Transcriptional and epigenetic mechanisms of gene regulation in cellular differentiation

The hematopoietic paradigm

Giorgio Lucio Papadopoulos

Department of Biology

University of Crete

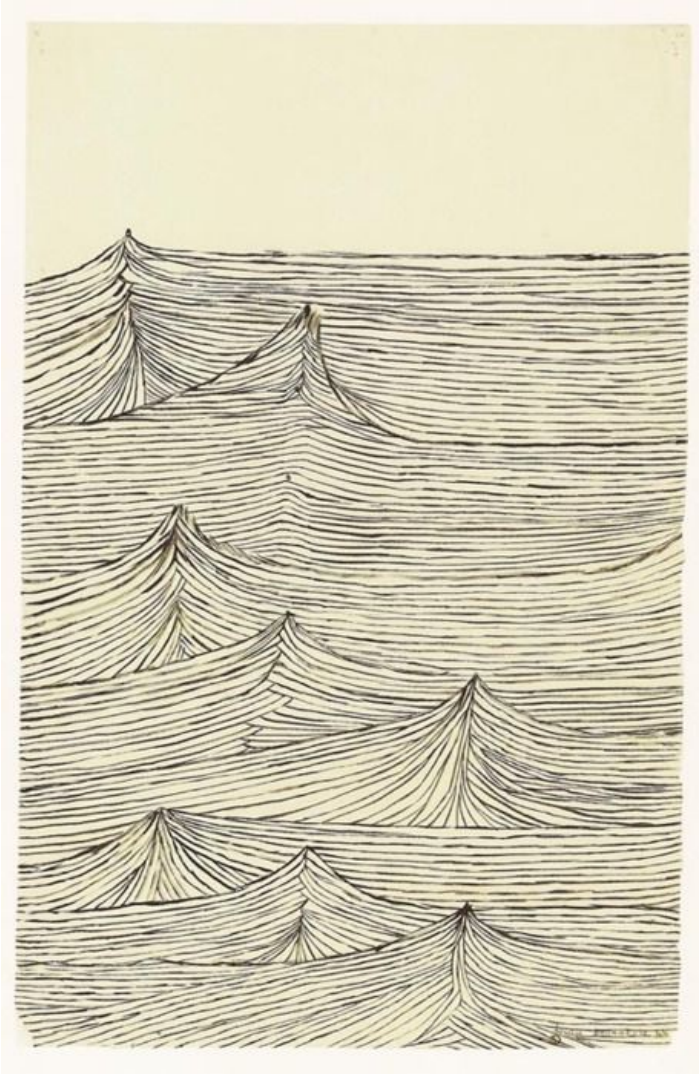
This dissertation is submitted for the degree of

Doctor of Philosophy

Scientific Supervisor:
Dr John Strouboulis
Academic Supervisor:
Prof George Garinis

July 2015

Dedicato a miei primi
insegnanti di biologia...



Louise Bourgeois, 'Waves'

Abstract

Cellular commitment and differentiation in multicellular organisms depend on the concerted action of transcription factors and epigenetic modifications in regulating differential patterns of gene expression. Understanding the molecular basis of such complex regulatory events has been greatly facilitated in recent years by the advent of next generation sequencing (NGS) technologies for the high throughput, high resolution, genome wide characterization of a multitude of transcriptional and epigenetic regulatory factors in various cellular and developmental models. Deciphering these data in extracting biological meaning has been a major challenge in the application of NGS technologies in gene regulation.

Our main research interest is to elucidate the transcriptional regulatory events underlying hematopoiesis by specifically focusing on red blood cell differentiation (erythropoiesis). To these ends, the work described here entails the development of a computational approach in analyzing and integrating a large number of comprehensive NGS datasets of multiple genomic characteristics (transcription factor binding, epigenetic modifications etc.) in murine and human hematopoiesis. Our computational analysis relies on the combination of supervised (RandomForest regression modeling) and unsupervised (hierarchical clustering) machine learning approaches, in producing highly structured gene wide distribution patterns of chromatin features in different hematopoietic cell populations.

We fist applied this approach in characterizing the genome-wide occupancy profiles of the master erythroid transcription factor GATA1 which we obtained in mouse fetal liver erythropoiesis (Papadopoulos et al., 2013). Integration of GATA1 occupancy profiles with available genome-wide transcription factor and epigenetic profiles in fetal liver erythroid cells, showed that GATA1 binding preferentially associates with specific epigenetic modifications, such as H4K16Ac and H3K27Ac or H3K4me2. Furthermore, we were able to classify GATA1 target genes into three distinct clusters, each associated with specific epigenetic signatures and functional characteristics, thus suggesting distinct GATA1 associated regulatory mechanisms.

Next, we applied our computational approach in utilizing available genomic data to investigate the differential transcriptional and epigenetic events underlying the specification of the erythroid and megakaryocytic lineages, deriving from a common progenitor. We identified a large group (~1000) of genes with active promoter marks in hematopoietic

stem cell (LSK cells), which become specifically inactive in erythroid cells but not in megakaryocytes. Comparison of DNase hypersensitivity profiles available for all erythroid differentiation stages, indicated that inactivation of these promoters initiates before the stage of early erythroid commitment ($CD71^{+}$/$Ter119^{-}$ cells), thus representing an early step of the erythroid specification process. By comparing expression profiles of erythroid-megakaryocytic progenitors (MEPs), erythroid cells and megakaryocytes, we also identified erythroid specific epigenetic modifiers that may serve as candidates in regulating erasure of this epigenetic signature in erythroid cells.

We also focused on the genome wide occupancies of transcription factors with essential functions in both erythroid and megakaryocytic differentiation, such as GATA1, GATA2, TAL1 and LDB1. By analyzing genome wide occupancies in LSK (HSCs), $Ter119^{+}$ (erythroid) and $CD41^{+}$ (megakaryocytic) cells, we found, firstly, that GATA1 binding patterns in erythroid and megakaryocytic cells appear to be largely distinct. Secondly, we found that the GATA1 erythroid specific binding profile is closely reflected by the TAL1 and LDB1 binding profiles in LSK cells, thus showing upstream specification of erythroid GATA1 binding by TAL1/LDB1.

Finally, we developed Ariadne (aegeas.imbb.forth.gr/Ariadne/) as a web based comprehensive tool to compare gene-wide relational profiles of multiple NGS datasets analyzed using our computational approach and in order to visualize primary sequencing data within single gene loci.

Table of contents

INTRODUCTION

Development and differentiation

Temporal and spatial control of gene expression allows for differential cellular identity and specialization. This becomes crucial in multicellular organisms given the fact that all cells share the same genetic material. Several lines of evidence show that this differential gene expression leads to a decreasing differentiation potential that peaks at the zygote stage (totipotent cells) and becomes 'extinguished' in fully differentiated (highly specialized) cells (Figure 1), (Waddington, 1957). Higher eukaryotes ontogeny is organized in distinct developmental and differentiation 'ancestry based' trees, composed of distinct lineages (branches) that terminate with the cycling production of fully specialized cells (leaves). Interestingly, differentiation pathways do not follow a simple, linear path for the production of each specialized cell but present with 'functional' ramifications, stemming from the production of intermediate, multipotent, progenitor cells that will give rise to several, functionally related but distinct, cell populations.

Hematopoiesis

Hematopoietic stem cells (HSCs) represent an excellent example of functional ramification of the ontogeny tree. In fact, HSCs represent a discrete cell population, characterized by the ability to carry out all terminal differentiation programs of mature blood tissue. More specifically, HSCs reside as rare cells in the bone marrow in adult mammals and sit atop a hierarchy of progenitors that become progressively restricted to several or single lineages (Orkin, 2000). As with all other stem cells, HSCs are also capable of self-renewal, a process that ensures the production of additional HSCs, thus enabling their replenishment in the adult hematopoietic tissue.

In mammals, the sequential sites of blood production include the yolk sac ('primitive' hematopoiesis), an area surrounding the dorsal aorta termed the aorta-gonad mesonephros (AGM) region, the fetal liver, and finally the bone marrow ('definitive' hematopoiesis) (Figure

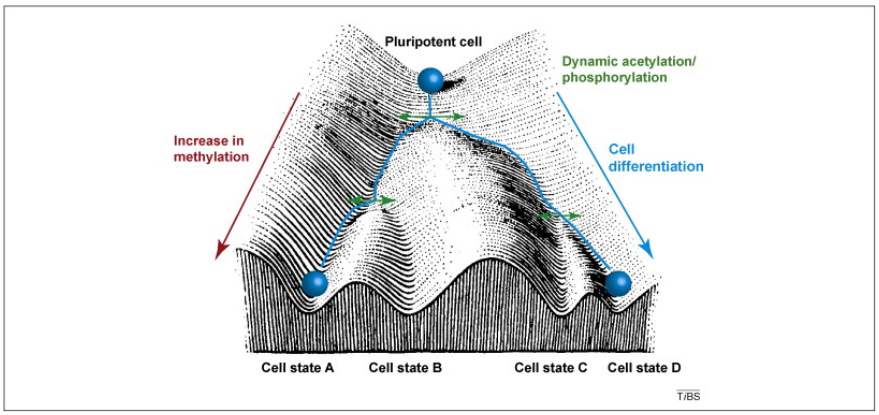


Figure 1: Integrating different histone modification types in Waddington's Epigenetic Landscape. The model presented by Conrad Waddington shows how a cell becomes more and more determined during development and that the possibility for differentiation decreases later in development. He compared the development of a cell with a ball (depicted in blue) rolling down the illustrated landscape and making its way through different valleys and elevations to different end points. Different types of histone modifications are incorporated in this model, arguing that they can influence how cells develop. Dynamic histone modifications such as acetylation and phosphorylation can change the fate of the cell to a small degree at a certain timepoint (green arrows), whereas stable lysine methylations accumulate during development and have more influence on final destiny of a cell (red arrow). (Barth and Imhof, 2010)

2). The primary function for primitive hematopoiesis is the production of red blood cells that facilitate tissue oxygenation of the embryo and is rapidly replaced by adult-type definitive hematopoiesis. Definitive hematopoiesis involves the colonization of the fetal liver, thymus, spleen, and ultimately the bone marrow. Importantly, none of these sites is accompanied by *de novo* HSC generation. Rather, their niches support expansion of populations of HSCs that migrate to these new sites. However, until recently, there has been no evidence by fate mapping or direct visualization that HSCs from one site colonize subsequent sites.

Stem cells depend on their microenvironment, the niche, for regulation of self-renewal and differentiation. In fact, the properties of HSCs in each subsequent hematopoietic site differ, presumably reflecting diverse niches that support HSC expansion and/or differentiation and intrinsic characteristics of HSCs at each stage. For instance, HSCs present in the fetal liver are in cycle, whereas adult bone marrow HSCs are largely quiescent. How niches modulate self-renewal is a challenge for future studies.

Remarkably, the site of hematopoiesis is not conserved in vertebrate evolution. For instance, the site of adult hematopoiesis in fish is the kidney, in frogs adult blood is formed in the liver whereas birds and mammals form blood in the marrow. Remarkably, in the frog *Rana temporaria*, the site of hematopoiesis switches between the liver and bone marrow depending on the season (Maslova and Tavrovskaia, 1993). Despite the differences observed in the

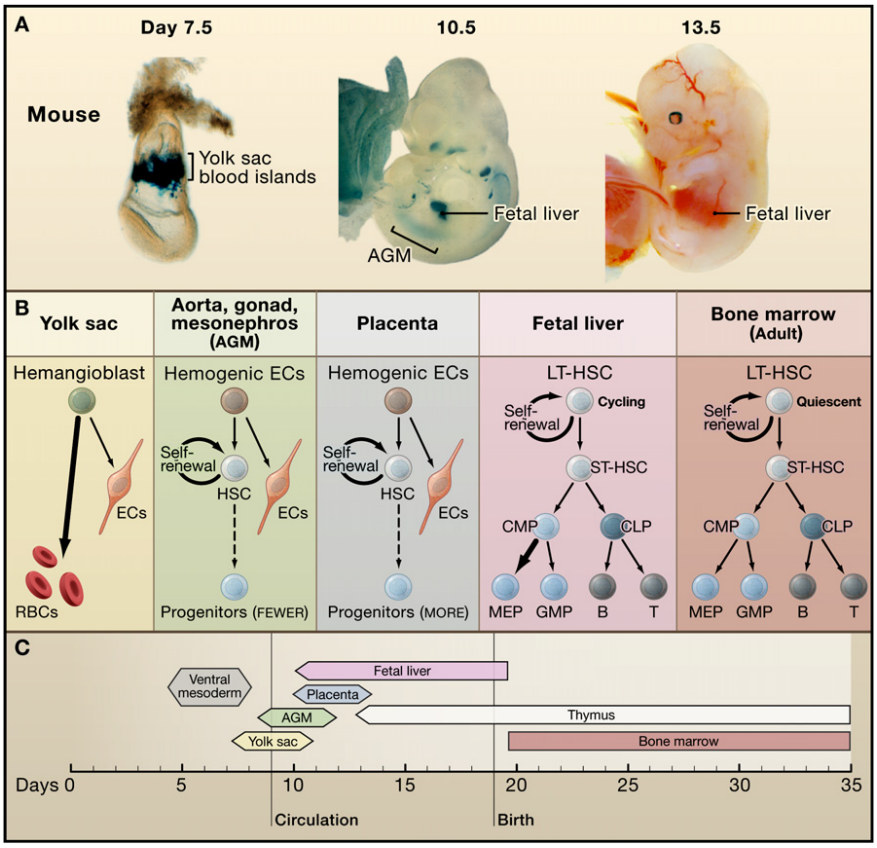


Figure 2: (A) Hematopoiesis occurs first in the yolk sac (YS) blood islands and later at the aorta-gonad mesonephros (AGM) region, placenta, and fetal liver (FL). YS blood islands are visualized by LacZ staining of transgenic embryo expression GATA1-driven LacZ. AGM and FL are stained by LacZ in Runx1-LacZ knockin mice. (Photos courtesy of Y. Fujiwara and T. North.) (B) Hematopoiesis in each location favors the production of specific blood lineages. Abbreviations: ECs, endothelial cells; RBCs, red blood cells; LT-HSC, long-term hematopoietic stem cell; ST-HSC, short-term hematopoietic stem cell; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte/macrophage progenitor. (C) Developmental time windows for shifting sites of hematopoiesis. (Orkin and Zon, 2008)

hematopoietic tissues, the hematopoietic process itself is generally conserved throughout vertebrate evolution, thus, manipulation of animal models, such as the mouse and zebrafish, can and has been used to complement and greatly extend our understanding of human hematopoiesis.

Lineage commitment of hematopoietic stem cells

All blood cell lineages derive from a common hematopoietic stem cell (HSC) responsible for life-long and balanced blood cell production, in man amounting to millions of cells per second in steady state (Ogawa, 1993). There are at least eight types of blood cells, varying

in their appearance and function (Figure 3). The most abundant cells in the blood are the red blood cells or erythrocytes, occupying 45% of its volume, whereas the rest of cell types -white blood cells- occupy about 1% of the blood volume and include the platelets, the granulocytes (neutrophils, eosinophils and basophils), the monocytes and the lymphocytes.

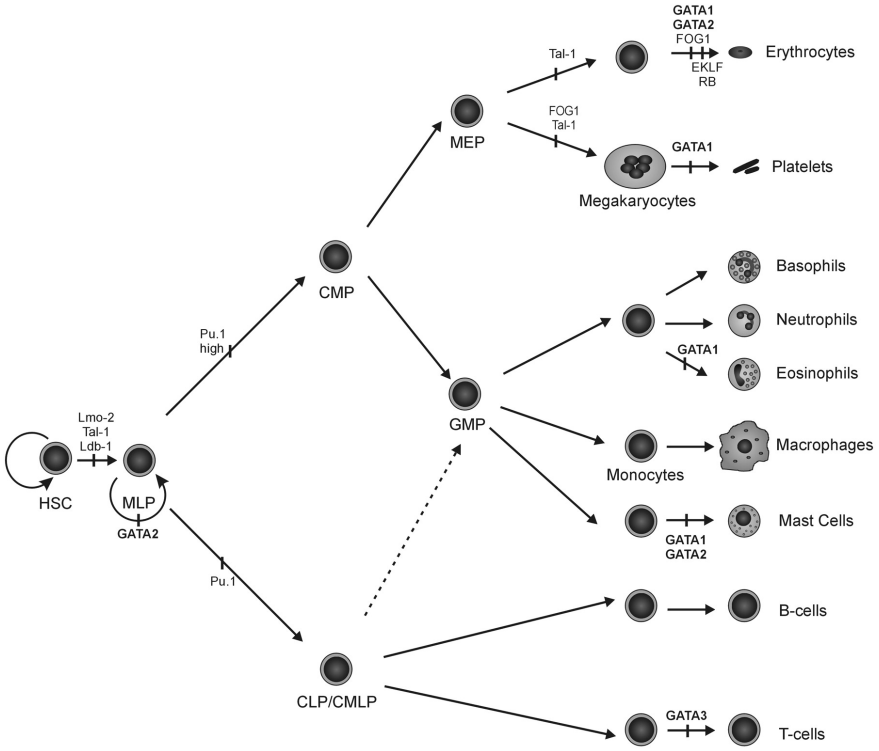


Figure 3: Schematic representation of the main lineage commitment steps in hematopoiesis. The hematopoietic stem cell (HSC) is the basis of the hematopoietic hierarchy and gives rise to multilineage progenitors (MLP), which can differentiate into all the hematopoietic lineages. MLPs become lineage restricted to the lymphoid and myeloid lineages in the common lymphoid progenitor (CLP) and common myeloid progenitor (CMP), respectively. CLPs can give rise exclusively to B and T cells, while CMPs can give rise to megakaryocyte-erythrocyte progenitors (MEP) and granulocyte-monocyte progenitors (GMP). Alternatively, it is also believed that the first lineage commitment separates myeloid and erythroid potential, in the CMP, from myeloid-lymphoid potential, in the common myeloid lymphoid progenitor (CMLP). CMLPs can than further differentiate in B cells, T cells, and GMPs (dashed line). Hematopoietic GATA factors and GATA1 cofactors relevant for the development of particular hematopoietic lineages are indicated. (Ferreira et al., 2005)

Virtually all HSC activities in adult mouse bone marrow (BM) have been shown to reside in the small Lin$^-$SCA-1$^+$KIT^{hi} (LSK) HSC compartment (0.1% of all BM cells; (Ikuta and Weissman, 1992; Li and Johnson, 1995; Spangrude et al., 1988; Weissman et al., 2001)). A number of improved assays for evaluating lineage potential, as well as tools for candidate progenitor purification, such as fluorescence activated cell sorting (FACS), have been developed. These have facilitated the identification, prospective purification

and characterization of distinct progenitors (Reya et al., 2001), such as common myeloid progenitors (CMPs) (Akashi et al., 2000) and common lymphoid progenitors (CLPs) (Kondo et al., 1997), which lack lymphoid and myeloid lineage potentials, respectively. CLPs can give rise exclusively to B and T cells, while CMPs can give rise to megakaryocyte-erythrocyte progenitors (MEP) and granulocyte-monocyte progenitors (GMP) (Ferreira et al., 2005). These findings were instrumental in establishing the 'classical' or CMP/CLP commitment model, which proposes that the first lineage restriction or branching point from an HSC might result in strict separation of subsequent myeloid and lymphoid development (Reya et al., 2001). Importantly, the lineage potentials attributed to CMPs and CLPs were also confirmed at the single cell level (Akashi et al., 2000; Kondo et al., 1997). Subsequent gene expression analyses showed that CMPs coexpress granulocyte/macrophage (GM) and megakaryocyte/erythroid (MegE) but not lymphoid affiliated genes, whereas CLPs coexpress B and T lymphoid but not myeloid associated genes (Mikkola et al., 2003), providing further validation to the early myeloid/lymphoid separation.

Interestingly, the degree to which the identified CMPs and CLPs represent obligatory intermediates for myeloid and lymphoid development in adult hematopoiesis has not been established. More recent studies, have revealed significant heterogeneity in the LSK multipotent progenitors (MPP) compartment. These studies suggested the existence within the adult mouse of an LSK MPP compartment of lymphoid-primed MPPs (LMPPs) or common myeloid-lymphoid progenitors (CMLPs) with combined granulocyte/macrophage and B and T lymphoid potentials, but little or no megakaryocyte/erythroid potential. Most notably multiplex single cell PCR analysis confirmed the combined GM-MegE priming of single HSCs and, in striking contrast, GM-lymphoid priming of single LMPPs, with an almost mutual exclusion of the MegE and lymphoid programmes throughout the HSC hierarchy (Adolfsson et al., 2005).

In the megakaryocyte and erythroid lineages, the final stage of noncommitted progenitor cell type is thought to be a common megakaryocyte-erythroid progenitor (MEP). Progenitor cells, such as the MEP, are exceedingly rare and difficult to isolate because of their transient nature, although several lines of evidence support the existence of such a cell type in vivo. MEPs were firstly identified as a human $CD34^{+}CD38^{low}$ cell population that was capable of giving rise to colonies that contain both erythroid and megakaryocytic cells (Nurden et al., 2006). Subsequent studies isolated cell populations capable of generating cells of either megakaryocytic or erythroid phenotype in single-cell differentiation experiments from mouse bone marrow (Akashi et al., 2000) or spleen tissues (Vannucchi et al., 2001). Furthermore, Adolfsson et al (2005) proposed the specification of the MEPs as one of the earliest branch points during hematopoietic differentiation.

Erythroid Differentiation

Through their oxygen delivery function, red blood cells are pivotal to the healthy existence of all vertebrate organisms. These cells are required during all stages of life — embryonic, fetal, neonatal, adolescent, and adult. In the adult, red blood cells are the terminally differentiated end-product cells of a complex hierarchy of hematopoietic progenitors that become progressively restricted to the erythroid lineage. During this stepwise differentiation process, erythroid progenitors undergo enormous expansion, so as to fulfill the daily requirement of $2x10^{11}$ new erythrocytes.

