediate layer protein 2	2.00	0.001953
CXCR7	chemokine (C-X-C motif) receptor 7	-2.02	0.003424
ASAP2	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2	-2.05	0.005281
TMEFF1	transmembrane protein with EGF-like and two follistatin-like domains 1	-2.09	0.000108
CALN1	calneuron 1	-2.10	0.000011
SMAGP	small trans-membrane and glycosylated protein	-2.10	0.000038
EMP1	epithelial membrane protein 1	-2.10	0.006898
FAM111B	family with sequence similarity 111, member B	-2.16	0.011623
RASGRP1	RAS guanyl releasing protein 1 (calcium and DAG-regulated)	-2.16	0.001808
PLCB4	phospholipase C, beta 4	-2.17	0.000004
GUCY1B3	guanylate cyclase 1, soluble, beta 3	-2.18	0.000195
SH3BP5	SH3-domain binding protein 5 (BTK-associated)	-2.19	0.009220
NPR3	natriuretic peptide receptor C/guanylate cyclase C (atrionatriuretic peptide receptor C)	-2.24	0.000253
GNAI1	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1	-2.27	0.000131
KIAA0125	KIAA0125	-2.27	0.000239
IL12RB2	interleukin 12 receptor, beta 2	-2.34	0.000333
CHRDL1	chordin-like 1	-2.34	0.001850
NAP1L3	nucleosome assembly protein 1-like 3	-2.36	0.001473
CD34	CD34 molecule	-2.38	0.000183
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	-2.41	0.000056
TNFSF4	tumor necrosis factor (ligand) superfamily, member 4	-2.43	0.000365
CTHRC1	collagen triple helix repeat containing 1	-2.43	0.000037
TGFB1I1	transforming growth factor beta 1 induced transcript 1	-2.46	0.004316
HTR1F	5-hydroxytryptamine (serotonin) receptor 1F	-2.47	0.000132
C5orf23	chromosome 5 open reading frame 23	-2.53	0.000016
RAI14	retinoic acid induced 14	-2.53	0.000382
GBP1	guanylate binding protein 1, interferon-inducible, 67kDa	-2.69	0.017717
HEMGN	hemogen	-2.74	0.000021
AKR1C3	aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II)	-2.79	0.010452
GZMA	granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)	-2.84	0.001536
IFI44	interferon-induced protein 44	-2.86	0.000719
CNN3	calponin 3, acidic	-2.97	0.000004
SPINK2	serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)	-3.00	0.000094
SLC16A14	solute carrier family 16, member 14 (monocarboxylic acid transporter 14)	-3.07	0.001477

Gene	Gene Name	FC	P value
MMRN1	multimerin 1	-3.09	0.000266
CRHBP	corticotropin releasing hormone binding protein	-3.25	0.000809
BEX2	brain expressed X-linked 2	-3.26	0.000961
HLF	hepatic leukemia factor	-3.30	0.002439
TMSB15A	thymosin beta 15a	-3.37	0.007020
ARMCX2	armadillo repeat containing, X-linked 2	-3.52	0.000017
TMEM200A	transmembrane protein 200A	-3.55	0.000282
DLK1	delta-like 1 homolog (Drosophila)	-3.96	0.000169

6.4. Lists of promoters

Complete lists of promoters identified using CAGE-seq are available at http://www.webcitation.org/6ZBVNp55l

6.5. Lists of active enhancers

Complete lists of active enhancers identified using ChIP-seq are available at http://www.webcitation.org/6ZBVNp55l

ACKNOWLEDGEMENTS

This work has been supported by FIRB (RBFR10OSG, “Analisi genome-wide dei promotori e degli enhancer utilizzati durante il differenziamento delle cellule staminali ematopoietiche”) and EPIGEN Flagship Project (“Mappatura degli elementi di regolazione trascrizionale attivi in cellule staminale somatiche umane”).

First of all, I would like to thank all the people that contributed to this project: Dr. Clelia Peano (ITB-CNR), Dr. Annarita Miccio (Università di Modena), Ms. Oriana Romano (Università di Modena), Dr. Guidantonio Malagoli Tagliazucchi (Università di Modena), Dr. Gianluca De Bellis (ITB-CNR), Prof. Silvio Bicciato (Università di Modena) and Prof. Fulvio Mavilio (Università di Modena). Nothing here would have been possible without them.