As mentioned above, erythrocytes represent the most common cell type in adult blood. Human blood contains $5x10^{6}$ erythrocytes per microliter; these cells have an average life span of 120 d, hence the great demand for new erythrocytes being constantly produced in the bone marrow. A series of intermediate erythroid precursors can be recognized that progressively

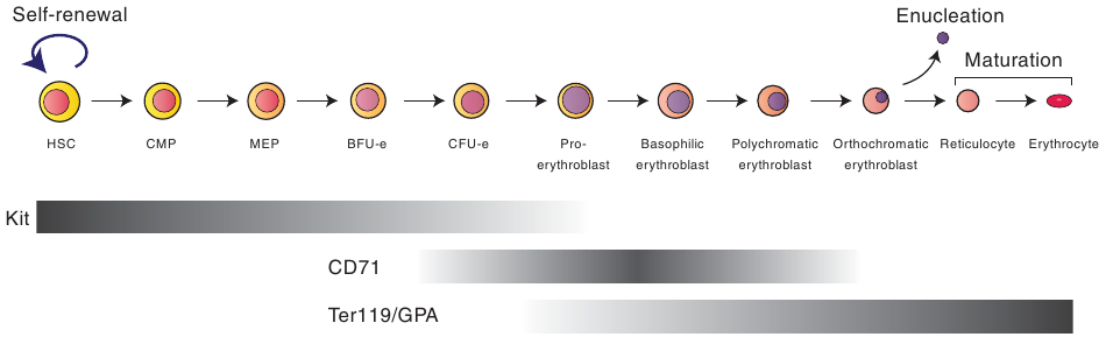


Figure 4: The expression of the most commonly used cell surface markers to identify the various stages is indicated by the bars. Gray, low expression; black, high expression; HSC, hematopoietic stem cell; CMP, common myeloid progenitor; MEP, megakaryocyte-erythroid progenitor; BFU-e, burst-forming unit, erythroid; CFU-e, colony-forming unit, erythroid. (Dzierzak and Philipsen, 2013)

gain erythroid characteristics (Figure 4). Traditionally, erythropoiesis has been divided into 3 stages: early erythropoiesis, terminal erythroid differentiation, and reticulocyte maturation (Dzierzak and Philipsen, 2013).

Early erythroid-restricted progenitors are identified as the burst forming unit-erythroid (BFU-E), so-called because its earliest progeny are motile, giving rise to a multi-subunit colony (or burst). More mature erythroid progenitors, colony-forming units-erythroid (CFU-E), consist of small colonies of 16 – 125 cells that appear after 2 – 3 d (mouse) or 5– 8 d (human) in methylcellulose culture. They are fivefold to eightfold more abundant than BFU-E in bone marrow and under normal circumstances do not appear in the circulation.

Terminal erythroid differentiation begins with proerythroblasts differentiating into basophilic, then polychromatic, then orthochromatic erythroblasts that enucleate to become

reticulocytes. Numerous changes occur during terminal erythroid differentiation. Erythroblasts decrease in size, activate hemoglobin synthesis, undergo extensive membrane reorganization (resulting in the characteristic discoid shape of erythrocytes), lose intracellular organelles and initiate a chromatin condensation process prior to enucleate (Hattangadi et al., 2011; Wong et al., 2011).

As progenitors undergo the differentiation process, their numbers increase, with their proliferative potential simultaneously decreasing. Under steady-state conditions, 1% of the erythrocytes are cleared every day and replaced by new cells. To maintain the red blood cell count in the 5 liters of blood of an adultindividual, $2.4x10^6$ new erythrocytes have to be produced each second.

Megakaryocytic Differentiation

The average platelet count in humans ranges from 150 to $350x10^9$/L, with some diurnal variation, although the level for any individual is maintained within fairly narrow limits when adjusted for time of day. The range of tolerable platelet counts is broad, but once platelet counts fall below $50x10^9$/L, the risk of pathological bleeding rises substantially.

As the second most abundant cell in the blood, their primary role is to maintain hemostasis and instigate wound healing on vascular damage. An adequate supply of platelets is essential to repair the minor vascular damage that occurs with daily life, and to initiate thrombus formation in the event of overt vascular injury. Platelets also play a critical role in cardiovascular disease (Yousuf and Bhatt, 2011), wound repair (Nurden, 2011), the innate immune response (Semple et al., 2011), and the biology of metastatic cancer (Gay and Felding-Habermann, 2011).

Once derived from a bi-potent erythroid–megakaryocyte progenitor cell, the megakaryoblast undergoes a series of divisions, during which time the cytoplasm begins to express platelet-specific proteins (e.g., β1-tubulin), the cell surface membrane becomes decorated by a number of platelet-specific adhesive proteins (e.g., integrin αIIb/β3, integrin αII/β1, glycoprotein GpIb, GpVI), cytoplasmic granules and their constituents (e.g., platelet factor 4, transforming growth factor β1, von Willebrand factor, P-selectin) assemble, and internal membranes (specialized for rapid calcium flux or for proplatelet formation) begin to form. After four to six cycles of cell division, mitosis begins to abort in anaphase. As DNA synthesis continues despite aborted mitosis, a process termed endomitosis, the megakaryocyte becomes highly polyploid. During the endomitotic phase of the megakaryocyte life cycle, gene transcription becomes synchronized on all copies of the platelet-specific structural genes, resulting in massive translation of critical platelet proteins, required for the impressive growth in cell volume during this phase of megakaryocyte development. The result is an

extremely large mature megakaryocyte that contains 64, 128, or even 256 times the normal chromosome complement. At this point, exvaginations of internal membranes form, driven by a breakdown of the circumferential actin cytoskeleton and projection of long filaments of β1-tubulin; these proplatelet processes then branch and fragment into platelets.

Lineage Specific Transcription Factors

As intrinsic determinants of cellular phenotype, transcription factors (TFs) provide an entry point for unraveling how HSCs develop during embryogenesis and how lineage-restricted differentiation is programmed (Orkin, 2000). The basic-helix-loop-helix (bHLH) factor SCL/Tal-1 is essential for development of both the primitive and definitive hematopoietic systems, since no blood cells are generated in its absence (Kim et al., 2007). Despite the fact that SCL/Tal1 is an obligate factor for hematopoietic fate specification during development, conditional inactivation in adult HSCs has surprisingly little consequence on maintenance or self-renewal of HSCs and multipotent progenitors (Mikkola et al., 2003). Under these circumstances, the role of this factor in maturation of erythroid and megakaryocytic cells is revealed.

Going downstream the hematopoietic tree (committed progenitors) cell-surface phenotypes (defined by monoclonal antibodies) and the subset of hematopoietic TFs expressed in these cells can be largely conveniently matched. For instance megakaryocytic/erythroid progenitors (called MEPs) that give rise to megakaryocyte and red blood cell precursors highly express the 'erythroid factor' GATA1, whereas a 'myeloid factor', such as C/EBPa, is present in granulocyte/macrophage progenitors (GMPs).

However, this relationship breaks down at earlier stages in the hierarchy. A simple one-to-one correspondence of lineage-restricted TFs and progenitors is challenged by the fact that earlier multipotential progenitors and HSCs express markers of disparate lineages even within single cells, albeit generally at low levels (Orkin, 2003). This phenomenon, termed lineage priming, suggests that the fate of these immature cells is not sealed and that lineage selection is largely a process in which alternative possibilities (differentiation potential) are extinguished rather than one in which new programs are imposed on an otherwise blank slate. Lineage priming may be an efficient means by which chromatin invested in important hematopoietic programs is maintained in an available or open configuration in HSCs, thus, maintaining the inherent plasticity of multipotential progenitors. Principal hematopoietic regulators (lineage-restricted TFs) are endowed with the complementary tasks of promoting their own lineage differentiation while simultaneously acting against factors favoring other

choices. This combination of positive and antagonistic roles between major regulators provides an efficient means for lineage commitment and differentiation.

Numerous examples of this principle of lineage programming have been described. One of the best described examples is the antagonistic action of GATA1 and PU.1 in promoting erythroid/megakaryocytic/eosinophil and myeloid differentiation, respectively (Figure 5). In fact, it has been shown that inhibition of GATA1 expression shifts hematopoietic progenitors to a myeloid fate, whereas the inhibition of PU.1 expression shift hematopoietic progenitors towards lymphoid fate (Galloway et al., 2005; Rhodes et al., 2005). Other examples of direct antagonism by hematopoietic transcription factors include the relative relationships of C/EBP and FOG1 with respect to eosinophil and multipotential cell fates (Querfurth et al., 2000), KLF1 (EKLF) and FLI1 for erythroid and megakaryocytic choice (Starck et al., 2003) and others.

GATA factors are illustrative of mechanistic regulatory models by which transcription factors directly interact within protein complexes (Kim and Bresnick, 2007). More specifically, GATA1 (or its close relative, GATA2) are part of a larger multimeric protein complex that includes SCL/Tal1, LMO2 and Ldb1. The functional significance of this complexes during cell fate decisions are illustrated by the fact that knockout of either LMO2 or SCL/Tal1 leads to the absence of any hematopoietic progenitors in the early embryo. Moreover, forced expression of GATA1/2, SCL, and LMO2 efficiently converts *Xenopus* mesoderm to a hematopoietic fate. Remarkably, although the GATA/SCL/LMO2/Lbd1 complex recognizes a composite DNA-binding site, later studies demonstrated that the DNA-binding activity of SCL/Tal1 (in a heterodimer with E2A) is dispensable for hematopoietic specification but required for full erythroid and megakaryocytic cell maturation (Porcher et al., 1999).

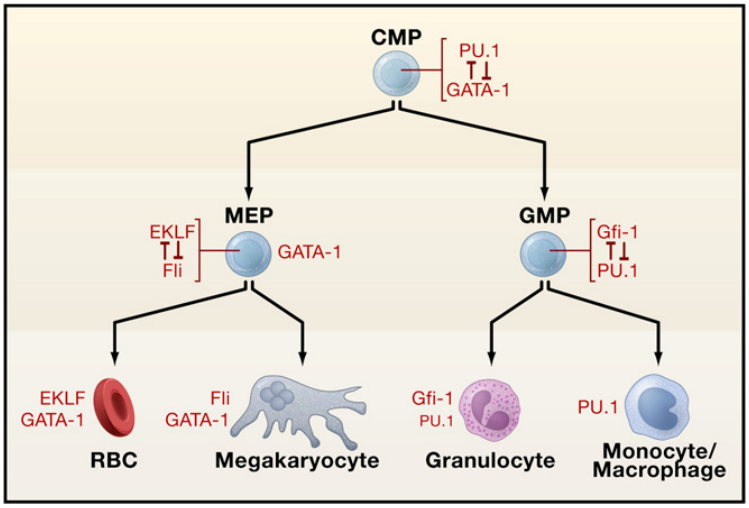


Figure 5: Examples of antagonism are depicted in red. The transcription factors present in the mature precursors following choice of specific lineage are shown at the bottom in black. Abbreviations: CMP, common myeloid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte/macrophage progenitor; RBCs, red blood cells. (Orkin and Zon, 2008)

Going Genome-wide

As discussed above, several transcriptional regulators and molecular triggers of the hematopoietic process have been discovered. However, the genome-wide regulatory events that determine cellular commitment and differentiation processes remain poorly understood. TF occupancy to DNA ultimately controls the activity of enzymes that catalyze production of messenger RNA. This complex process entails TF-mediated recruitment of a variety of non-RNA polymerase enzymes to specific sites within the genome, where they mediate covalent modifications of DNA bases or DNA-bound proteins, in particular the tails of histone proteins. Recent development of Next Generation Sequencing (NGS) allowed for an unprecedented view of protein-DNA interactions, mRNA quantitation and epigenetic features in a genome-wide manner. More specifically, Chromatin ImmunoPrecipitation followed by massive parallel sequencing (ChIPseq) is a technique for genome-wide profiling of DNA-binding proteins and epigenetic modifications. Importantly, ChIPseq offers higher resolution, less noise and greater coverage than its array-based predecessor ChIP–chip, becoming an invaluable tool for studying gene regulation and epigenetic mechanisms. On the other hand, ChIPseq experiments generate large quantities of data, and effective computational analysis is becoming crucial for uncovering biological mechanisms and indispensable for biological interpretation of the data. As a result, the main question raised by current approaches to understanding gene regulatory mechanisms during cell fate decisions shifted from "Is it possible to generate genome-scale information?" to "What can we actually learn from genome-scale datasets?"

Hematopoiesis represents one of the most experimentally tractable mammalian organ systems and, therefore, has historically tended to be at the forefront of applying new technologies within biomedical research. As a result, several major erythroid TFs have been recently analyzed by ChIPseq. The first such factor is the archetypal erythroid TF GATA1, resulting in the production of several genome-wide binding profiles, reported by several groups in both mouse and human primary cells and cell lines (Cheng et al., 2009; Fujiwara et al., 2009; Papadopoulos et al., 2013; Soler et al., 2010; Wontakal et al., 2012; Yu et al., 2009).

GATA1

GATA1 (also known as Eryf-1, NF-E1, NF-1 and GF-1) is a critical transcription factor for erythroid differentiation. It is the founding member of the GATA family of proteins, which consists of six transcription factors, GATA1 to GATA6 (Lowry and Mackay, 2006; Morceau et al., 2004). They all bind to the DNA consensus sequence (A/T)GATA(A/G) by two characteristic C4 (Cys-X2-Cys-X17-Cys-X2-Cys) zinc-finger motifs, referred to as the

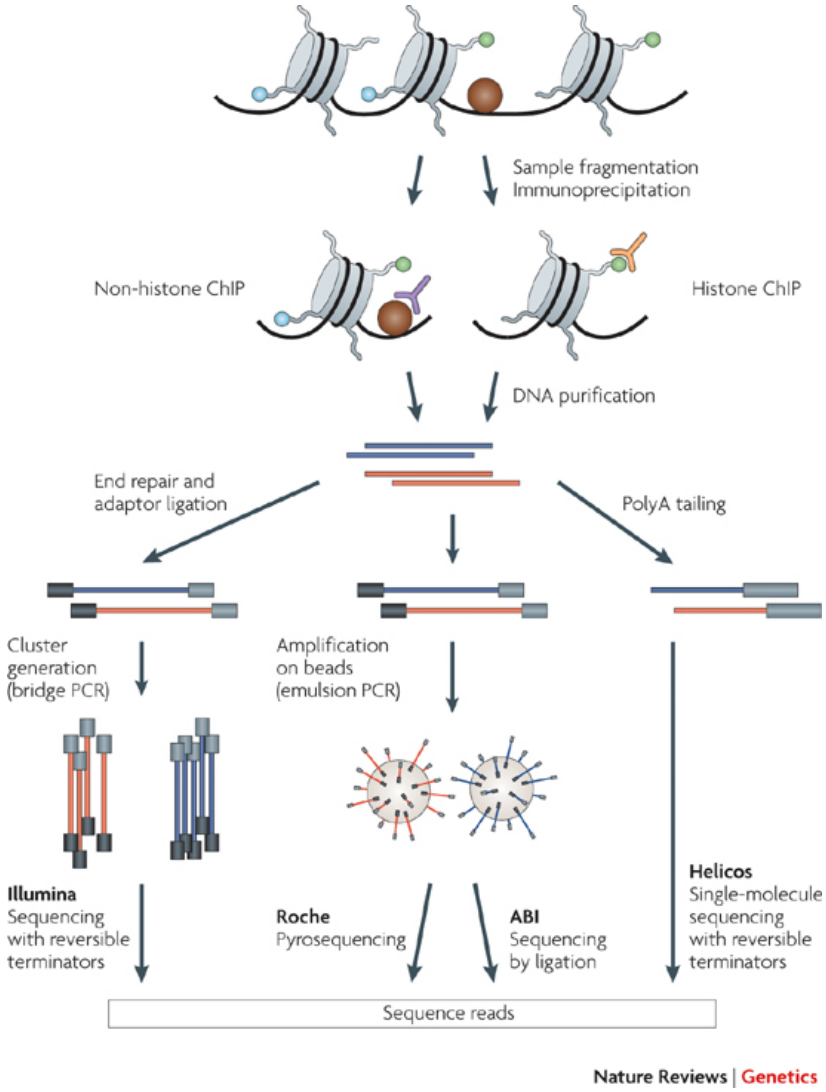
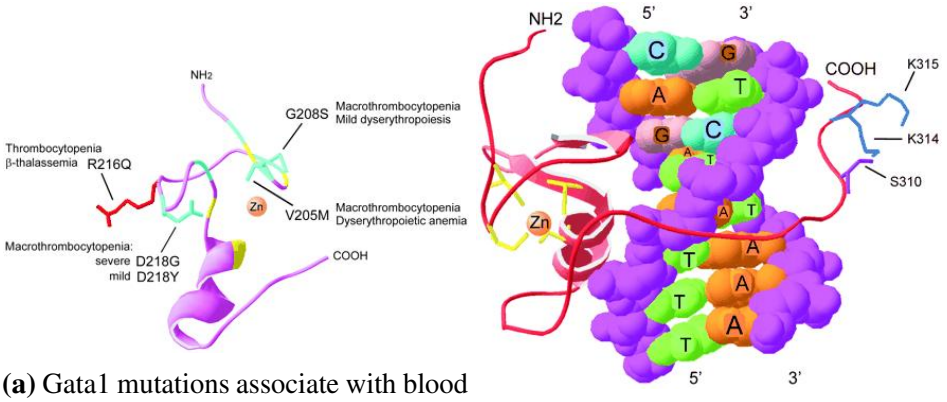


Figure 6: Overview of a chiP–seq experiment. Using ChIP followed by massively parallel sequencing, the specific DNA sites that interact with transcription factors or other chromatin-associated proteins (non-histone ChIP) and sites that correspond to modified nucleosomes (histone ChIP) can be profiled. The ChIP process enriches the crosslinked proteins or modified nucleosomes of interest using an antibody specific to the protein or the histone modification. Purified DNA can be sequenced on any of the next-generation platforms. The basic concepts are similar for different platforms: common adaptors are ligated to the ChIP DNA and clonally clustered amplicons are generated. The sequencing step involves the enzyme-driven extension of all templates in parallel. After each extension, the fluorescent labels that have been incorporated are detected through high-resolution imaging. On single-molecule sequencing platforms such as the HeliScope by Helicos (bottom right), fluorescent nucleotides incorporated into templates can be imaged at the level of single molecules, which makes clonal amplification unnecessary. (Park, 2009)

GATA fingers (Figure 7). All members of the GATA family share sequence homology in the zinc finger regions only; outside those, the conservation between GATA factors is low.



(a) Gata1 mutations associate with blood related pathologies

(b) Gata1 ZnF DNA binding

Figure 7: (a) Mutations in the N-terminal finger of GATA1 causing human disease. Blue, the mutations V205 M (75), G208S (64), D218G (25), and D218Y (26) interfere with FOG-1 binding; red, the mutation R216Q (133) interferes with DNA binding. (1GNF.pdb) (b) Three-dimensional (3D) representation of the C-terminal finger (66-aa) of chicken GATA1 (red) bound to DNA (5′-CAGATAAA-3′). The C-terminal extension of the zinc-finger makes extensive contacts with the minor groove of the DNA. The side chains of S310, K314, and K315 point away from the DNA, suggesting that they might be accessible to other proteins when GATA1 is bound to DNA. (2GAT.pdb) (Ferreira et al., 2005).

GATA1 is mainly expressed in the hematopoietic system, whereas, GATA2 and GATA3 are expressed in various tissues, including the hematopoietic system. GATA1 was isolated as a protein with binding specificity to the 3' enhancer of the β-globin locus and was cloned from a mouse erylthroleukemic (MEL) cell line cDNA library. The human homologue was cloned soon after and assigned to the X-chromosome at position Xp21-11. The murine homologue is also located on the X-chromosome. GATA1 is expressed in erythroid cells, megakaryocytes, mast cells, eosinophils and dendritic cells in hematopoiesis and in the Sertoli cells in the testis.

The promoter of the murine Gata1 gene has a CACC box and a double palindromic GATA site, which suggests an autoregulatory role for GATA1. The net effect of this autoregulation, whether it is positive or negative, remains controversial as it depends on the cell type and/or the stage of differentiation of the cells. For example, a targeted deletion of the palindromic GATA site results in loss of the eosinophil lineage only (Yu et al., 2002) whereas overexpression of GATA1 in the erythroid lineage causes downregulation of the endogenous GATA1 (Whyatt et al., 2000).

The Gata1 gene has two tissue specific promoters: a distal promoter, which contains the first exon and is specifically transcribed in the Sertoli cells and a proximal promoter, which encodes a shorter transcript than the one transcribed in the testis and is found in the erythroid cells only. Three functional domains have been described for GATA1 protein: an N-terminal

domain reported to act as a transcriptional activation domain in transient transcription assays and two zinc finger domains located toward the C-terminus of the protein. The C-terminal-most, zinc finger is responsible for binding to DNA, whereas the N-terminal zinc finger modulates and stabilizes DNA binding, for example, to more complex palindromic GATA motifs. Both zinc fingers are involved in protein-protein interactions.

GATA1 protein-protein interactions

Transcriptional regulation of erythropoiesis by GATA1 is partly accomplished by protein interactions with other transcription factors or co-factors. GATA1 has been reported to interact with many transcription factors, such as FOG1, EKLF, SCL/Tal1, PU.1 and cofactors such as CBP/p300, Brg1, Med1, MeCP1/NuRD, and others ((Cantor and Orkin, 2002; Ferreira et al., 2005; Lowry and Mackay, 2006).

Using an in vivo biotinylation tagging approach coupled with mass spectrometry, Rodriguez et al. (2005) isolated and characterized nuclear GATA1 protein complexes in differentiated MEL cells. This work showed that GATA1 forms two independent complexes with FOG1, with and without the repressive MeCP1/NuRD chromatin remodeling/histone deacetylase complex. In addition, GATA1 forms distinct complexes with the hematopoietic transcription factors SCL/Tal1 (and associated partners) or Gfi-1b and with the chromatin remodeling complex ACF/WCRF.

Importantly, most of our current understanding regarding the molecular mechanisms of action of lineage specific transcription factors relies on gene knock-in and knock-out genetic experiments on mouse models. Furthermore, because of experimental limitations, most of these mechanisms have been tested on a small number of established target genes (e.g. α- and β-globin, Gata1, Gata2 and cKit *loci*), thus raising the question of their genome wide and generalized validity.

SCL/Tal1

Stem-cell leukemia (SCL, also known as Tal1, SCL/Tal1) is required for blood cell development and has ben shown to participate in protein complexes with GATA factors at activating sites (Aplan et al., 1992; Shivdasani et al., 1995; Tripic et al., 2009). SCL/Tal1 is required for the specification of all hematopoietic lineages, whereas scl/tal1 null ($SCL^{-/-}$) mice die at approximately embryonic day 9.5 (Porcher et al., 1996; Robb et al., 1995, 1996; Shivdasani et al., 1995). As mentioned above, despite its critical role in HSC generation, SCL/Tal1 is not required for HSC survival and multipotency in mice. However, in the absence of

SCL/Tal1, differentiation of MEPs into both erythroid and megakaryocytic progenitors is severely impaired (Mikkola et al., 2003).

SCL/Tal1 functions as a core member in the GATA-nucleated regulatory complex, which also includes LMO2, LDB1, E2A, and sometimes ETO2. At genomic loci where GATA1 represses transcription, the SCL complex is absent (Tripic et al., 2009). In general, SCL/Tal1, LMO2, and LDB1 compose a complex that is primarily activating, whereas ETO2 serves as a corepressor (Fujiwara et al., 2009; Goardon et al., 2006; Hamlett et al., 2008; Tripic et al., 2009). Interestingly, mice that harbor a congenital mutation in the basic helix-loop-helix DNA-binding domain of SCL (SCL^{RER}) do not all die as early as the complete knockouts, suggesting that SCL/Tal1 DNA binding is not always direct or that SCL/Tal1 has non–DNA binding functions. More specifically, it appears that DNA binding is dispensable for the HSC specification functions of SCL/Tal1, but not for SCL/Tal1 functions related to erythroid maturation (Kassouf et al., 2010).

LDB1

Another component of the GATA/SCL complex, LDB1, has also been shown to be required for chromatin loop formation at the Hbb *locus* in murine cell lines (Song et al., 2007). LDB1 null mice do not make red blood cells and die of various morphologic abnormalities and anemia at approximately embryonic day 9.5 (Mukhopadhyay et al., 2003). Recent work has identified a requirement for LDB1 throughout embryonic and adult erythroid and megakaryocytic development. The same study also suggested that LDB1 is a critical downstream effector of the transcriptional activation network of GATA1 (Li et al., 2010). The LDB1 chromatin occupancy repertoire was also recently interrogated in MEL cells. LDB1 binds to genomic sites primarily as a part of an activating complex and is also thought to have a role in the formation of chromatin loops and long-range interactions between the Hbb locus and other LDB1 bound genes on chromosome 7 (Soler et al., 2010). However, the purpose of these long-range LDB1 interactions remains unclear.

KLF1

The CACCC-binding nuclear factor, KLF1 (also known as EKLF), is the founding member of the mammalian Kruppel-like family of transcription factors (Miller and Bieker, 1993). Expression of KLF1 is critical for activation of β-globin transcription and mice lacking KLF1 die in utero (Donze et al., 1995; Nuez et al., 1995; Perkins et al., 1995). KLF1 has been shown to interact with the histone acetyltransferases CBP and p300, which acetylate KLF1 and enhance its transcriptional activity *in vitro* and *in vivo* (Zhang and Bieker, 1998). Most

interestingly, KLF1 has been shown to be directly involved with nearly all aspects of the heme synthesis and iron procurement pathway in maturing erythroid cells, establishing a crucial molecular role for KLF1 in the establishment of functional erythrocytes (Tallack et al., 2010).

A role for KLF1 in MEPs and the megakaryocyte-erythroid lineage fate decision has been also proposed. During the mesodermal specification of in vitro differentiated murine embryoid bodies, Klf1 is activated in a GATA1–independent manner by a GATA2 and Smad5-nucleated complex (Lohmann and Bieker, 2008). Using a GFP reporter under the control of the Klf1 promoter, Lohmann et al showed that cells that express GFP, and not the erythroid-specific marker Ter119, are capable of ultimately differentiating into either megakaryocytic or erythrocytic colonies, suggesting that KLF1 has a role in MEPs (Lohmann and Bieker, 2008). Furthermore, gain- and loss-of-function studies in embryoid bodies differentiation systems showed that enforced KLF1 expression selectively blocks megakaryocyte development (to the benefit of erythroid development) when KLF1 is expressed during a short window before megakaryocyte-erythroid lineage choice (Frontelo et al., 2007). Together, these reports demonstrate a role for KLF1 in the MEP and the megakaryocyte-erythroid lineage switch.

FLI1

Homozygous loss of functional Fli1 alleles in mice leads to embryonic lethality because of severe defects in fetal megakaryopoiesis and a high incidence of embryonic hemorrhaging. The latter phenotype is presumably the result of coagulation defects secondary to impaired megakaryopoiesis as well as inefficient blood vessel formation, as FLI1 is also essential for vascular endothelium and hemangioblast specification (Hart et al., 2000; Liu et al., 2008; Spyropoulos et al., 2000).

In HSCs, FLI1, together with GATA2 and SCL/Tal1 compose a fully connected triad, wherein each factor directly activates the expression of each of the other 2 factors, leading to a robust and consistently active network module (Pimanda et al., 2007). In the multipotent HPC-7 line, FLI1 occupancy closely paralleled genome-wide occupancy of the closely related ETS factor ERG. In addition, FLI1 was identified as a core constituent of a regulatory heptad of transcription factors (ERG, FLI1, GATA-2, LMO2, LYL1, RUNX1, and SCL). Within a population of HPC-7 cells, these factors co-occupy genomic regions associated with hundreds of genes that are enriched for cell death, cell cycle, signaling, and transcriptional control. Although, there is no evidence that all of these factors are bound to the same span of DNA at the same time within the same cell (Wilson et al., 2010).

In megakaryocytic cells, FLI1 binds to and directly regulates the genes encoding several proteins that are essential for terminal megakaryocytic maturation. Specifically, in primary fe-

tal liver-derived megakaryocytes FLI1 binds and, potentially, regulates Itga2b (which encodes the CD41 antigen) as well as Gp1ba (CD42), Gpix (glycoprotein 9), Mpl (thrombopoietin receptor), and Cxcl4 (platelet factor 4) (Pang et al., 2006).

Epigenetic changes in chromatin during erythroid differentiation

In many developmental systems, post-translational modifications of histones are crucial in regulating gene expression (Suganuma and Workman, 2011; Talbert and Henikoff, 2010). Although some modifications tend to be associated with gene activation or repression states, the actual situation is generally more complex.

For example, the H3K4me3 modification at transcriptional start sites (TSSs) is commonly associated with gene activation and transcription initiation by RNA polymerase II (RNApol2). However, the H3K4 modification often colocalizes with H3K27me3 modifications, which are associated with repressed genes. Such chromatin regions marked both by H3K27me3 and H3K4me3 are termed "bivalent domains." Such domains are found in embryonic stem cells on genes encoding key developmental transcription factors (Bernstein et al., 2006). Similar bivalent domains occur in human primary HSCs/progenitor cells ($CD133^+$) that can differentiate into CD36-expressing erythrocyte precursors (Cui et al., 2009).

Several epigenetic regulatory mechanisms control gene induction and repression during erythroid development (Figure 8). Changes in gene expression are not accompanied by significant changes in histone modifications, such as H3K4me3 and H3K27me3, after GATA1 is reintroduced into a GATA1-null erythroblast cell line to induce differentiation (Wu et al., 2011). Changes in H4K16ac and H3K79me2 levels, rather than H3K4me3 and H3K27me3, are most predictive of the direction in changes in gene expression during terminal fetal liver erythroid differentiation (Wong et al., 2011). Because H3K4me3 is usually associated with transcriptional initiation whereas H3K79me2 is tightly correlated with transcription elongation, control of RNApol2 elongation could be a mechanism for regulating erythroid gene expression. This may be mediated by GATA1 or SCL/TAL1 or their associated complexes, especially because nearby regions of genes induced during terminal erythroid development are co-occupied by these factors (Wu et al., 2011). The precise timing of chromatin switch(es) associated with erythroid differentiation is unclear. Several modifications at the β-globin locus (DNA demethylation, formation of DNase I hypersensitive sites, and onset of activation-associated histone modifications) occur during the S phase of an early erythroid cell cycle after stimulation of CFU-E proliferation (Pop

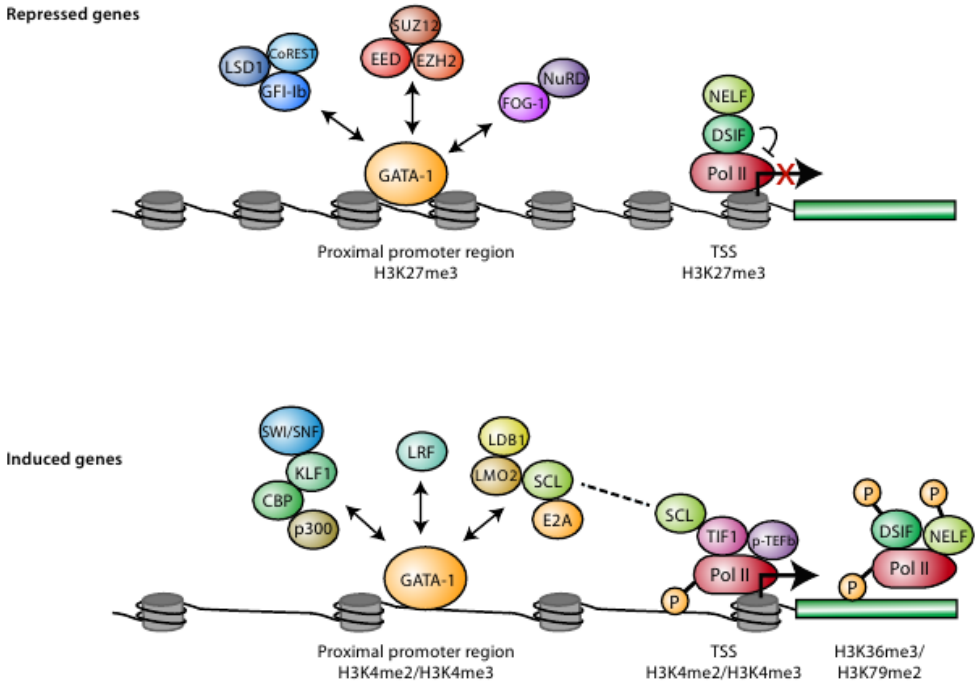


Figure 8: Transcription factors, Pol II status, and histone modifications associated with actively transcribed and repressed genes in erythroid cells. For genes activated during differentiation (bottom panel), the GATA1 activation-associated proteins exist in different complexes (as indicated by the double arrows) through binding to both distal regulatory regions and promoter regions near the TSS. H3K4me2 and H3K4me3 are enriched in some of these bound regions. TIF1 and the SCL complex recruit pTEF-b to promote RNApol2 elongation by phosphorylating DSIF, NELF, and RNApol2. SCF could be the link between the GATA1 complex and RNApol2 elongation machinery. H3K79me2 and H3K36me3 are often enriched in the transcribed portion of the gene. Among repressed genes (top panel), the GATA1 repressor complexes are less clear and could exist in different forms as indicated by the double arrows. H3K27me3 is enriched near the TSS. Pol II is either not bound or is paused around the TSS. In the absence of TIF-1, recruitment of p-TEFb is impaired, and DSIF and NELF inhibit the phosphorylation of RNApol2 and RNApol2 elongation. (Hattangadi et al., 2011)

et al., 2010). This raises the possibility that one window of time during DNA replication allows structural changes in chromatin associated with newly synthesized DNA.

Certain histone modifications affect not only binding of regulatory proteins but also the stability of chromatin itself. Global levels of several acetylation marks on histones, including H3K9Ac, H4K5Ac, H4K8Ac, and H4K12Ac, are significantly reduced during the terminal stages of erythroid differentiation, concomitant with a decrease in the levels of the histone acetyltransferase GCN5 (Jayapal et al., 2010). Administration of histone deacetylase (HDAC) inhibitors to human erythroid precursors cultured from $CD34^+$ cells inhibited terminal differentiation (Fujieda et al., 2005). The HDAC most important for erythropoiesis is probably HDAC2 because either specific pharmacologic inhibition or shRNA knockdown

of HDAC2 blocked chromatin condensation and enucleation of mouse fetal liver erythroblasts *in vitro* (Ji et al., 2010).

Scope of this study

The main focus of this study is the integrative analysis and interpretation of genome wide biological data to identify distinct regulatory events, such as, differential chromatin states and transcription factor occupancy profiles, that distinguish differentially regulated gene activity during the blood lineage specification and differentiation processes (hematopoiesis).

Some of the main issues that we needed to engage in this process are: 'How to extract biological information and testable hypotheses from genome-scale experiments?' and, importantly, 'How to extract biological information from genome-scale experiments using a data-driven approach as opposed to a hypothesis-driven validation approach?'. In addition, despite the huge accumulation of NGS data on transcription factor occupancies and epigenetic modifications, most of our current knowledge of gene regulatory mechanisms relies on the extensive analysis of only a few gene *loci* at a time. Thus, there is a pressing need to develop approaches that allow for a biologically and functionally meaningful categorization of large scale datasets into distinct, data based, relational contexts, allowing for the interpretation of their genome wide relations.

In addressing these questions we focused on erythroid lineage specification and terminal differentiation as our main model system. In order to extend our understanding of the erythroid differentiation process, we first focused on the erythroid master transcription factor GATA1 by producing genome-wide occupancy maps using chromatin immunoprecipitation coupled to massive parallel sequencing (ChIPseq) of mouse primary erythroid cells.

In the first results chapter we describe the comprehensive analysis of the in vivo GATA1 genome-wide occupancy profiles in fetal liver derived Ter119^{-} proerythroblasts and Ter119^{+} erythroblasts (Papadopoulos et al., 2013). We developed a Random Forest based systematic approach to accurately identify potential target genes and a genomic similarity based clustering approach to identify functionally distinct subsets of GATA1 target genes. More specifically, the computational analysis pipeline applied here relies on the combination of supervised (RandomForest regression) and unsupervised (hierarchical clustering) machine learning algorithms to produce highly structured gene wide distribution patterns of chromatin features in different cell populations.

In the second results chapter we applied a similar approach to investigate the differential transcriptional and epigenetic events underlying the specification of the erythroid and megakaryocytic lineages, which derive from a common progenitor and differentiate into

fully distinct mature cell types. Based on our results, one of the main events associated with the erythroid lineage specification process is the highly specific loss of active promoter chromatin marks from a large group of genes, functionally associated with alternative blood lineages, such as megakaryocytes and lymphoid cells. Importantly, the active promoter state of these genes appears to be present in hematopoietic stem cells and is maintained throughout the megakaryocytic specification and differentiation processes. Interestingly, comparison of the epigenetic signatures identified in mouse and human erythroid specification revealed both similar and distinct characteristics.

Finally, to facilitate the interrogation and interpretation of the results we developed a web based portal that allows the user to navigate through the gene wide distribution patterns of chromatin features in different hematopoietic cell populations. In addition, the portal includes an interactive graphical display of the read density profiles of single gene *loci*, allowing for the generation of hypotheses about the mechanisms of differential gene regulation of specific genes, as well as for the design of validation experiments.

Notably, the computational pipeline proposed in this study was successfully applied in a wide range of biological processes (erythroid differentiation and lineage commitment) and settings (different subsets of genes, differential gene expression and differential chromatin levels) and using different combinations of NGS datasets (histone modifications, TF occupancy profiles, DNA methylation, DNase hypersensitivity), thus suggesting its general applicability in the analysis and interpretation of complex genomic data.

Chapter 1

GATA1 genome wide occupancies in erythroid cells

Abstract

In this chapter [1] we report the genomic occupancy profiles of the key erythroid hematopoietic transcription factor GATA1 in mouse fetal liver derived pro-erythroblasts and mature erythroid cells. Integration of the identified GATA1 occupancy profiles with available genome-wide transcription factor and epigenetic profiles assayed in fetal liver erythroid cells enabled us to assess the involvement of GATA1 in modulating local chromatin structure of target gene *loci* during terminal erythroid differentiation.

The results presented here suggest that GATA1 associates preferentially with specific epigenetic modification changes, such as acetylation of lysine 16 and lysine 27 of histone H4 and H3, respectively, and di-methylation of lysine 4 of histone H3. To further dissect the epigenetic and transcriptional regulatory patterns established by GATA1 we used random forest (RF) non-linear regression to predict changes in the expression levels of its target genes by including all genomic features available for pro-erythroblasts and mature fetal liver-derived erythroid cells. Remarkably, our prediction model explained a high proportion (62%) of variation in gene expression, within the identified GATA1 target genes.

Hierarchical clustering of the proximity values calculated by the RF model produced a clear separation of upregulated versus downregulated genes and a further separation of downregulated genes in two distinct groups, thus, suggesting two distinct molecular mechanisms of GATA1 target gene downregulation. In conclusion, our study of GATA1 genome-wide occupancy profiles in mouse primary erythroid cells and their integration with

[1] Most of the work presented here is reproduced from the Papadopoulos et al manuscript published in 2013 in Nucleic Acids Research.

global epigenetic marks reveals three clusters of GATA1 gene targets, each associated with specific epigenetic signatures and functional characteristics.

1.1 Introduction

1.1.1 Erythroid Differentiation

Erythropoiesis is a dynamic complex multistep process involving the terminal differentiation of erythroid progenitors to enucleated red blood cells (reviewed in Tsiftsoglou et al. (2009)). Erythroid cell differentiation is a well characterized process and thus makes for an ideal model system to study the molecular events driving terminal cell differentiation. The various differentiation stages of committed erythroid cells are distinguishable by the differential expression of specific cell surface markers (Socolovsky et al., 2001) and unique morphologies (Gutierrez et al., 2005). One of the key cell surface erythroid-specific antigens expressed primarily by terminally differentiating erythroblasts is Ter119 which is widely used to separate mature erythroid cells from proerythroblasts (Kina et al., 2000; Socolovsky et al., 2001). Commitment of multipotent progenitor cells (MEPs) to committed erythroid progenitors ($CD71^+$) involves both the activation of the erythroid specific transcription program (heme production, erythroid cell membrane proteins, oxidative stress etc.), as well as the repression of early hematopoietic multipotentiality and of alternative hematopoietic lineage transcription programs (MEGs, lymphoid cells) (reviewed in Hattangadi et al. (2011)).

1.1.2 GATA1: Erythropoiesis' Master Regulator

GATA1 is a critical transcription factor that is essential for the terminal differentiation of erythroid cells and of other hematopoietic lineages, such as megakaryocytes, mast cells and eosinophils (reviewed in Cantor and Orkin (2002); Crispino (2005); Ferreira et al. (2005)). The erythroid specificity of GATA1 is highlighted by the fact that all presently known erythroid specific genes include GATA binding sites in their promoters, including those for Gata1 itself and for the transcription factors Gata2, Klf1 and Scl/Tal-1 (reviewed in Tsiftsoglou et al. (2009)). GATA1 is involved in both the activation or repression of the erythroid and alternative lineage transcription programs, respectively, that occurs with terminal erythroid differentiation (Rodriguez et al., 2005; Welch et al., 2004).

Several lines of evidence have suggested that GATA1 binding leads to changes in the epigenetic landscape of target genes (Letting et al., 2003; Steger et al., 2008; Yu et al., 2009). Furthermore, GATA1 has been reported to interact with several hematopoietic transcription factors as well as chromatin remodelling and modification complexes, such as the NuRD

complex, histone acetyl transferases and Polycomb-Group members (Ferreira et al., 2005; Tsiftsoglou et al., 2009; Yu et al., 2009). Significantly, enforced ectopic GATA1 expression in highly purified murine progenitor cells (myeloid or lymphoid), has been shown to reprogram them towards the erythroid and megakaryocytic lineages that GATA1 normally regulates (Graf, 2002; Heyworth et al., 2002; Iwasaki et al., 2003). Thus, GATA1 is capable of imposing an erythroid transcription program in myeloid-derived hematopoietic lineages, establishing it as a "master" erythroid transcription factor. As erythropoiesis progresses, GATA1 protein levels vary from basal in the hematopoietic progenitors, to maximal in committed erythroid cells and low during terminal differentiation (Ferreira et al., 2005). Moreover, altered GATA1 genome wide occupancy leads to changes in both gene expression levels and epigenetic changes in chromatin structure, in order to allow for the terminal erythroid maturation program to be completed (reviewed in Bresnick et al. (2010); Nakajima (2011); Tsiftsoglou et al. (2009); Wickrema and Crispino (2007)).

1.2 Characterization of GATA1 genome wide occupancy in erythroid cells

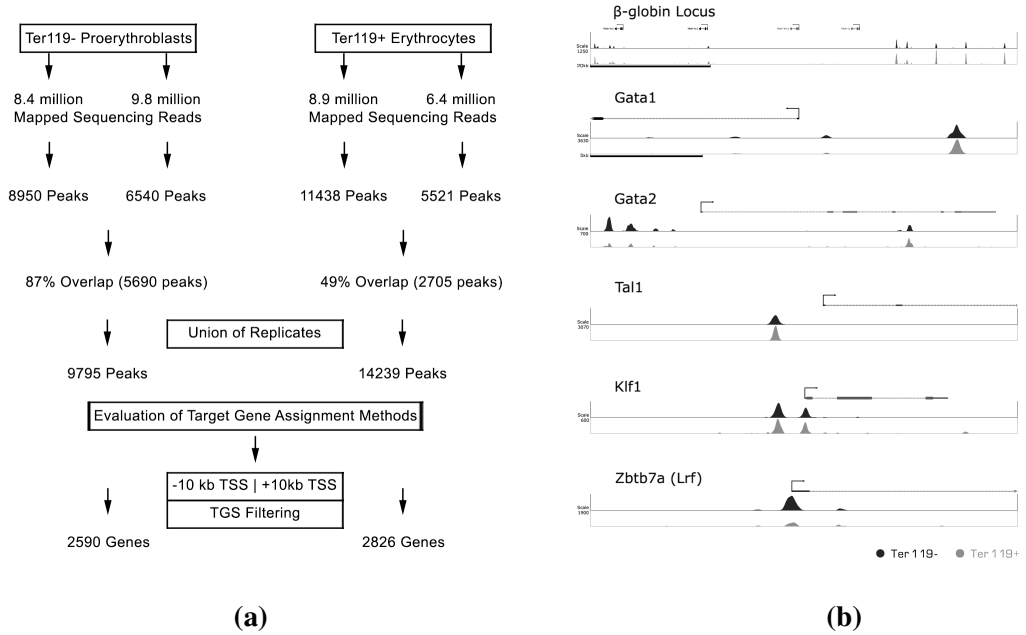
GATA1 genome-wide occupancy by ChIP-seq has been previously described in both mouse (Cheng et al., 2009; Soler et al., 2010; Yu et al., 2009) and human (Fujiwara et al., 2009) erythroid cell lines or in in vitro differentiated mouse ES cells (Wontakal et al., 2012). These studies agree in that (i) GATA1 binds mostly to sequences that are distal to promoters, probably corresponding to enhancer elements; (ii) GATA1 targets include genes that are both activated and repressed with differentiation; (iii) GATA1 gene targets are enriched for histone H3K4 methylation marks; (iv) there is a strong positive correlation between activated GATA1 target genes and binding of the SCL/TAL1 complex (Kassouf et al., 2010). Integration of GATA1 ChIP-seq data with those for SCL/TAL1 and KLF1 led to the identification of a few hundred gene targets that are common to all three factors (Kassouf et al., 2010; Pilon et al., 2011; Tallack et al., 2010; Wontakal et al., 2012), which have been proposed to represent a core erythroid network enriched for genes involved in erythroid differentiation (Wontakal et al., 2012).