I would like to thank the University of Milan, the PhD School in Molecular Medicine and Prof. Cristina Battaglia for the great opportunity to start my PhD and for the training support.

I thank all the Genomics Group at Institute of Biomedical Technologies (ITB-CNR) of the National Research Council: Santosh Anand, Roberta Bordoni, Giada Caredda, Ingrid Cifola, Clarissa Consolandi, Gianluca De Bellis, Eleonora Mangano, Clelia Peano, Alessandro Pietrelli, Eva Pinatel, Simone Puccio, Ermanno Rizzi, Marco Severgnini for sharing a fundamental step of my education. Thank you everyone!

I thank the people at Center for Biomedical Informatics at Harvard Medical School, in particular Dr. Francesco Ferrari, Prof. Peter V. Kharchenko and Prof. Peter J. Park, for the invaluable opportunity to be a visiting PhD student there.

PUBLICATIONS

[1] Peano C, Wolf J, Demol J, Rossi E, **Petiti L**, De Bellis G, Geiselmann J, Egli T, Lacour S, Landini P. *Characterization of the Escherichia Coli σS Core Regulon by Chromatin Immunoprecipitation-Sequencing (ChIP-Seq) Analysis*. Nature Publishing Group 2015: 1–15.

[2] Poletti V, Delli Carri A, Malagoli Tagliazucchi G, Faedo A, **Petiti L**, Mazza EMC, Peano C, De Bellis G, Bicciato S, Miccio A, Cattaneo E, Mavilio F. *Genome-Wide Definition of Promoter and Enhancer Usage During Neural Induction of Human Embryonic Stem Cells*. PLoS ONE 2015; 10(5): e0126590.

[3] Delvillani F, Sciandrone B, Peano C, **Petiti L**, Berens C, Georgi C, Ferrara S, Bertoni G, Pasini ME, Dehò G, Briani F. *Tet-Trap, a Genetic Approach to the Identification of Bacterial RNA Thermometers: Application to Pseudomonas Aeruginosa*. Rna 2014; 20(12): 1963–1976.

[4] Peano C, Pietrelli A, Consolandi C, Rossi E, **Petiti L**, Tagliabue L, De Bellis G, Landini P. *An Efficient rRNA Removal Method for RNA Sequencing in GC-Rich Bacteria*. Microb Inform Exp 2013; 3(1): 1.

[5] Ghignone S, Salvioli A, Anca I, Lumini E, Ortu G, **Petiti L**, Cruveiller SEP, Bianciotto V, Pietro Piffanelli, Lanfranco L, Bonfante P. *The Genome of the Obligate Endobacterium of an AM Fungus Reveals an Interphylum Network of Nutritional Interactions*. The ISME Journal 2011; 6(1): 136–145.

POSTERS

[1] Pietrelli A, Previtali G, Semeraro R, **Petiti L**, Tattini L, Magi A, De Bellis G, Battaglia C, Cifola I. *DriveMe: a web framework for gene burden analysis on NGS data.* BITS annual meeting, University of Milan Bicocca, Milan, June 3-5, 2015.

[2] Miccio A, Peano C, Romano O, Malagoli Tagliazucchi G, **Petiti L**, Cifola I, Rizzi E, Severgnini M, Bicciato S, De Bellis G and Mavilio F. *Genome-wide analysis of promoters and enhancers usage in hematopoietic stem cell differentiation.* Epigenetics & Chromatin: Interactions and processes conference, Boston, MA, USA, March 11-13, 2013.

[3] Peano C, Miccio A, Romano O, Malagoli Tagliazucchi G, **Petiti L**, Cifola I, Rizzi E, Severgnini M, Bicciato S, De Bellis G and Mavilio F. *Defining the transcriptome and the epigenome of human hematopoietic stem/progenitor cells and their lineage-restricted progeny.* EPIGEN Project - NGS

Sequencing and Epigenomics Workshop, Bari, Italy, December 05-07, 2012.

[4] Ghignone S, Anca I, Friard O, **Petiti L**, Lumini E, Salvioli A, Bianciotto V, Piffanelli P, Lanfranco L and Bonfante P. *Genomics and comparative genomics of Candidatus Glomeribacter gigasporarum.* FEMS 2009 - 3rd Congress of European Microbiologist, Gothenburg, Sweden, June 28 - July 02, 2009.