Importantly, the genome-wide GATA1 binding patterns in mouse primary fetal liver-derived erythroid cells have only been reported in the context of providing limited validation for GATA1 ChIP-seq data in erythroid cell lines (Soler et al., 2010; Wu et al., 2011), therefore a complete analysis of GATA1 genome wide occupancies in primary erythroid cells had not been carried out at the time.

In the work described here, we provide for the first time a comprehensive analysis of the *in vivo* GATA1 occupancy profiles in fetal liver derived Ter119$^-$ proerythroblasts and Ter119$^+$ mature erythroid cells.

1.2.1 *In vivo* GATA1 genomic occupancy in primary erythroid cells

In order to identify genome-wide differential GATA1 binding patterns during erythroid differentiation in vivo, we performed GATA1 ChIP on Ter119$^-$ proerythroblasts and Ter119$^+$ mature erythroid cells fractionated from day E12.5 mouse fetal liver cells, followed by high throughput massive parallel sequencing. ChIPed DNA from Ter119$^-$ and Ter119$^+$ cells was sequenced in replicates to generate a total of 18.2 million and 15.3 million uniquely mapped sequence reads, respectively (Figure 1.1a).



(a) **(b)**

Figure 1.1: Determination of GATA1 chromatin occupancy in Ter119$^-$ and Ter119$^+$ cells by ChIP sequencing. **a)** Description of the pipeline followed for the analysis of raw deep sequencing results of GATA1 ChIP-seq leading to the identification of occupancy sites (peaks) and potential target genes. **b)** *Bona fide* GATA1 binding sites as determined by GATA1 ChIP-seq in Ter119$^-$ and Ter119$^+$ cells. Scale refers to normalized read counts.

Using the QuEST peak-calling algorithm (Valouev et al., 2008) we assembled the unique, non-redundant sequence reads for each replicate into peaks to identify potential GATA1 bound regions across the genome. For both samples, we took the union of the peaks of the two replicates, resulting in 9,795 peaks and 14,239 peaks for the Ter119$^-$ and Ter119$^+$ samples, respectively (Figure 1.1a). Visualization of the sequencing read density profiles in

known GATA1 target gene *loci*, such as β-globin, Gata1, Gata2, Klf1, Zbtb7 or Scl/Tal-1 gene *loci* (Kassouf et al., 2010; Martowicz et al., 2005; Rodriguez et al., 2005; Valverde-Garduno et al., 2004; Yu et al., 2009) provided early validation for our sequencing data in both the Ter119^{-} and Ter119^{+} cell populations (Figure 1.1b).

To explore the genome wide distribution of GATA1 occupancy with respect to Transcription Start Sites (TSSs) we interrogated the distance distribution of all identified peaks from annotated genes. This analysis led to the conclusion that the majority of the peaks clustered proximally (within 5kb) to gene TSSs (Figure 1.2a and Figure 1.2b).

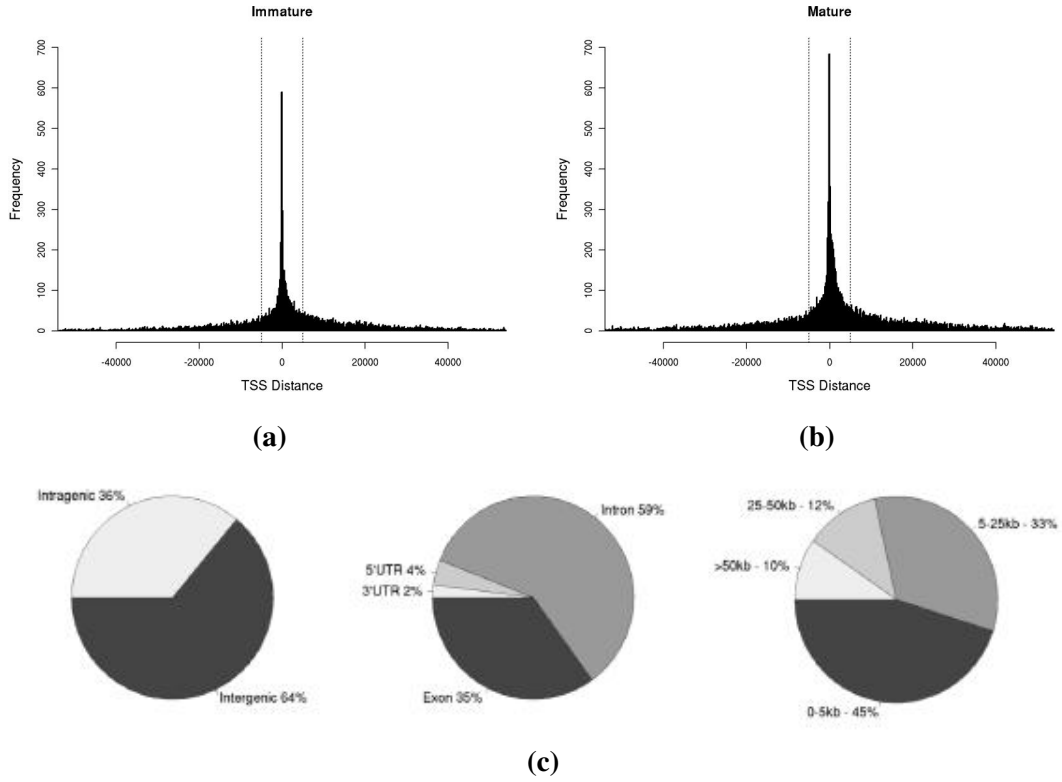


Figure 1.2: Distribution of GATA1 chromatin occupancy in Immature and Mature erythroid cells. **a)** and **b)**: Distribution of the distances between GATA1 identified peaks from annotated gene TSSs. **c)** Location analysis of GATA1 peaks. Left: Percentage of intragenic and intergenic peaks. Middle: Distribution of intragenic GATA1 peaks in 5'UTR, 3'UTR, introns and exons. Right: Distribution of intergenic GATA1 peaks in 0-5kb, 5-25kb, 25-50kb and >50kb windows from TSS.

Furthermore, in both Ter119^{-} and Ter119^{+} samples, we find a similar distribution of approximately 64% and 36% of peaks falling within intergenic and intragenic regions, respectively (Figure 1.2c). Of the intragenic peaks, 59% fall within introns and 35% within exons, with clear clustering of peaks towards the 5'-most introns and exons (Figure 1.2c,

right panel and not shown). By contrast only 4% of GATA1 peaks fall within 5'UTRs and 2% within 3' UTRs in vivo (Figure 1.2c, middle panel).

By and large, our data on GATA1 peak distribution are consistent with previous data obtained in erythroid cell lines (e.g. MEL cells (Yu et al., 2009)). Overall, we do not observe any differences in intergenic or intragenic GATA1 peak distribution between Ter119$^-$ proerythroblasts and Ter119$^+$ erythroid cells.

1.2.2 Peak assignment to potential GATA1 gene targets

We next sought to assign specific genes to the GATA1 peaks identified in the Ter119$^-$ and Ter119$^+$ datasets. This is usually done by nearest gene assignment or by assigning peaks that fall within a given window around a gene's transcription start site (TSS) and/or transcription end site (TES) (MacQuarrie et al., 2011). As this has led to differences in target gene assignments in different GATA1 ChIP-seq studies (Kerenyi and Orkin, 2010), we approached it in a more systematic way. More specifically, we tested a series of different target gene assignment parameters to identify one that provides the most significant association between GATA1 occupancy and changes in the expression level of the assigned target gene. In order to quantify the association between GATA1 occupancy and changes in the expression levels of the target genes identified by each assignment method, we constructed a series of Random Forest (RF)-based non-linear regressors (Breiman, 2001), using GATA1 occupancy features as predictors of the target genes' expression fold change. The rationale behind this approach is that the gene assignment method that will capture the best association between GATA1 occupancy and its potential target genes will maximize the models' accuracy in predicting the target genes' expression change.

Gene expression fold changes were determined based on the RNA-sequencing expression data obtained in Ter119$^-$ and Ter119$^+$ mouse fetal liver erythroid cells by Wong et al. (2011). In order to determine our training datasets we scored for GATA1 peaks found within windows of increasing size (i.e. ±1kb, ±2kb, ±5kb, ±10kb, ±20kb) around a gene's TSS, or within a region extending from -20kb from a gene's TSS to +10kb from a gene's TES, or by assigning peaks to the nearest TSS (Figure 1.3). The number of potential GATA1 target genes thus identified varied from 919 to 4,551 expressed genes in Ter119$^-$ cells and from 1,008 to 5080 in Ter119$^+$ cells, depending on the different assignment parameters (Figure 1.3).

Potential GATA1 target genes identified in this way, were next ascribed a number of features (predictors) based on the GATA1 occupancy profiles in Ter119$^-$ and Ter119$^+$ cells. These predictors included the total gene score (TGS), defined as the sum of the GATA1 peak scores assigned to it, the difference in TGS score between Ter119$^-$ and Ter119$^+$ cells, the highest GATA1 peak score assigned to the gene, the minimum and maximum distances of

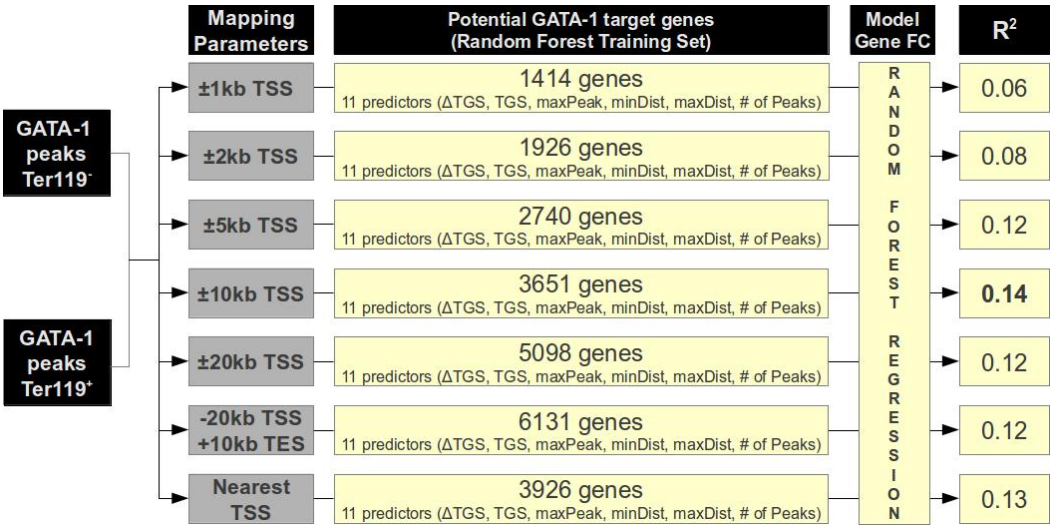


Figure 1.3: Schematic representation of the pipeline used to individually evaluate the association of different gene assignment methods with differential gene expression. The number of genes reported refers to the union of GATA1 target genes identified in Ter119$^-$ and Ter119$^+$ cells

assigned GATA1 peaks from the gene's TSS and the total number of assigned GATA1 peaks, thus resulting in a total of 11 features per gene in Ter119$^-$ and Ter119$^+$ cells (Figure 1.3).

The different RF regression models were ranked according to the coefficient of determination achieved by each model (R^2, see Materials and Methods). As a result, all proposed regressors showed a moderate degree of gene expression variation explained by the different models, in line with the assumption that GATA1 occupancy is directly associated with the modulation of gene expression during erythroid differentiation.

Furthermore, the difference in accuracy of the trained models show that a systematic choice of the target gene assignment method can identify differences in the accuracy of the ensemble of target genes, provided that an additional functional feature is available (i.e. gene expression profiles). Overall, the most accurate ensemble of GATA1 target genes in erythroid cells was obtained by assigning genes harboring a GATA1 peak within a ±10kb window of their TSS (R^2=0.14, 3651 genes).

1.2.3 Analysis of GATA1 target genes

Based on the ±10kb mapping, a total of 2,590 and 2,826 potential GATA1 target genes were identified in the Ter119$^-$ and Ter119$^+$ datasets, respectively. The union of the two datasets yielded a total of 3,651 potential GATA1 target genes, of which 1,765 genes were common to both Ter119$^-$ and Ter119$^+$ datasets thus giving an intersection of 48.3% (Figure 1.4). By

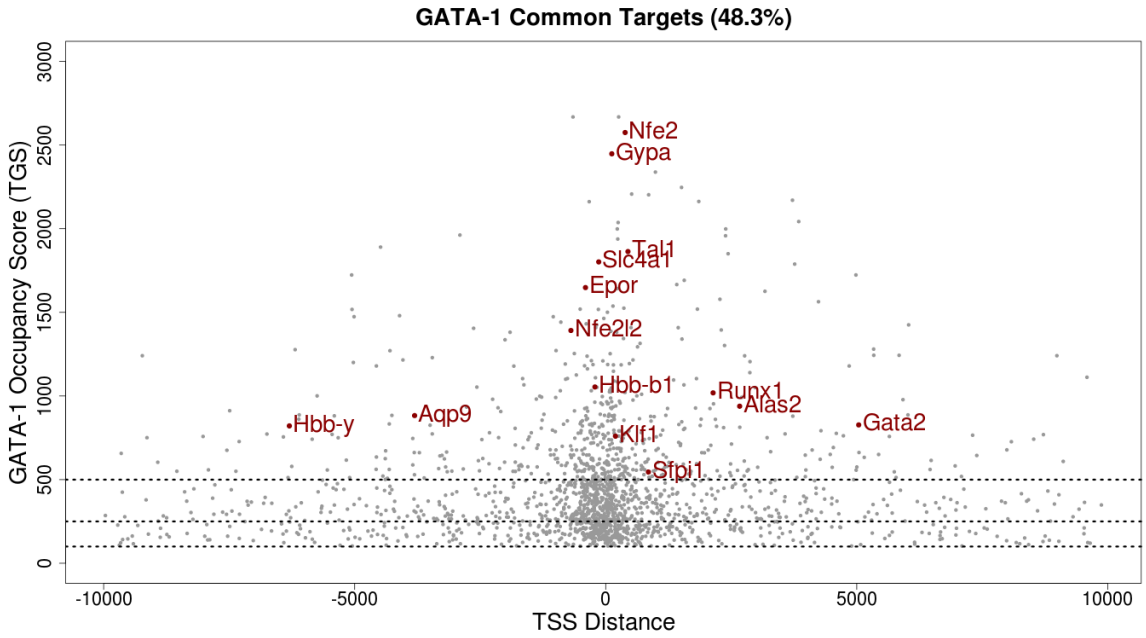
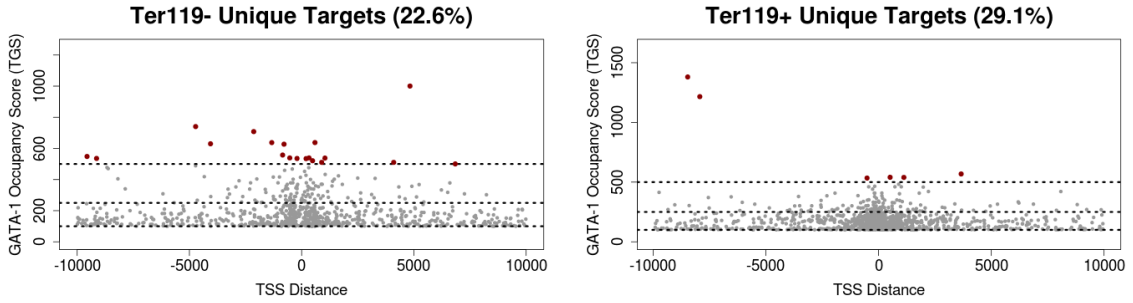


Figure 1.4: Scatterplot of GATA1 potential target genes identified in both early and late erythroid progenitors. Horizontal lines define score thresholds for the 3 TGS classes. Most highly enriched (Class I) target genes are found in the intersection of the two datasets and comprise most of the GATA1 target genes described in the bibliography (highlighted genes in the plot).

contrast, 825 (22.6%) and 1,061 (29.1%) genes were unique to the Ter119$^-$ or Ter119$^+$ cells, respectively (Figure 1.5a and 1.5b).



(a) **(b)**

Figure 1.5: GATA1 potential target genes identified uniquely in Ter119$^-$ **a)** or Ter119$^+$ **b)** cells. Despite the fact that target genes unique in either Ter119$^-$ or Ter119$^+$ cells are poor in Class I genes (red dots) they still show a prominent clustering of GATA1 occupancy sites near the identified target gene TSS.

These data reveal a considerable conservation of GATA1 target genes throughout erythroid differentiation.

In order to further facilitate the differential analysis of potential GATA1 gene targets in Ter119$^-$ and Ter119$^+$ cells, we arbitrarily divided all genes into three categories on the basis of their TGS:

Class I includes genes with a TGS greater than 500, Class II includes genes with a TGS of 500 to 250 and Class III includes genes with a TGS of 250 to 100. Inspection of the three classes of genes led to a number of observations. First, it is clear that Class I includes most of the well-established GATA1 erythroid-specific target genes. For example, genes like the Gata1 locus itself, Gata2, the β-globin locus (especially the Locus Control Region (LCR)), EpoR, Nfe2, Slc4a1, Gypa, Tal1, Lrf, Klf1, Nrf2, Runx1, Alas2 all have a TGS score greater than 500 (Figure 1.4). Furthermore, Class I target genes are also markedly enriched in erythroid cell-related ontology terms (Figure 1.6). Thus, Class I genes, corresponding to approximately 15% of all identified GATA1 target genes (Figure 1.4) most likely represent the erythroid transcription program.

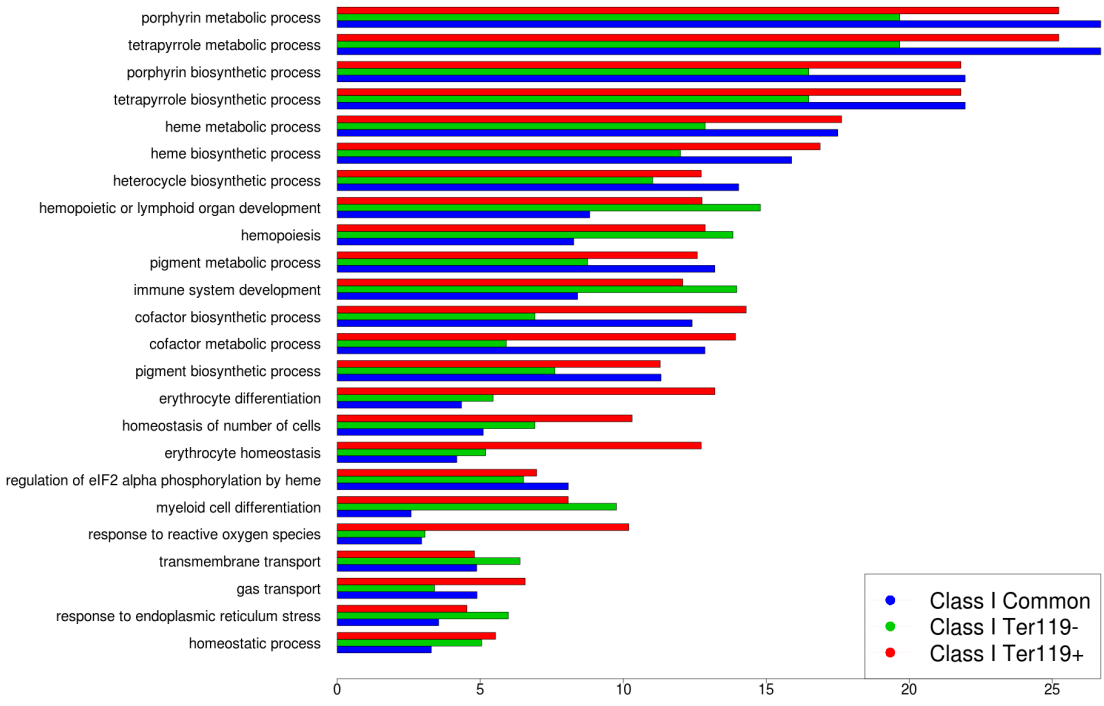


Figure 1.6: GO analysis of GATA1 target genes composing the TGS Class I. Results are shown for genes identified in Ter119$^-$ (green bars) and/or Ter119$^+$ (red bars). Blue bars refer to genes identified in both Ter119$^-$ and Ter119$^+$ cells with a TGS score higher than 500 (Class I threshold).

A second observation arising from this analysis is that Class II and III genes include most of the GATA1 targets that are unique to the Ter119$^-$ or Ter119$^+$ cells (806/825 and 1,055/1,061 genes, respectively; Figure 1.7a and 1.7b). More specifically, of the 1,765 genes that are bound by GATA1 in both Ter119$^-$ and Ter119$^+$ cells, 480 genes (27%) show reduced GATA1 binding in mature Ter119$^+$ cells compared to Ter119$^-$ cells, whereas, 353 genes (20%) transitioned to a higher class as a result of higher enrichment for GATA1 binding with erythroid differentiation (Figure 1.7b). The differences in the numbers between these two datasets suggest that low binding GATA1 targets are being lost from Ter119$^-$ cells and new ones are being acquired in Ter119$^+$ cells with erythroid differentiation (Figure 1.5 and 1.7a). Thirdly, mobility of an appreciable fraction of GATA1 targets within the three Classes is seen as erythroid differentiation proceeds from Ter119$^-$ to Ter119$^+$ cells.

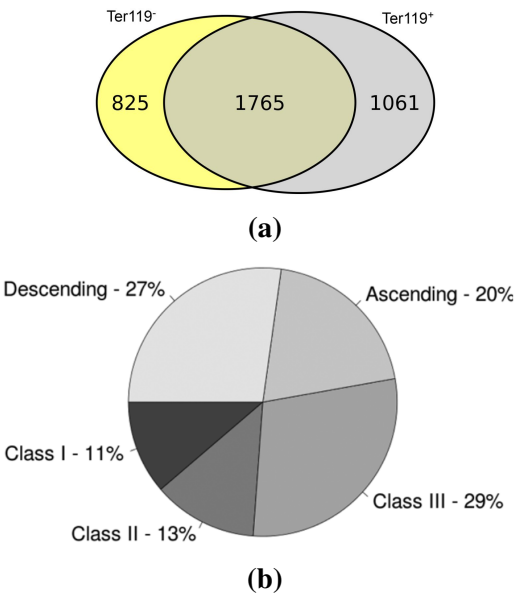


Figure 1.7: Comparison of GATA1 target genes with erythroid differentiation. **a)** Venn diagram showing the overlapping and unique GATA-1 target genes identified in Ter119$^-$ and Ter119$^+$ erythroid cells. **b)** Distribution of GATA1 potential target genes in Classes. Pie chart illustrating the distribution of common genes in three TGS classes and also the presence of a large quota of genes showing dynamic changes in GATA1 binding throughout erythroid differentiation (Ascending and Descending).

Interestingly, Gene Ontology (GO) analysis using DAVID (da Huang et al., 2009) of genes transitioning to lower categories with erythroid differentiation, revealed a relative enrichment for genes involved in immune and early hematopoietic pathways, myeloid differentiation and immune response activation (data not shown), for example, Kit, Hhex, and Zfp36 genes. The GO analysis of genes transitioning to a higher category showed a relative enrichment in oxygen response pathways, chromatin organization and modification and cell cycle regulation (data not shown), which are all processes associated with mature erythroid physiology. Examples include the Slc4a1, Cat and Urod genes.

Overall, we find that genes representing the erythroid transcription program are highly enriched for GATA1 binding throughout differentiation. By contrast, reduced GATA1 binding during differentiation appears to be associated with genes implicated in early hematopoietic and alternative lineage programs.

1.3 Association of GATA1 with differential chromatin states

1.3.1 Epigenetic landscape of GATA1 target genes

In order to obtain a more global insight into the regulatory events taking place during erythroid differentiation, we integrated our GATA1 occupancy profiles with a series of genome-wide TF occupancy and histone modification profiles that were publicly available for Ter119^{-} and Ter119^{+} fetal liver cells at the time (Figure 1.8). As a result, a total of

ChIP Seq Target	Ter119-	Ter119+	Function	
GATA-1	+	+	Transcription Factors	
TAL1	+	+		
PU.1	+	+		
KLF1	+	+		
H3K27Ac		+	Chromatin Marks	Enhancer
H3K4me1		+		
H3K4me2	+	+		Activation
H3K4me3	+	+		
H3K9Ac	+	+		
H4K16Ac	+	+		
H3K36me3	+	+		Elongation
H3K79me2	+	+		
RNA Pol II	+	+		
H3K27me3	+	+		Silencing
DNA Methylation	+	+		
RNA Seq	+	+	Expression	
Total	16 (14 both conditions)			

Figure 1.8: Genomic features available for Ter119^{-} and/or Ter119^{+} fetal liver erythroid cells.

28 ChIP-seq datasets comprising 4 TFs, 9 histone tail modifications, RNA polymerase II, DNA methylation ratios and gene expression profiling by RNA seq (Kassouf et al., 2010; Kowalczyk et al., 2012; Pilon et al., 2011; Shearstone et al., 2011; Tallack et al., 2010; Wong et al., 2011; Wu et al., 2011) were incorporated into a single database. Importantly, with the exception of two datasets (H3K27Ac and H3K4me1) obtained from Ter119^{+} cells only, all other data were obtained from both Ter119^{-} and Ter119^{+} fetal liver erythroid cells (Figure 1.8). For all subsequent analyses, the TGS scores were based on the read density profiles produced for each experiment within a 10kb window around each gene's TSS (see Materials and Methods). As a first approach to characterize the epigenetic landscape of GATA1 occupied regions we calculated the pairwise Pearson correlation between TGS scores of GATA1 target genes and the TGS score calculated for each of the other TF occupancy profiles and epigenetic marks (Figure 1.9). Based on this analysis we observe that GATA1 occupancy strongly correlates with SCL/TAL1 binding ($R_{Ter119neg}$ = 0.53, $R_{Ter119pos}$ = 0.49), as has been previously reported (Kassouf et al., 2010), whereas a much weaker correlation

is observed with KLF1 occupancy profiles ($R_{Ter119neg}$ = 0.15, $R_{Ter119pos}$ = 0.07) and PU.1 ($R_{Ter119neg}$ = 0.1, $R_{Ter119pos}$ = 0.08), as also seen by Pilon et al. (2011). Furthermore, most of the histone modifications show a considerable correlation with GATA1 binding (Figure 1.9). Interestingly, GATA1 occupancy correlates highly with the levels of H4K16Ac mark in both early and late stages of erythroid differentiation ($R_{Ter119neg}$ = 0.49, $R_{Ter119pos}$ = 0.58) and with the levels of the enhancer related H3K27Ac and H3K4me1 marks (the latter were only available for Ter119$^+$ cells) ($R_{Ter119neg}$ = 0.54, $R_{Ter119pos}$ = 0.61 and $R_{Ter119neg}$ = 0.46, $R_{Ter119pos}$ = 0.5, respectively).

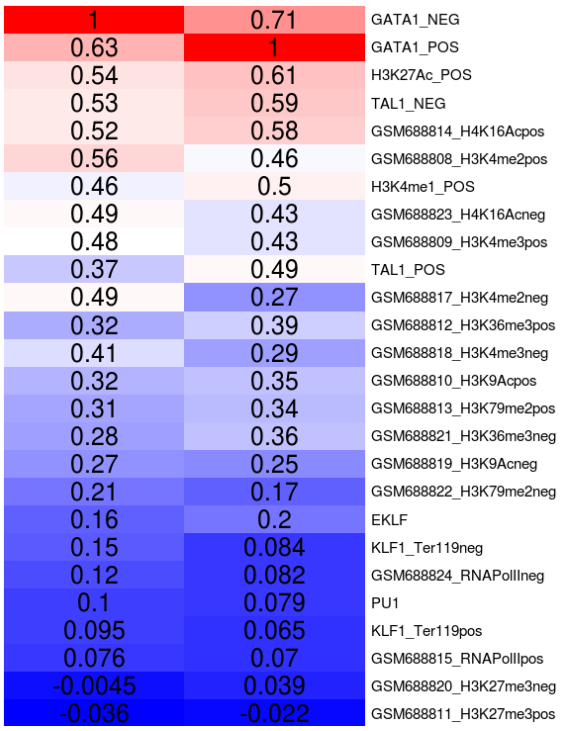


Figure 1.9: Heatmap showing the pairwise Pearson correlations between GATA1 TGS and epigenetic mark TGSs in Ter119$^-$ and Ter119$^+$ erythroid cells.

These data are consistent with the observations by Kowalczyk et al. (2012) showing that sequences enriched in H3K27Ac are predominantly bound by GATA1 (and other transcription factors) in erythroid cells. By contrast, we find a clear underrepresentation of H3K27me3 marks in genome-wide GATA1 occupied regions ($R_{Ter119neg}$ = -0.004, $R_{Ter119pos}$ = -0.02). Hence, the association of GATA1 binding with the H3K27me3 mark seen by Yu et al. (2009) in a subset of repressed GATA1 target genes in MEL cells does not appear to be reflected at the genome-wide level in fetal liver-derived erythroblasts. As previous studies have associated GATA1 with the acquisition of the H3K79 methylation mark (Steger et al., 2008) and with the formation of an erythroid specific histone H3 and H4 acetylation pattern in erythroid cells (Letting et al., 2003), we tested for a possible role for GATA1 in predicting the variation in specific histone modifications.

To this end, we used Random Forest based non-linear regression modeling (Breiman, 2001) to quantify the amount of variation in the levels of histone tail modifications between Ter119$^-$ and Ter119$^+$ cells that can be explained by GATA1 occupancy. A high degree of variation explained by GATA1 binding would provide an indirect indication of GATA1 modulating specific aspects of the epigenetic landscape in differentiating erythroid cells.

The performance of each model was quantified by the R^2 (% of explained variation) value calculated for each histone tail modification. Based on the results of these analyses GATA1 occupancy could be related, to varying extents, to the variation of all tested histone tail modifications, with changes in specific modifications being more predictable by GATA1 occupancy (Figure 1.10). Specifically, the regressors learned for the H3K79me2, H3K4me2,

Histone Tail Modification	GATA-1 R^2	All TFs R^2	Gain %
H3K4me2 (ΔTGS)	0.35	0.55	57.6
H3K4me3 (ΔTGS)	0.29	0.47	61.7
H3K9Ac (ΔTGS)	0.20	0.43	109.8
H4K16Ac (ΔTGS)	0.28	0.48	72.7
H3K36me3 (ΔTGS)	0.19	0.29	52.7
H3K79me2 (ΔTGS)	0.31	0.46	48.3
H3K27me3 (ΔTGS)	0.07	0.19	182.0
H3K9Ac (Ter119$^+$ TGS)	0.30	0.66	120.5
H4K16Ac (Ter119$^+$ TGS)	0.41	0.58	41.8
H3K27Ac (Ter119$^+$ TGS)	0.45	0.64	43.4
H3K4me1 (Ter119$^+$ TGS)	0.30	0.52	71.6
H3K4me2 (Ter119$^+$ TGS)	0.42	0.64	52.7
H3K4me3 (Ter119$^+$ TGS)	0.35	0.58	65.6

Figure 1.10: Percentage of variation explained (R^2) by the GATA1 (first column) and GATA1/TAL1/KLF1 (second column) trained RF regression models. Third column refers the percentage of the increase in R^2 values between the two models.

H3K4me3 and H4K16Ac histone marks resulted in a relatively high amount of variation being explained by the chromatin occupancy profiles of GATA1 in erythroid cells (Figure 1.11). Importantly, our results can be related to previous observations showing GATA1 associations with specific histone modifying enzymes such as CBP/p300, Dot1l and HDACs (Blobel et al., 1998; Steger et al., 2008; Watamoto et al., 2003).

1.3.2 GATA1 occupancy associates with variation of specific histone marks

The results obtained for the H3K79me2 modification (R^2=0.31) provide validation for our approach, as this modification is the only experimentally determined association between GATA1 occupancy and the deposition of a specific epigenetic mark (Steger et al., 2008). Additionally, GATA1 has been implicated in the tissue-specific acetylation pattern of histones H3 and H4 (Letting et al., 2003). Our results extend these observations in that GATA1 seems to preferentially associate with the H4K16 acetylation mark rather than H3K9 acetylation ($R^2_{H4K16Ac}$ = 0.28 and R^2_{H3K9Ac} = 0.20). Given the lack of data for genome-wide levels of H3K27Ac in Ter119$^-$ cells and in order to compare the involvement of GATA1 across all the acetylation events available in our dataset, we also produced regression models of the histone acetylation levels measured only in Ter119$^+$ cells. Interestingly, GATA1 occupancy

is a very good predictor for all 3 acetylation marks, with H3K27Ac showing the highest degree of correlation and H3K9Ac the lowest ($R^2_{H3K27Ac}$ = 0.45, $R^2_{H4K16Ac}$ = 0.41 and R^2_{H3K9Ac} = 0.30). These observations are in agreement with GATA1 interacting directly with the CBP/p300 acetyltransferase (Blobel et al., 1998), the latter having a specificity for acetylating both H4K16 and H3K27, but not H3K9 (Galvez et al., 2011; Jin et al., 2011). Additionally, GATA1 itself can be acetylated by the CBP/p300 complex, potentially altering its affinity for DNA (Boyes et al., 1998). The crosstalk between GATA1 acetylation and deposition of specific histone acetylation marks could provide the basis for the nucleation of different GATA1 complexes resulting in differential gene regulation between acetylated and non acetylated forms of the GATA1 protein.

In addition to the histone tail acetylation profiles, we also found a high correlation between GATA1 occupancy and the variation in methylation levels of histone H3. Notably, GATA1 seems to be associated more with the variation of the di-methyl mark rather than the tri-methylated lysine 4 ($R^2_{H3K4me2}$ = 0.35 and $R^2_{H3K4me3}$ = 0.29), again showing a preferential association of GATA1 with highly specific epigenetic events during erythroid differentiation. Interestingly, while H3K4me2 and H3K4me3 are highly correlated in terms of their localization at active genes, previous evidence suggests a tissue-specific regulatory role for H3K4me2 independently of H3K4me3 (Pekowska et al., 2010). In erythroid cells, Wong et al. (2011) described different dynamics between the two H3K4 methylation states, with increasing levels of H3K4me2 manifesting along the gene body of up-regulated genes, rather than near the gene TSS. In addition, genome wide studies in a multipotential hematopoietic cell line showed that H3K4me2 marks identify a population of lineage-specific transcriptionally poised genes (Orford et al., 2008). The H3K4 methylation state of these poised genes is regulated throughout erythroid differentiation of these cells, reflecting their developmental potential (Orford et al., 2008). Our results suggest a direct implication of GATA1 in the dynamic regulation of these poised genes. By contrast, and in accordance with the results obtained from the pair-wise correlations data above, GATA1 occupancy is a poor predictor of the variation observed in the levels of the H3K27me3 mark (R^2 = 0.07).

In order to assess whether the occupancies of the SCL/TAL1 and KLF1 transcription factors could provide additional information on a specific epigenetic mark's variation, we also built a second series of regression models by including the occupancy profiles of SCL/TAL1 and KLF1 (Kassouf et al., 2010; Pilon et al., 2011; Tallack et al., 2010; Wu et al., 2011) with those of GATA1 in the RF training datasets. We noticed a higher performance for all the regression models tested compared to GATA1 alone (Figure 1.10 and 1.11), suggesting that SCL/TAL1 and KLF1 may be involved together with GATA1 in modulating

epigenetic modifications. Importantly, the additional information derived from the inclusion of SCL/TAL1 and KLF1 occupancy varies within the different modifications.

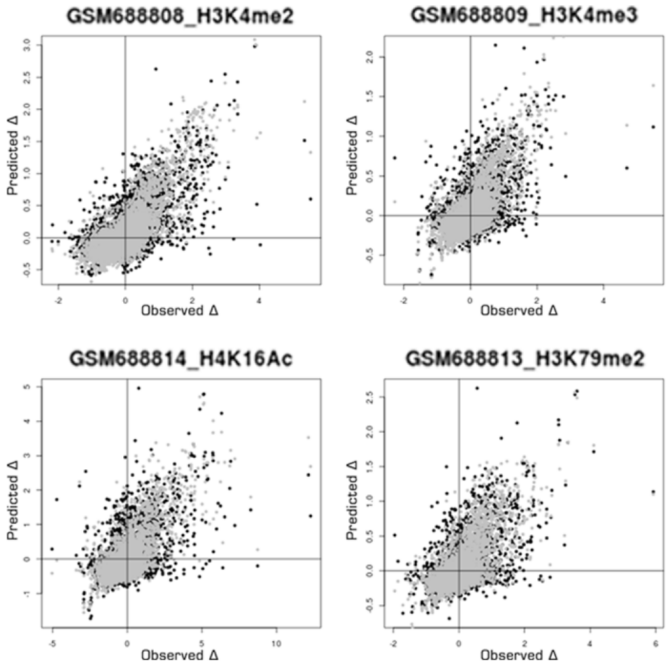


Figure 1.11: Scatterplots of observed and RF regression predicted values of selected histone mark variation between Ter119$^-$ and Ter119$^+$ cells. Black dots represent the predicted values of the GATA1 trained model, grey dots represent the values predicted by the GATA1/TAL1/KLF1 trained model.

The highest overall increase was observed for the H3K27me3 (182%) resulting in an R^2 value of 0.19, whereas the acetylation of H3K9 showed an increase of 101% resulting in an R^2 value of 0.43. These observations raise the possibility that distinct erythroid TF complexes are implicated in the deposition of specific epigenetic marks.

Overall, our results show that GATA1 is involved in the regulation of the epigenetic landscape of a large subset of target genes through the variation of specific epigenetic events and further suggest that GATA1 binding preferentially associates with specific histone tail modifications, such as H4K16 and H3K27 acetylation and H3K4 methylation.

1.3.3 Modeling gene expression of GATA1 gene targets

Of the 3,651 genes identified as GATA1 target genes, 321 genes are up-regulated by more than 2-fold with differentiation, 1941 genes are down-regulated by more than 2-fold and 1390 genes show a less than 2-fold variation between Ter119$^-$ and Ter119$^+$ erythroid cells. Since both GATA1 occupancy and the epigenetic landscape are involved in the regulation of GATA1 differentially expressed target genes (2258 genes), we integrated all of the available information (Figure 1.8) to model by RF non-linear regression (Breiman, 2001) the changes in their expression levels during erythroid differentiation. In doing so, we used as predictors a comprehensive database of all genome-wide TF occupancy and histone modification profiles presently available (28 in total) for Ter119$^-$ and Ter119$^+$ fetal liver cells.

Additional predictors included the distance between the point of maximum enrichment of each feature and the target gene TSS and, where possible, the difference in TGS scores between Ter119$^+$ and Ter119$^-$ cells. As a result, the final training dataset included a total of 62 features for both Ter119$^-$ and Ter119$^+$ determined values. The prediction model produced by this approach resulted in a remarkably high proportion (62%, r = 0.8, Figure 1.12a) of variation in gene expression being explained by the binding signals of the 4 TFs, 9 histone modifications, RNA pol II and DNA methylation levels measured in Ter119$^-$ and Ter119$^+$ cells. An important feature of RF is that it provides an internal measure of variable importance that can be used to rank variables. Ranking is based on the percentage of increase in mean square error (MSE) when a specific variable is randomized (%IncMSE) (Figure 1.12b).

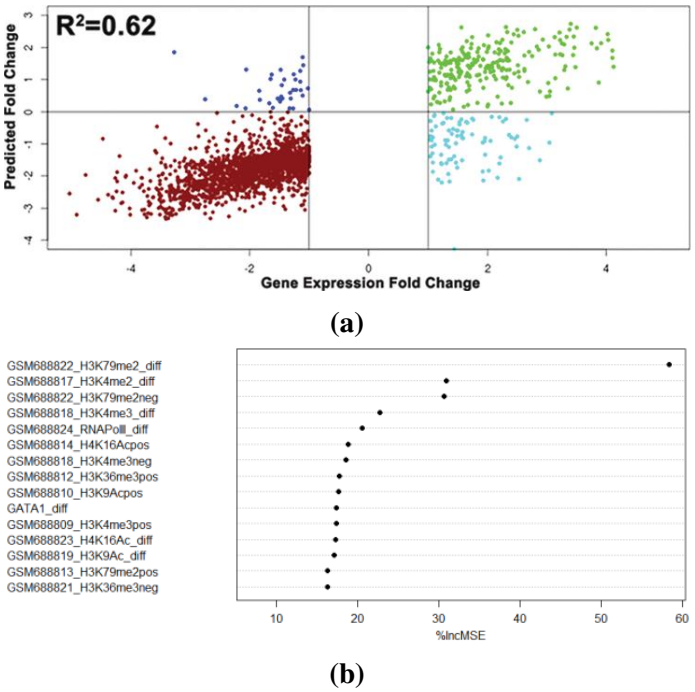


Figure 1.12: Modeling differential expression of GATA1 target genes. **a)** Scatterplot of observed and predicted values of gene expression change between Ter119$^-$ and Ter119$^+$ cells of differentially regulated GATA1 target genes. Red and green dots refer to values correctly predicted as decreasing or increasing mRNA levels, respectively. Blue and cyan dots refer to GATA1 target genes with inverted predicted values. **b)** Variable importance measures (%IncMSE) in predicting gene expression changes between Ter119$^-$ and Ter119$^+$ erythroid cells. Only the 15 top ranked features are plotted.

Based on the RF ranking, the most predictive feature of gene expression variation in erythroid differentiation is the variation in the levels of the H3K79me2 elongation mark, in accordance with the findings of Wong et al. (2011). The variation of H3K4 methylation levels closely followed, whereas variation in GATA1 occupancy was found to be in a group of almost equally important features comprising H3K9Ac, RNApolII and H4K16Ac. It is interesting to note that the most predictive features (H3K79me2 and H3K4 methylation) can be, at least in part, associated with GATA1 itself, as shown above. This observation further consolidates the notion that part of the GATA1 regulatory function is exerted through the modulation of the epigenetic landscape of its target genes. Another important feature of RF regressions is the ability to calculate proximity values which represent the degree of

similarity between sample points. In our context, the proximity value reflects the similarity of two genes (sample points) based on TF binding and enrichment in specific epigenetic marks.

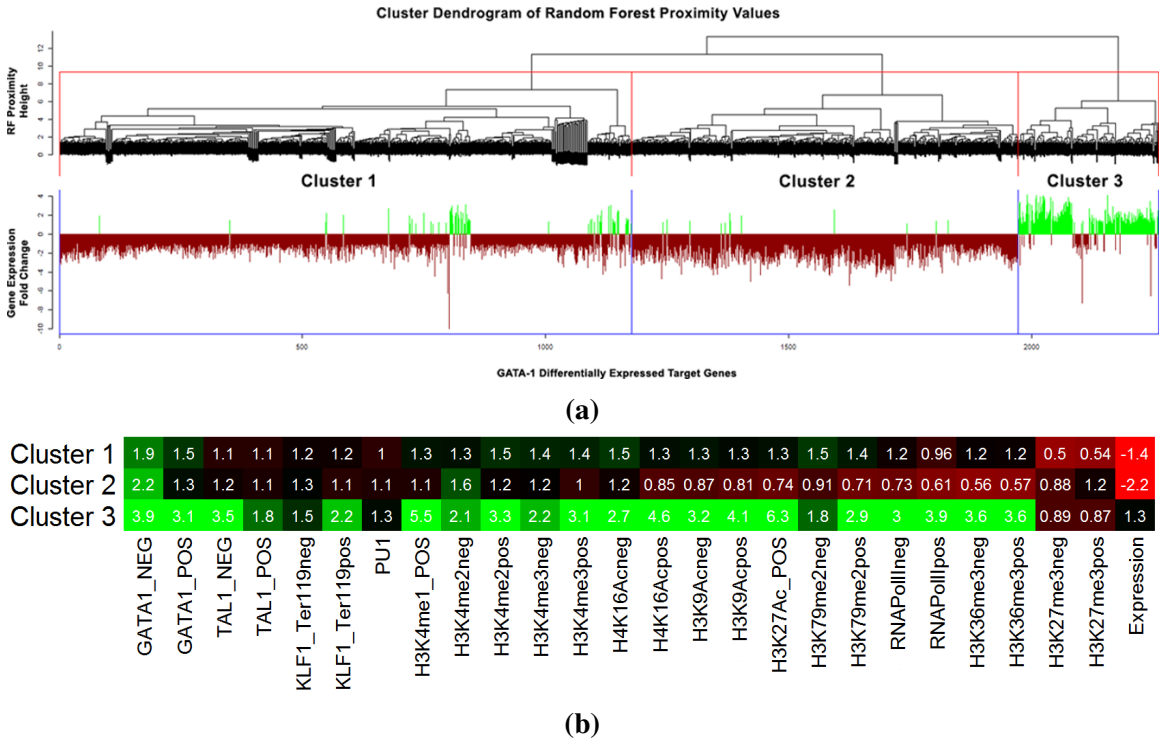


Figure 1.13: Evaluation of the gene expression RF regression model. **a)** Dendrogram showing clusters of GATA1 differentially regulated target genes according to the RF calculated proximity values. Under the dendrogram the corresponding gene expression fold change values of the proximity clustered GATA1 target genes are plotted. **b)** Heatmap illustrating the mean TGS values of the different occupancy profiles within the genes composing each cluster of GATA1 differentially regulated target genes.

In order to identify clusters of genes bearing similar epigenetic profiles, we converted the RF based proximity matrix into a euclidean distance matrix and then performed hierarchical clustering using Ward's method (Figure 1.13a). The first observation is that the proximity values calculated by the RF regression model efficiently separated up-regulated from down-regulated genes. Additionally, the branch corresponding to the down-regulated genes is further dissected in two distinct gene clusters. In order to assay for any functional distinction between these 3 clusters we performed GO analysis using the DAVID online tool (da Huang et al., 2009). Not surprisingly, the up-regulated genes' cluster (Cluster 3, 309 genes) showed a very high enrichment for all the heme biosynthetic processes and erythrocyte differentiation (Figure 1.14) and contained genes like α and β-globin, Scl/Tal1, Slc4a1 and Alas2. Interestingly, the analysis of the two clusters associated with down-regulated genes revealed

clearly distinct functional properties. Cluster 1 (1,186 genes) was highly enriched for genes involved in RNA processing, the translation machinery and ribosome biogenesis, whereas Cluster 2 (763 genes) was enriched in genes involved in hematopoiesis, immune system development, myeloid and lymphoid cell differentiation and cell proliferation and included genes like PU.1, c-Kit, Lyn, Cebp, Hif1a, Runx1 and members of the Stat and Smad families.

Cluster 1	
Term	Pvalue
RNA processing	1.2e-17
mRNA processing	8.3e-10
mRNA metabolic process	9.6e-10
translation	6.1e-9
RNA splicing	1.9e-7
ncRNA processing	3.2e-7
ribonucleoprotein complex biogenesis	9.2e-6
ribosome biogenesis	1.4e-5

Cluster 2	
Term	Pvalue
hemopoiesis	2.8e-7
immune system development	2.9e-7
hemopoietic or lymphoid organ development	9.7e-7
regulation of cell proliferation	2.7e-5
myeloid cell differentiation	1.1-4
cell activation	1.4e-4
myeloid leukocyte differentiation	3.3e-4
embryonic hemopoiesis	4.4e-4

Cluster 3	
Term	Pvalue
tetrapyrrole biosynthetic process	1.5e-12
porphyrin biosynthetic process	1.5e-12
porphyrin metabolic process	2.2e-12
tetrapyrrole metabolic process	2.2e-12
heme biosynthetic process	2.6e-10
erythrocyte differentiation	7.2e-6
erythrocyte homeostasis	1.1e-5
negative regulation of apoptosis	0.001

Figure 1.14: Most highly enriched Gene Ontology terms identified for each of the three GATA1 target gene clusters.

Significantly, the three clusters show distinct epigenetic signatures (Figure 1.13b).

Cluster 3: Erythroid specific genes

Cluster 3 (up-regulated genes) shows the highest levels of GATA1, SCL/TAL1 and KLF1 occupancy and also the highest levels of the activating and elongating histone marks. By contrast, Cluster 2, enriched in down-regulated genes involved in alternative hematopoietic lineages, was associated with the highest levels of H3K27me3 histone modification and the lowest enrichment levels for all three TFs and activating and elongating histone marks. It is of interest that the majority of the GATA1 target genes found by Yu et al. (2009) to also display H3K27me3 marks, e.g. GATA2, c-kit etc., partition within this cluster.

Cluster 1: Housekeeping genes

Cluster 1, enriched in down-regulated genes involved in house-keeping processes, is characterized by the lowest levels of H3K27me3, low levels of TF occupancy and intermediate to low levels of activating and elongating marks. Even though Cluster 1 is composed of genes that exhibit decreasing mRNA levels, the lack of the H3K27me3 mark and the persistence of activating and elongating marks could be an indicator of gene expression that is maintained at low levels, or in the process of being extinguished.

Cluster 2: Alternative blood lineages associated genes

By contrast, genes composing Cluster 2 show a more severe down-regulation with high levels of the Polycomb-Group H3K27me3 repressive mark, suggesting that a repressive epigenetic memory mechanism is in place. In fact, if we compare the absolute mRNA levels through stages R2 to R5 of erythroid differentiation (Wong et al., 2011), we find significantly lower mRNA levels for Cluster 2 genes (alternative lineages) compared to Cluster 1 (protein production) ($P < 2.2e\text{-}16$, Wilcoxon rank-sum test).

Our findings significantly extend previous observations made by Cheng et al. (2009) using a limited number of GATA1 target genes in G1E cells (GATA1-null proerythroblast cell line, (Rylski et al., 2003)), in which they divided repressed genes in two classes: one enriched in H3K27me3 marks and depleted for SCL/TAL1 binding and the second class depleted for H3K27me3 marks and enriched for SCL/TAL1 binding.

Summary

Collectively, our data reinforce previous observations for GATA1 regulating the erythroid differentiation process at multiple levels (reviewed in Hattangadi et al. (2011)).

Firstly, GATA1 positively regulates the expression of erythroid specific genes and genes involved in the production of mature hemoglobin molecules.

Secondly, it negatively regulates the expression of genes involved in early hematopoietic differentiation and alternative myeloid and lymphoid lineages, by completely shutting them down to allow terminal erythroid differentiation to proceed.

Thirdly, it is directly involved in the reduced expression of the mRNA maturation and translation machinery adjusting it to the reduced needs of the enucleated mature erythrocyte. Importantly, our work shows that specific epigenetic signatures are associated with functionally different subsets of GATA1 target genes, thus suggesting a degree of plasticity in the regulatory functions of GATA1.

Chapter 2

Analysis of erythroid lineage specification by computational integration of genomic data

Abstract

Commitment of hematopoietic stem cells into different hematopoietic lineages requires precise transcriptional and epigenetic regulatory events in order to produce the required diversity of terminally differentiated blood cell populations. In the previous chapter we focused on the different roles of the GATA1 transcription factor during terminal erythroid differentiation. Our approach involved the integration of a comprehensive set of genomic and transcriptomic observations to identify differential regulatory patterns of GATA1 function. In this chapter we applied a similar approach in order to identify lineage specific epigenetic patterns emerging during the lineage specification process of the erythroid and of the megakaryocytic lineages. Our approach results in the production of highly structured gene wide distribution patterns of chromatin features (NGS datasets), allowing for the direct inspection and comparison of several chromatin features at high (single gene locus) resolution, greatly simplifying the interpretation of the results.

Using this approach, we identify a large group of active gene promoters in HSCs specifically transitioning to an inactive promoter state during the erythroid specification process. Importantly, the active promoter state of these genes is maintained in the megakaryocytic specification process. Furthermore, comparison of the epigenetic signatures in mouse and human erythropoiesis revealed both similar and distinct characteristics.

2.1 Modeling differential gene expression between erythroid and megakaryocytic lineages

In order to compare the transcriptional, epigenetic and transcription factor occupancy profiles of the erythroid and megakaryocytic lineages, we used a comprehensive set of genomic and transcriptomic NGS (next generation sequencing) datasets produced by the mouse Encode consortium for mature erythroid (Ter119$^+$) and megakaryocytic (CD41$^+$) cells, in addition to datasets available for mouse hematopoietic stem cells (mHSCs, LSK), proerythroblasts (Ter119$^-$) and total bone marrow cells (BM) (Figure 1A). These include genome wide profiles for several histone tail modifications, for DNase I hypersensitive sites, for transcription factor occupancies including those for GATA1, TAL1, LDB1 and GATA2, for RNA polymerase II and RNA expression profiles (Figure 2.2a, and Table 3.2a). We developed a machine learning modeling and clustering approach to integrate the transcriptomic and genomic profiles of the three distinct hematopoietic populations (LSK, Ter119$^+$ and CD41$^+$) in uncovering differential patterns which may provide clues as to the regulatory mechanisms underlying the commitment and differentiation of the two lineages, i.e. erythroid cells and megakaryocytes, stemming from a common progenitor, the megakaryocytic-erythroid progenitor (MEP). To these ends, we trained a Random Forest (RF), (Breiman, 2001) regression model to predict differences in gene expression levels based on the epigenetic profile of their broader (± 25kb) promoter region in each cell population.

2.1.1 Optimization of RNAseq and ChIPseq signal values

In order to accurately define our training dataset we tested a series of differential expression analysis algorithms for RNAseq data, namely, Cufflinks, edgeR and Deseq (Anders and Huber, 2010; Robinson et al., 2010; Trapnell et al., 2010). The three algorithms returned significantly different results with the number of identified differentially expressed genes varying from ~2000 (edger, deseq) to ~3500 (Cufflinks) genes, using a 2-fold cut-off in differential expression levels between the erythroid and megakaryocytic lineages.

To identify the most accurate result we chose to maximize the consensus between expression and epigenetic data, applying a mathematical prediction based validation approach of the results produced by each algorithm. To this end we trained a series of non-linear RF-based regressors of gene expression fold change within the top 750 upregulated and downregulated genes (total 1500 genes in each case), as defined by either Cufflinks, DESeq and edgeR. We used two different series of predictor sets, one using only TF binding profiles and one using only histone tail modification profiles to avoid any bias introduced by the

selection of the chromatin features included in the training sets (Table 3.1). Given the fact that each predictor series was maintained constant in each differential expression regressor any difference observed in the result will depend solely on the accuracy of the Fold Change measure reported by each differential expression algorithm.

As a result, the differential expression levels produced by the DESeq algorithm resulted in the highest coefficient of determination for both the TF and histone based approach, thus showing a better agreement with the epigenetic and regulatory variation observed in the identified differentially expressed gene loci (Figure 2.1).

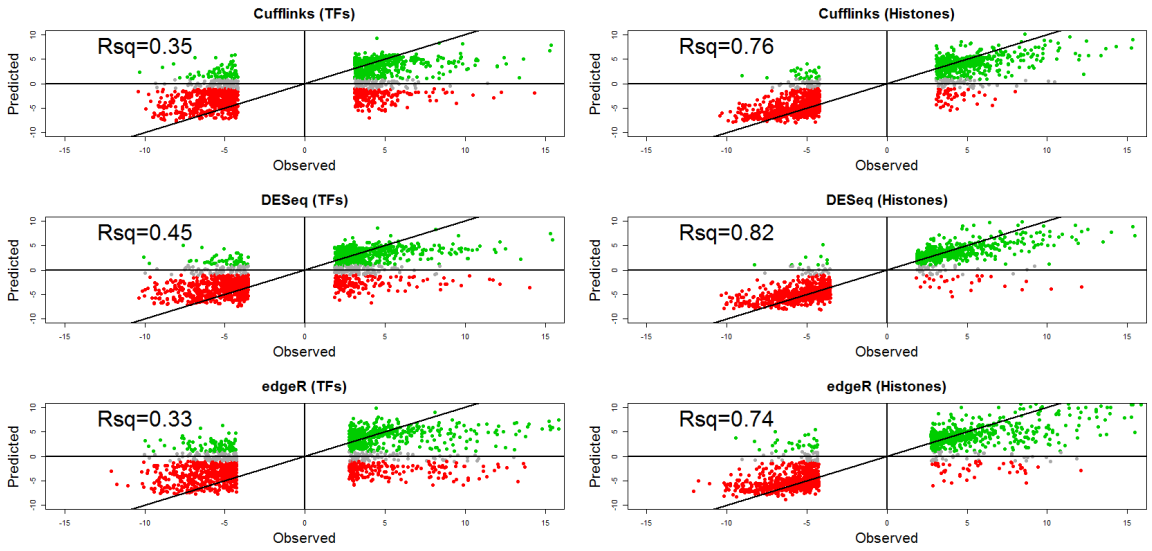


Figure 2.1: Scatterplots of observed versus RandomForest regressors predicted values of different RNAseq analysis algorithms based on TF occupancies (left) or histone tail modification profiles (right). The R^2 value achieved by each different regressor is reported in the upper left corner of each plot as Rsq.

Following this validation approach we trained a non-linear RF based regression model of all differentially expressed genes between the erythroid and megakaryocytic lineages using a 4-fold cut-off in gene expression and including both TF and histone tail modification ChIP signals as predictors. Importantly, our model reached a highly accurate coefficient of determination of over 80% (R^2 = 0.82, Figure 2.2b).

2.1.2 Visualization of the results

In order to identify distinct epigenetic signatures of erythroid and megakaryocytic-specific genes, we clustered the proximity values calculated by the RF model for the ~1600 genes that are differentially expressed between the two lineages. Proximity (or similarity) values

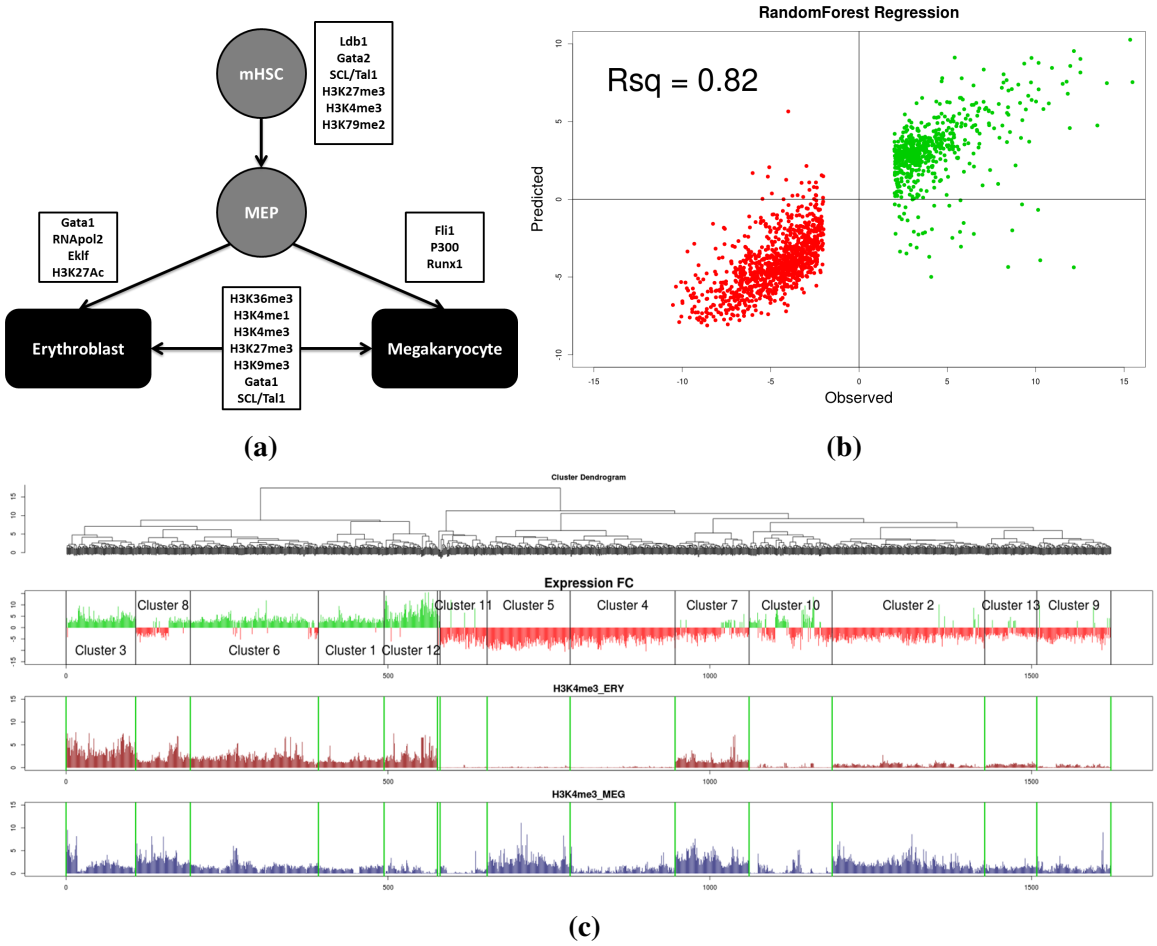


Figure 2.2: RF model of erythroid vs megakaryocyte specific gene expression based on genomic data **a)** NGS Datasets Available in the MegakaryocyteErythroid (MegE) differentiation branch. **b)** Performance of the RF based prediction model. **c)** Comparison of the H3K4me3 histone tail modification profiles in erythroid and megakaryocytic cells.

calculated during the RF training step quantify the overall similarity between all genes included in the training set and reflects both the expression and the epigenetic information available for each gene. We employed this similarity measure to produce a hierarchical clustering of all genes under investigation. We used an internal validation method (Dunn index) to identify the optimal number of clusters characterizing the dataset. Based on this approach, we identified 14 distinct gene clusters within this set of differentially expressed genes (Figure 2.2c).

Given the increased complexity resulting from the large number of clusters identified and also the fact that a 'heatmap' approach would reflect the mean values for each cluster, we opted for a more human readable visualization format. To do this we plotted the total gene score (TGS; defined as the sum of the ChIPseq peak scores assigned to it) of each gene based

on the hierarchical clustering order. This approach produces a much higher resolution for each dataset, whereby the ChIPseq signal of each gene is represented by a single vertical line, the height of which corresponds to its TGS (Figure 2.2c).

2.2 Chromatin state variation during terminal erythroid differentiation

As expected by the very high coefficient of determination achieved by our model, the hierarchical clustering of the similarity values clearly separated the erythroid specific genes (green vertical lines, positive log2(FC)) from the megakaryocyte specific genes (red vertical lines, negative log2(FC)) (Figure 2.2c, 'Expression FC' panel).

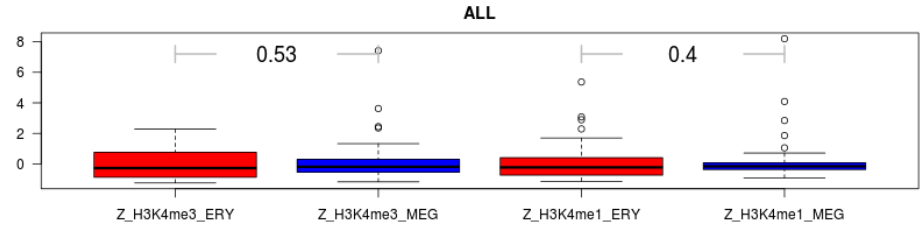
In trying to characterize the epigenetic signatures of each cluster we first compared the H3K4me3 status, an epigenetic mark for active promoters, in erythroid and MK cells. Interestingly, genes with megakaryocytic specific expression display some heterogeneity in their epigenetic patterns.

2.2.1 Loss of active promoter epigenetic state in erythroid lineage differentiation

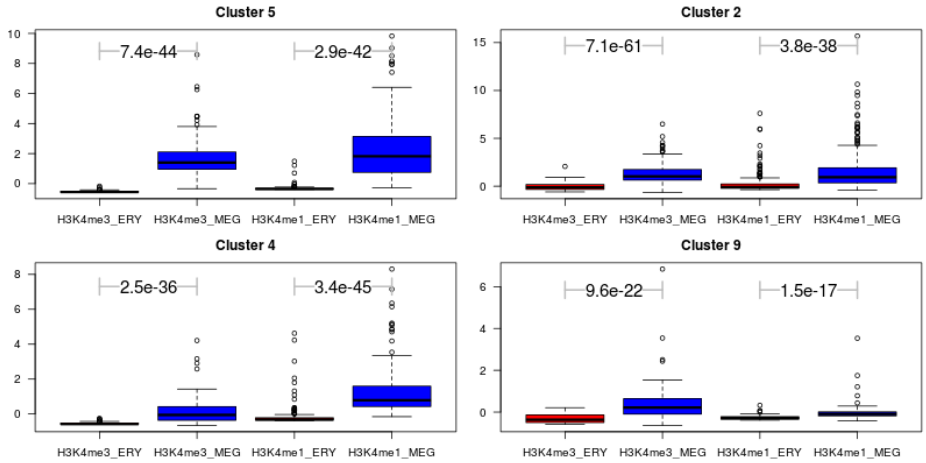
More specifically, Cluster 5 (129 genes) is characterized by a distinct absence of H3K4me3 and H3K4me1 histone modification marks specifically in erythroid cells, whereas the same genes are enriched for H3K4me3 and H3K4me1 marks in megakaryocytes (Figure 2.2c). A similar pattern, though to a lesser extent, can be seen in Clusters 2, 4 and 9 (237, 163 and 115 genes, respectively Figure 2.3b).

Clusters with erythroid specific expression include clusters 3, 12, 6 and 1 (108, 83, 199 and 102 genes, respectively). Interestingly, none of these clusters show a clear absence of active marks in megakaryocytes, but rather they appear to gain in H3K4 mono- and tri-methylation and in H3K36me3 marks in erythroid cells (Figure 2.3c).

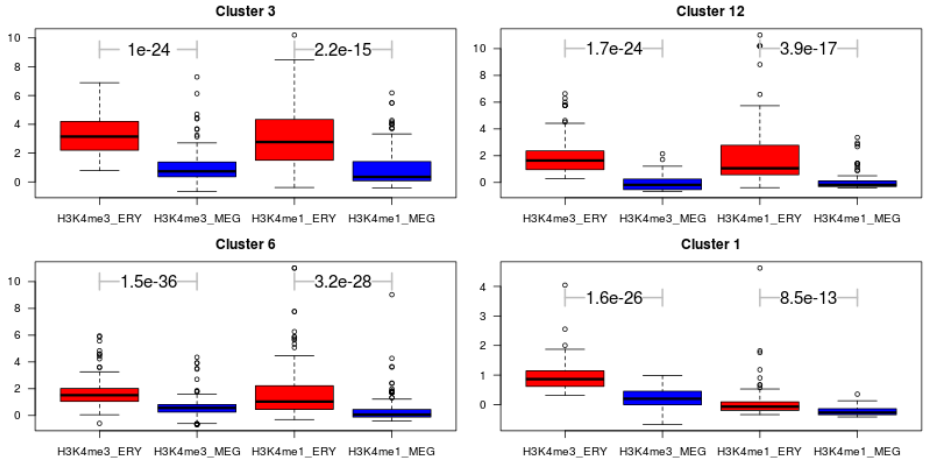
Functional analysis by EnrichR (Chen et al., 2013) of the selected clusters identified Cluster 5 and Cluster 3 as the main megakaryocytic and erythroid specific gene subsets, respectively (Figure 2.4). More specifically, genes composing Cluster 5 are highly enriched in hemostasis, blood coagulation and platelet activation Gene Ontology terms and include megakaryocytic specific genes such as Itgb3 (CD61), Pdgfb and Plek. Genes composing Cluster 3 are highly enriched in heme biosynthetic and metabolic Gene Ontology terms and include erythroid specific genes such as Hemgn, Alad and Fech.



(a) All Clusters (full dataset)



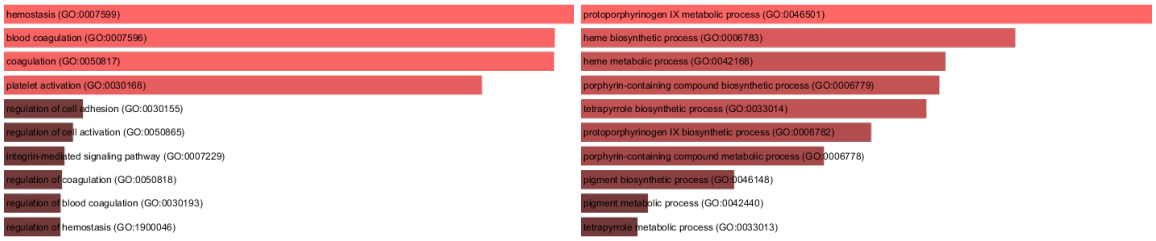
(b) MK specific clusters



(c) Erythroid specific clusters

Figure 2.3: Differential distribution of H3K4me1/me3 histone modification marks between erythroid (red) and megakaryocytes (blue) in selected gene clusters. P values are calculated using a two-sided Wilcoxon rank sum test between the TGS distribution of each mark among the genes of each cluster.

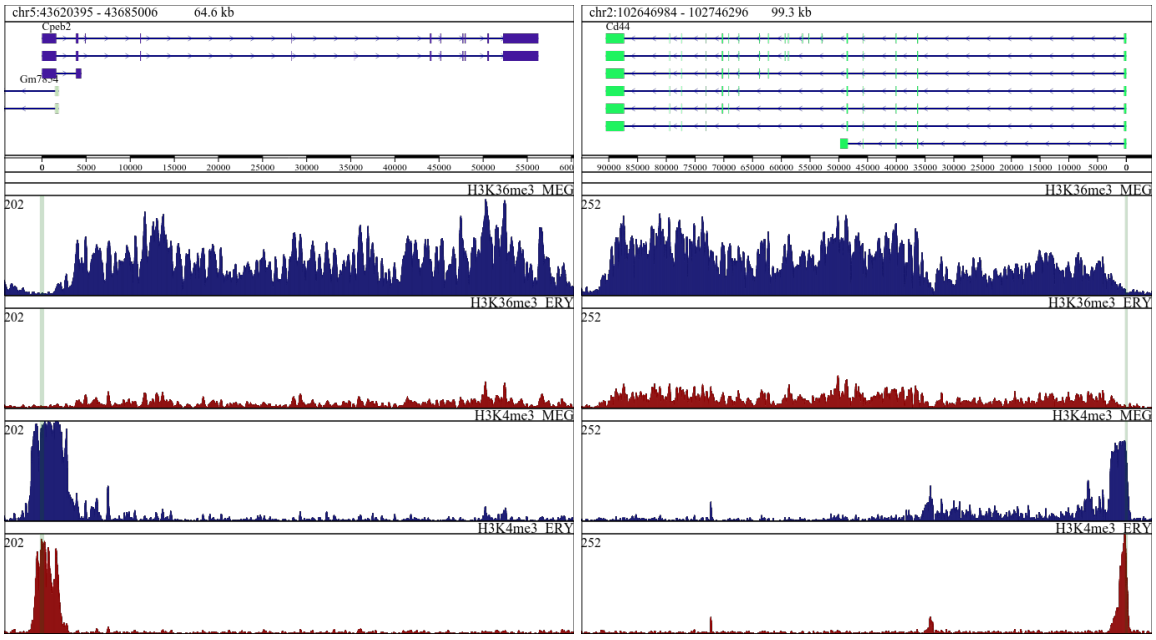
Cluster 7 is also of interest as genes in this cluster display high levels of the H3K4me3 mark in both erythroid and MK cells. However, these genes do not appear to be transcribed in erythroid cells, as evidenced by the very low levels of H3K36me3, a mark of transcriptional elongation (Figure 2.5 shows two representative examples,the Cpeb2 and CD44 *loci*).



(a) Cluster 5 (129 genes) **(b)** Cluster 3 (108 genes)

Figure 2.4: Megakaryocytic and erythroid specific functional signatures of Cluster 5 and Cluster 3.

Thus, using our approach we have identified distinct clusters of differentially expressed genes in erythroid and megakaryocytic cells with clearly distinguishable epigenetic signatures.



(a) Cpeb2 **(b)** Cd44

Figure 2.5: Representative examples of Cluster 7 epigenetic signature. Read density profiles refer to H3K36me3 (elongation mark) and H3K4me3 (active promoter mark) in erythroid (red) and megakaryocytic (blue) cells.

Expanding the model to non differentially expressed genes

To further explore the extent of the different H3K4me3 signatures and to eliminate any bias introduced by the strict differential expression thresholds used in the selection of genes included in the modeling step, we expanded our analysis to include all genes expressed in the

erythroid and/or megakaryocytic population (0 > log2(FC) > 0, ~12000 genes). Despite the lower coefficient of determination achieved by the gene wide regression model (0.55 and 0.82 for all expressed and differentially expressed genes, respectively), hierarchical clustering of the new proximity values still produced a highly structured profile for the H3K4me3 signal in each population (Figure 2.6, panels 4-6). More importantly, we could still identify a clear difference between the erythroid and megakaryocytic H3K4me3 profiles, as exemplified by Cluster 10, thus expanding the number of differentially marked genes from 129 to 978 (Figure 2.6, highlighted area).

The difference in the H3K4me3 signal identified between the erythroid and megakaryocytic lineages could be due to either *de novo* deposition of the H3K4me3 mark in these gene *loci* in MKs or due to specific H3K4me3 de-methylation in erythroid cells. To distinguish between these two possibilities, we again applied our approach in analyzing the gene-wide distributions of the H3K4me3 mark in HSCs and comparing them to those obtained in the erythroid and MK lineages. As can be seen in Figure 2.6, panels 3-6, the H3K4me3 profile of cluster 10 genes in HSCs closely resembles that obtained for this cluster in MKs. This also appears to be the case for clusters 3 and 7 (Figure 2.6). On the basis of this evidence, we suggest that the differences observed in the H3K4me3 marks between the erythroid and MK lineages are due to a highly specific loss of H3K4me3 marks in the erythroid lineage rather than their *de novo* acquisition in the MK lineage. It is also of interest to note that the overall genome-wide profile of H3K4me3 marks in HSCs closely resembles that of the megakaryocytic lineage (Figure 2.6, compare panels 3 and 4). In order to identify the stage at which the H3K4me3 marks are erased in gene clusters 10, 3, 5 and 7 in the erythroid lineage, we compared the H3K4me3 signatures relative to DNase hypersensitivity profiles during erythroid commitment and differentiation. The promoter regions of genes in these clusters are characterized by an open chromatin configuration in cKit$^+$/CD71$^-$ cells (LSK), followed by a progressive loss of overall chromatin accessibility which becomes evident as early as the proerythroblast (Ter119$^-$) stage (Figure 2.6 and 2.7). These observations suggest that the erythroid-specific loss of the H3K4me3 marks in genes

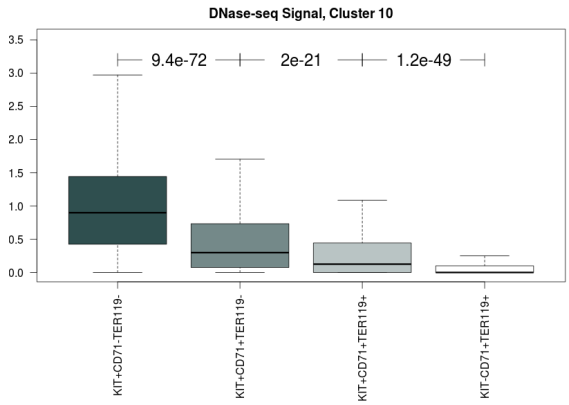


Figure 2.7: Progressive loss of overall chromatin accessibility within Cluster 10 gene promoters. (Wilcoxon test of sequential erythroid differentiation DNase-seq signal.)

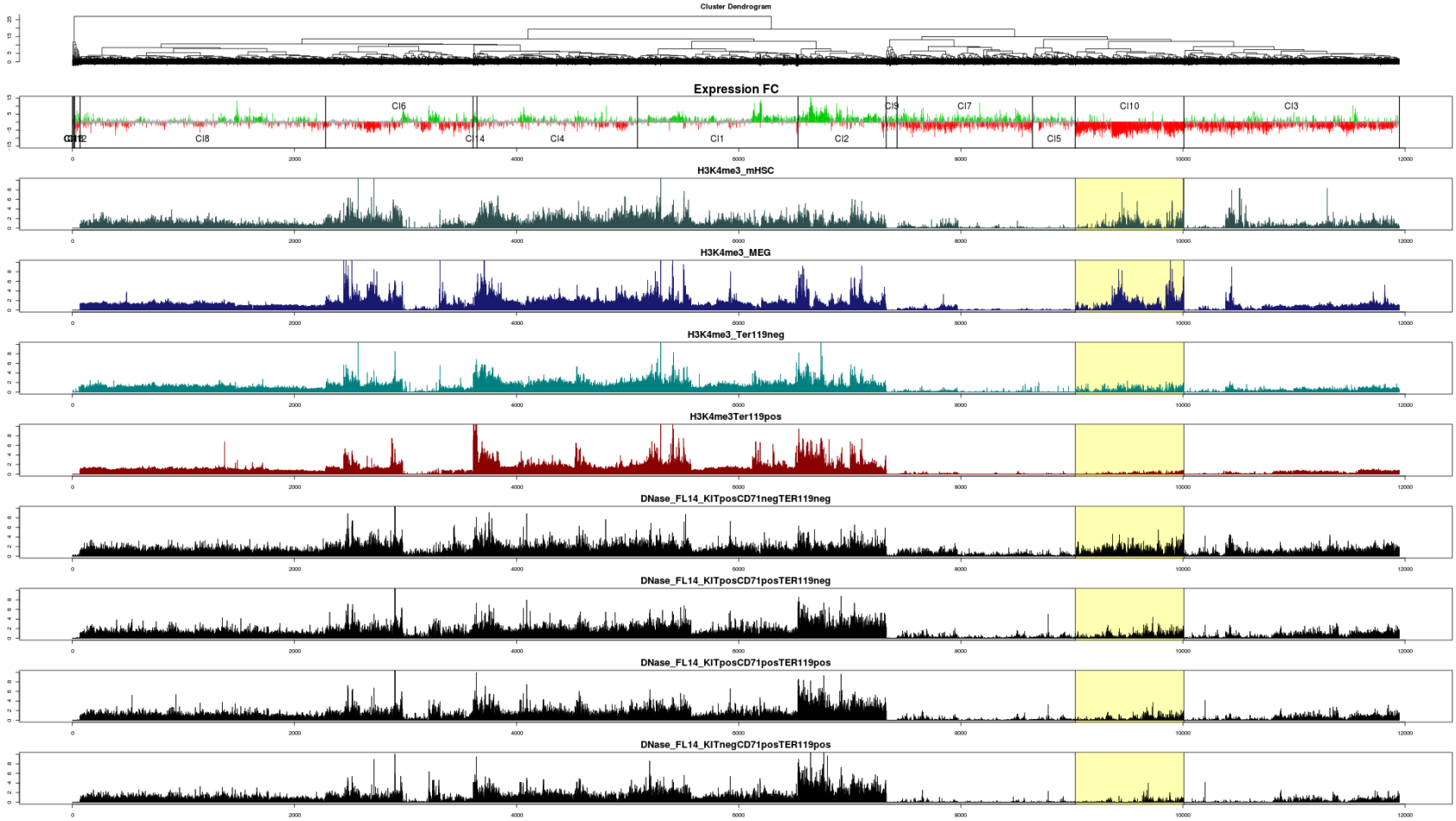


Figure 2.6: Progressive loss of H3K4me3 and overall chromatin accessibility with erythroid commitment and differentiation process. Hierarchical clustering of genomic similarity of genes expressed in erythroid and/or megakaryocytic lineages. Panels represent (Top to bottom): Dendrogram, ERYvsMEG Gene Expression FoldChange, H3K4me3 enrichment in mHSC, Megakaryocytes and Erythroid (Ter119^{-} and Ter119^{+}), DNase hypersensitivity in sequential erythroid differentiation stages (KIT^{+}CD71^{-}Ter119^{-}, KIT^{+}CD71^{+}Ter119^{-}, KIT^{+}CD71^{+}Ter119^{+}, KIT^{-}CD71^{+}Ter119^{+}).

within these clusters initiates prior to the emergence of CD71$^+$ cells, which represent the first stage of a committed erythroid cell population.

In conclusion, several lines of evidence suggest that one of the main characteristics that drive erythroid cell differentiation and specification is the loss of active state promoter marks in ~1000 genes.

2.2.2 Identification of potential erythroid specific epigenetic modifiers

In order to identify potential epigenetic factors with an active role in the chromatin reconfiguration differences between erythroid and megakaryocytic cells identified above, we interrogated a recently described comprehensive set of epigenetic modifiers (Huang et al., 2013). More specifically, based on RNAseq profiles (FPKM values) we searched for genes that are expressed in the erythroid cell population, at the same time presenting with very low mRNA levels in the MEP and megakaryocytic cell populations (Figure 2.8). Furthermore, erythroid specific expression of the identified genes was assessed by examination of the epigenetic profiles of their gene loci, and more specifically of the H3K4me3 (active promoter) and H3K36me3 (transcript elongation) profiles of the promoter and gene body, respectively (Figure 2.9).

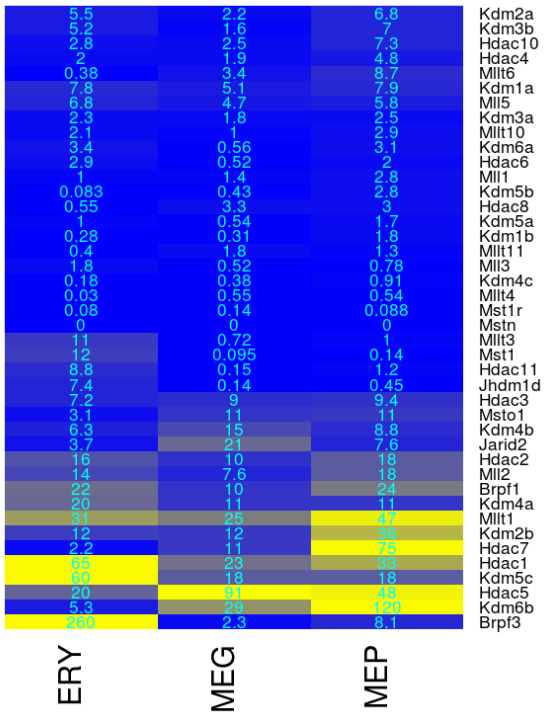


Figure 2.8: Identification of potential erythroid specific modifiers through comparison of their absolute expression levels in common progenitor and differentiated cells.

Based on this approach we identified 5 epigenetic modifiers with erythroid specific expression and epigenetic (regulatory) profiles (Figure 2.9). Among the selected genes we identify proteins with previously described functions in erythroid and megakaryocytic specification and differentiation, such as Mllt3 (Pina et al., 2008; Zini et al., 2012) and Mst1 (Nejigane et al., 2013), thus validating our approach. In addition, a novel set of epigenetic modifiers was identified bearing highly similar epigenetic and expression profiles, which included the lysine demethylase Jhdm1d/Kdm7a, the histone deacetylase Hdac11 and the ATP-dependent chromatin remodeling protein Cecr2. Interestingly, a similar approach applied in the identification of

megakaryocytic specific modifiers led to the identification of a single gene, Phf21a (Figure 2.9f), encoding a chromatin remodeling protein which has been shown to interact with and inhibit the action of the Lsd1 demethylase (Shi et al., 2005).

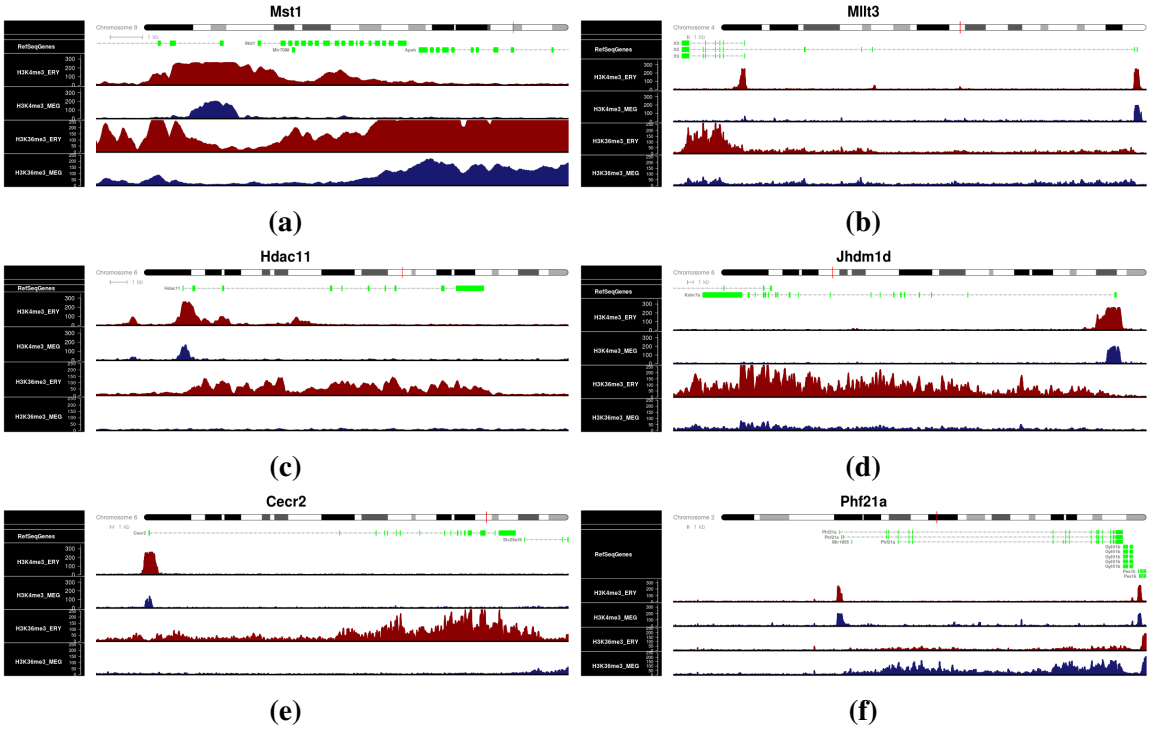


Figure 2.9: Epigenetic landscape of epigenetic modifiers with erythroid **(a-e)** or megakaryocytic **(f)** specific expression patterns. Read density profiles refer to active promoter mark H3K4me3 and transcription elongation mark H3K36me3 in erythroid (red) and megakaryocytic (blue) cells.

2.3 Lineage specific Transcription Factor binding profiles

We next focused on the gene-wide distributions of specific hematopoietic transcription factors (TFs) in the LSK, Ter119$^-$,Ter119$^+$ and CD41$^+$ cell populations. To allow for a clearer picture of TF-specific clustering we restricted our training set by excluding all histone modification profiles (Table 3.2b). This approach led to a relatively low coefficient of determination (0.25) but, nevertheless, produced very good clustering results (Figure 2.10). The low coefficient of determination could be explained by the fact that TF binding profiles are predictive of the direction of expression (erythroid or megakaryocytic specific genes) of only their target genes, which represent a much lower fraction of genes bearing widely distributed chromatin modification marks such as H3K4me3.

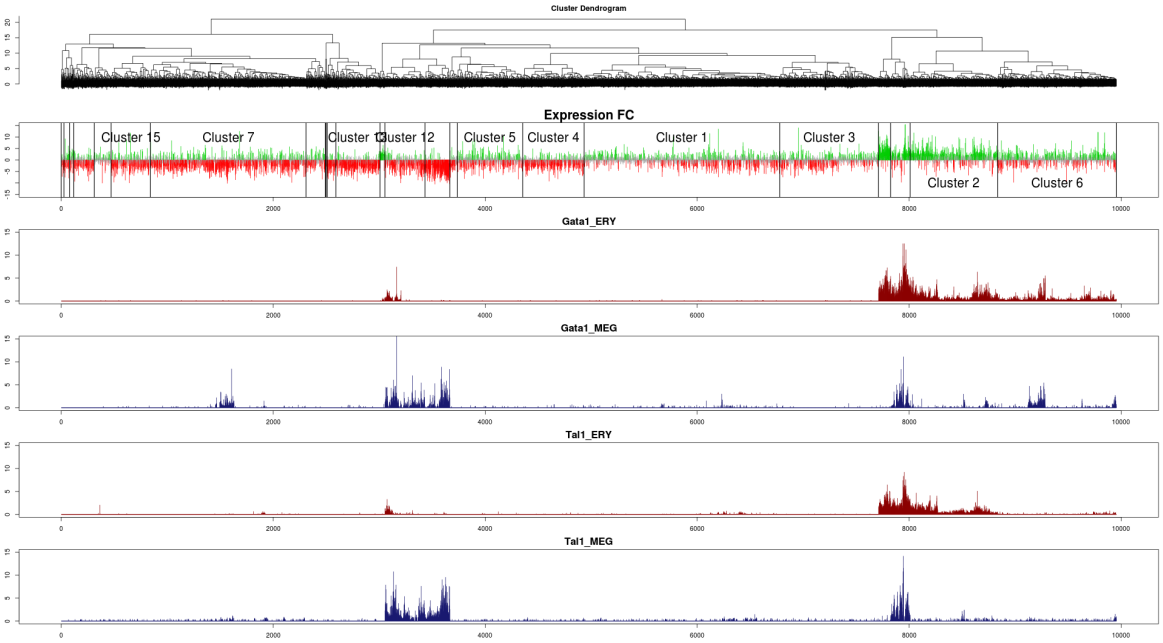


Figure 2.10: Distinct erythroid and megakaryocytic distribution of GATA1 and SCL/TAL1

2.3.1 Comparison of GATA1 profiles in erythroid and MK lineages

Based on this analysis we find that the GATA1 binding profiles appear to be largely distinct between the erythroid and megakaryocytic lineages, with clear quantitative differences even in regions of overlap, for example clusters 2, 6, 9 and 17 (Figure 2.10). Furthermore, the binding profile of the GATA1 interacting partner SCL/TAL1 in erythroid or megakaryocytic cells to a large extent, though not completely, mirrors that of GATA1 in these lineages.

Notably, there are clusters of genes in both erythroid and MK cells which are bound by GATA1 but not by SCL/TAL1, for example, clusters 6 and 12 in erythroid cells and clusters 6 and 7 in MK cells. This most likely reflects the binding of GATA1 with other protein partners such as FOG-1 in erythroid cells or RUNX1 in MK cells (REFS), providing further evidence for distinct TF subcomplexes.

2.3.2 Upstream specification of erythroid GATA1 binding signature

One question arising from these observations is how the differential binding profiles for GATA1 and TAL1 are established in erythroid versus megakaryocytic lineages? GATA1 shows a narrower expression profile during hematopoiesis with respect to TAL1 and LDB1. More specifically, both TAL1 and LDB1 are expressed and are functional in HSCs whereas GATA1 is not expressed at high levels until the Meg/erythroid progenitor (MEP) stage. Furthermore, GATA2, another member of the GATA family, is expressed and functional in

HSCs, but upon Gata1 upregulation in the erythroid lineage, Gata2 becomes specifically repressed (but remains active in megakaryocytes), whereas Tal1 and Ldb1 expression is maintained in both erythroid and megakaryocytic cell populations.

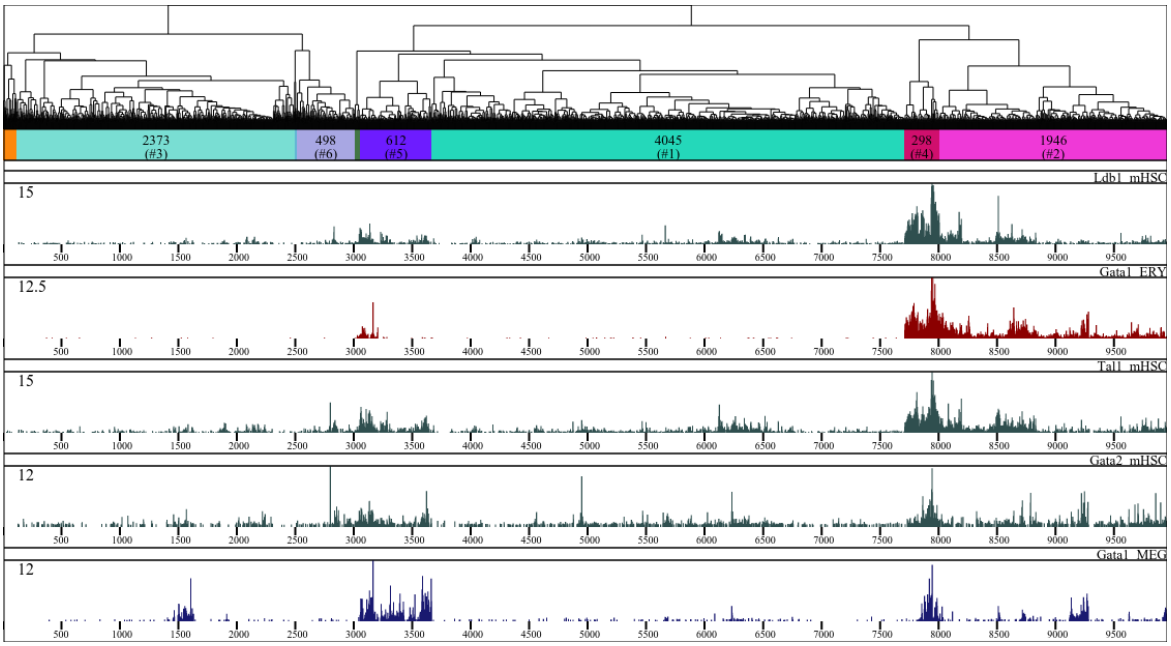


Figure 2.11: GATA1 erythroid occupancy priming by LDB1 and SCL/TAL1 in mHSCs (upper panels) and GATA1 megakaryocytic occupancy priming by GATA2 in mHSCs.

In trying to address the possible mechanisms that lead to the establishment of the differential GATA1 binding profiles in the erythroid versus megakaryocytic lineages, we compared the gene wide binding profiles of GATA1, TAL1, LDB1 and GATA2 in sequential and/or alternative differentiation stages, such as HSCs, proerythroblasts ($Ter119^-$), erythroblasts ($Ter119^+$) and megakaryocytic ($CD41^+$) cell populations).

Direct comparison of the TAL1, LDB1 and GATA2 genome wide binding profiles in mouse HSCs against those obtained for GATA1 in the erythroid and megakaryocytic lineages led to a series of observations. Firstly, the TAL1 and LDB1 binding profiles in mHSCs closely resemble those for GATA1 in erythroid cells, especially at the $Ter119^-$ proerythroblast stage (Figure 2.10). By contrast, the mHSC specific binding profile of GATA2 appears to be more closely related to that of GATA1 binding in the MK lineage, whereby gene promoters that are bound by GATA2 in HSCs are retained by GATA1 in MKs (Figure 2.11). In order to quantify the overall similarity of the gene wide binding profiles between GATA1, TAL1, LDB1 and GATA2 in the HSC, erythroid and megakaryocytic cell populations, we computed and clustered their linear correlation matrix (Figure 2.12). This led to a clear separation of the megakaryocytic GATA1 and TAL1 binding

profiles, whereas erythroid GATA1 and TAL1 co-clustered with LDB1 and TAL1 mHSC binding profiles ($R_{GATA1ery:LDB1mHSC}$ = 0.69 and $R_{GATA1ery:TAL1mHSC}$ = 0.62 compared to $R_{GATA1meg:LDB1mHSC}$ = 0.26 and $RG_{ATA1meg:TAL1mHSC}$ = 0.3, Figure 2.12). Based on these observations, we suggest that TAL1, LDB1 and GATA2 may serve as "bookmarking" factors for the GATA1 erythroid and megakaryocytic transcription programs, respectively. It is of interest to note that TAL1 assumes a different binding profile in the megakaryocytic lineage to that in mHSCs, suggesting a re-organization of its binding specificities/properties during hematopoiesis. Taken together these results show an erythroid bias of the TAL1 and LDB1 genome wide binding profiles in early, non-committed cells, priming erythroid specification, whereas GATA2 binding reveals a megakaryocytic bias, priming megakaryocyte specification.

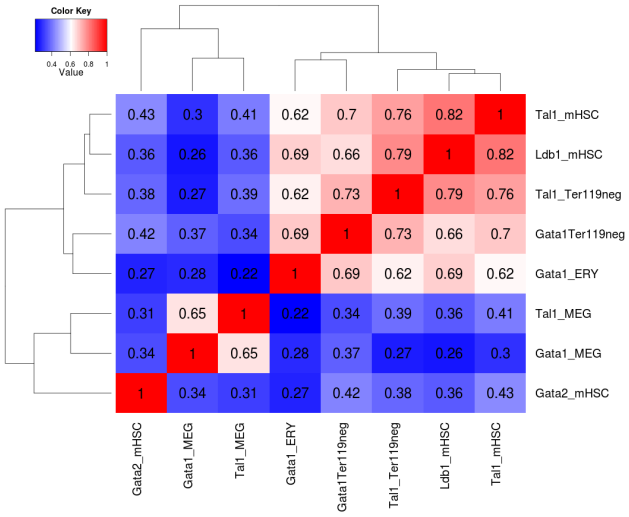
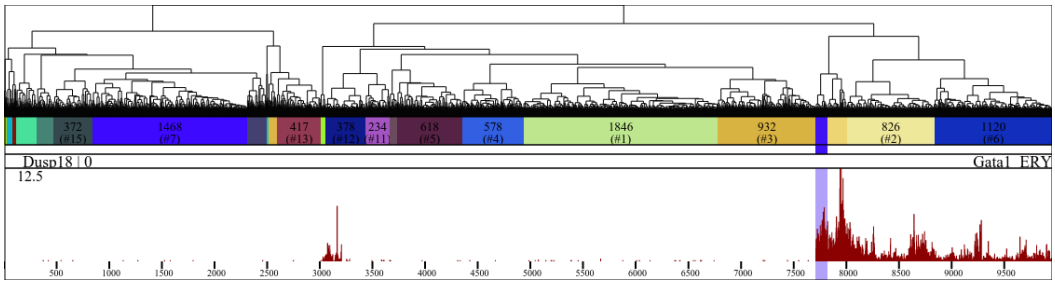


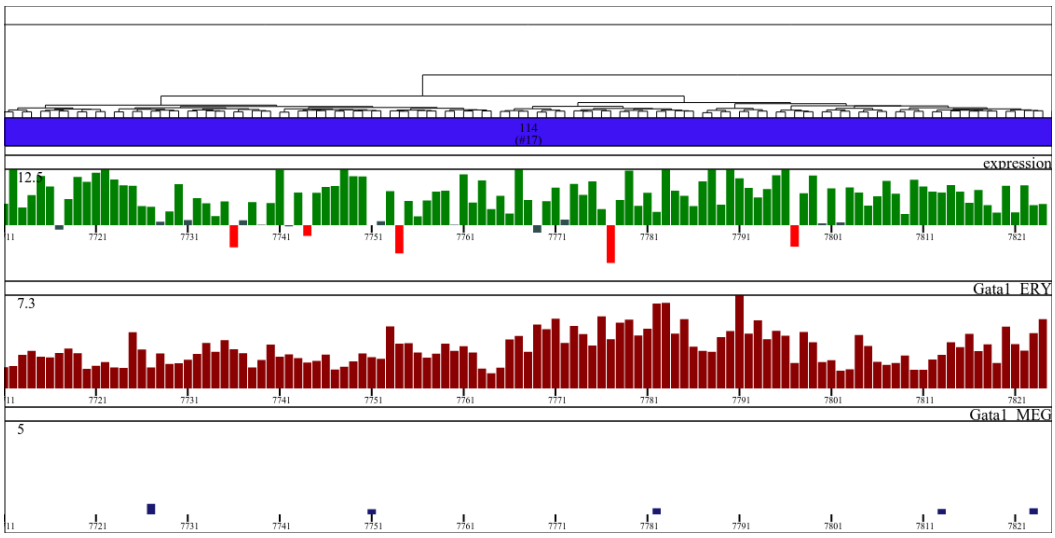
Figure 2.12: Correlation matrix of gene wide TF binding in different differentiation stages

2.3.3 Erythroid specific GATA1 binding cluster identifies epigenetic modifiers

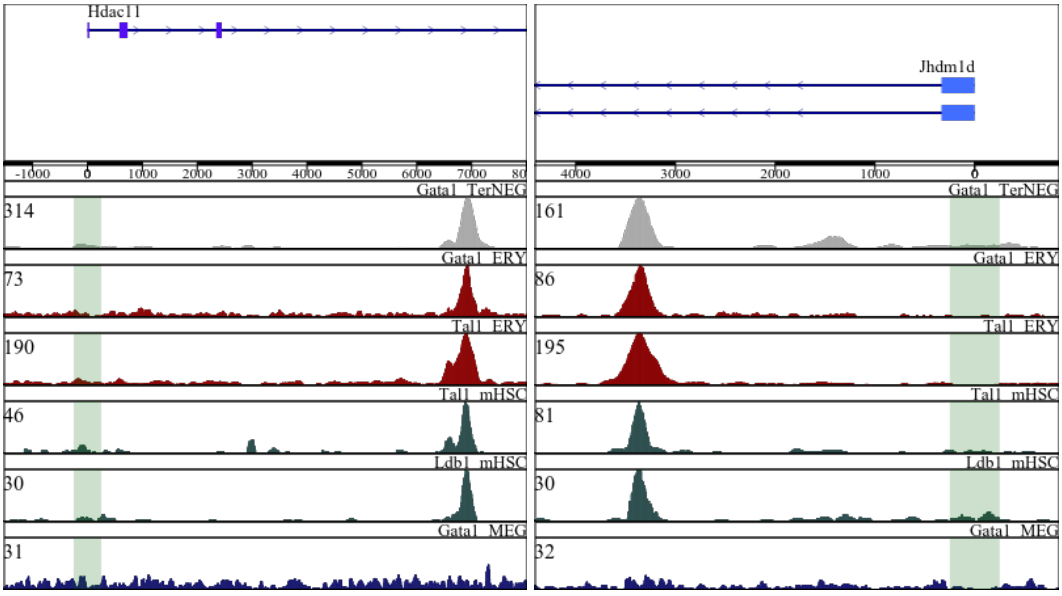
As shown previously, GATA1 binding profiles appear to be largely distinct between the erythroid and megakaryocytic lineages (Figure 2.10). Further dissection of the GATA1 bound cluster of genes in erythrocytes (Figure 2.13a, 2244 genes) leads to the identification of a subcluster of genes with highly specific erythroid expression profile (Figure 2.13b, 114 genes). Not surprisingly, this sub-cluster of GATA1 erythroid target genes is highly enriched for erythroid specific GO terms (not shown). Furthermore, it includes important erythroid specific TFs such as Lmo2, Klf1 (Eklf), Klf11(Emery et al., 2007) and Trim10, heme biosynthesis related enzymes such as Fech, Alad, Uros and Urod, β-globin gene Hbb-y, the erythroid-differentiation associated protein Hemgn (Yang et al., 2006), thus representing a highly specific erythroid signature. Importantly, the identification/isolation of this sub-cluster from the original full gene dataset (over 10000 genes) depends on the



(a)



(b)



(c) Hdac11

(d) Jhdm1d

Figure 2.13: Epigenetic landscape of epigenetic modifiers with erythroid specific expression patterns **a)** Full view of TF occupancy based clustering of genes expressed in erythroid and megakaryocytic cells. GATA1 erythroid specific binding is highlighted in blue. **b)** Sub-cluster view illustrating ERYvsMEG differential expression, GATA1 binding in erythroid and megakaryocytic cells. **c,d)** GATA1, LDB1 and TAL1 ChIP signal in mHSCs, erythrocytes and megakaryocytes in the Jhdm1d and Hdac11 genomic *loci*.

combinatorial genomic similarity (TF occupancy data) of these genes, rather than an *a priori* defined hypothesis (e.g. genes bount by GATA1 only in erythroid cells). Interestingly, within this highly erythroid specific cluster of genes we also find two of the previously identified epigenetic modifiers, Hdac11 and Jhdm1d, further supporting their involvement in erythroid differentiation. Furthermore, the TF occupancy profile of this sub-cluster indicates a potential role for, at least, GATA1 and GATA1 complexes in the gene expression regulation mechanism of these genes. In fact, both Hdac11 (Figure 2.13c) and Jhdm1d (Figure 2.13d) present with GATA1 and TAL1 peaks in erythroid cells. Moreover, there is no evidence of GATA1, TAL1 or FLI1 binding to these genes in MKs. Importantly, and in accordance with our previous observations on LDB1 and TAL1 'bookmarking' functions, TAL1 binding to these *loci* is already established in mHSCs.

2.4 Application on human hematopoiesis

The hematopoietic process is thought to be generally conserved throughout vertebrate evolution, thus, allowing for mouse models to complement our understanding of human hematopoiesis. In order to further extend and compare our findings for mouse hematopoietic genomic events, such as the loss of the active promoter state of a large group of genes during erythroid lineage specification, we applied our approach to human hematopoiesis. More specifically, we used gene expression and histone modification NGS data obtained during differentiation of multipotent human primary hematopoietic stem cells/progenitor cells (HSCs/HPCs, $CD133^+$) into erythrocyte precursors ($CD36^+$), (Cui et al., 2009). Transcription factor occupancy profiles during human erythroid development of fetal and adult peripheral blood-derived erythroblasts were also included (ENCODE, 2012; Xu et al., 2012).

As described above, we trained an RF-based non-linear regression model to predict differences in gene expression levels between $CD133^+$ and $CD36^+$ cells, based on the epigenetic profile of their broader (± 25kb) promoter region. Importantly, the RF regressor trained on differentially expressed genes (1832 genes, 58 predictors, Table 3.3) achieved an R^2 value of 0.73, whereas the model trained on all genes expressed in $CD133^+$ and/or $CD36^+$ cells (14111 genes, 58 predictors) achieved an R^2 value of 0.49. The high degree of variation explained by the models trained on both the mouse and human datasets indicates that our method can be generally applied.

In order to identify and compare the epigenetic signatures between $CD133^+$ and $CD36^+$ cells, we performed hierarchical clustering of the similarity values calculated by the RF model for the 1832 differentially expressed genes and for the full cohort of expressed genes (14111 genes). First, we compared the distributions of H3K4me3 marks between $CD133^+$

and CD36$^+$ cells (Figure 2.14). Based on this approach, we identify two distinct clusters of genes (Cluster 6 and Cluster 2, 341 and 202 genes, respectively) that present a marked reduction in H3K4me3 ChIPseq signal associated with the acquisition of the CD36$^+$ surface marker phenotype, closely resembling our findings on mouse hematopoiesis.

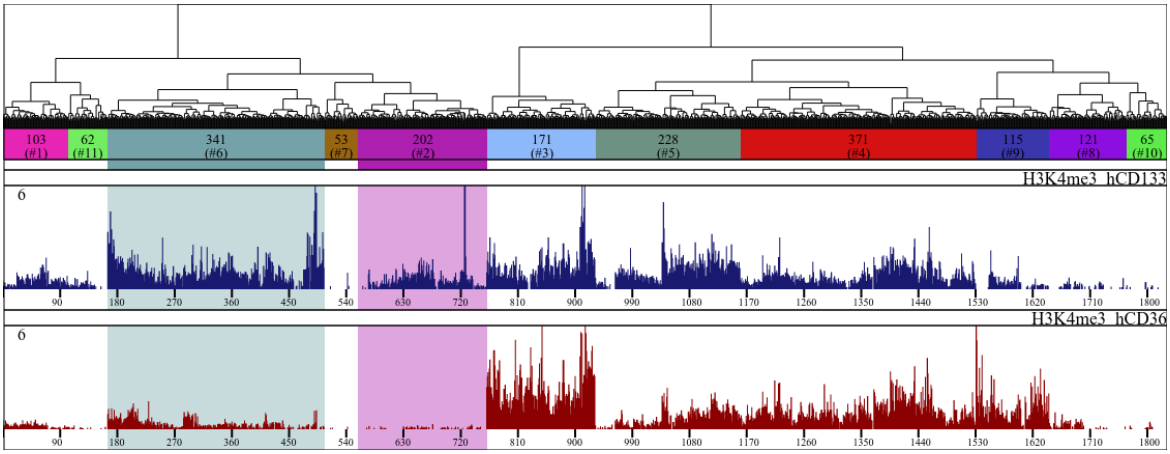


Figure 2.14: Loss of H3K4me3 mark during human erythroid commitment

Strikingly, and in contrast to mouse hematopoiesis, scaling up the differentially expressed genes trained model to include all genes expressed in either CD133$^+$ or CD36$^+$ cells, we find that the cluster characterized by loss of the H3K4me3 mark is restricted to a smaller fraction of genes (~240, Figure 2.15). Interestingly, the loss of H3K4me1 mark from a large subset of genes (1132 genes) seems to play a similar role in human hematopoiesis. Furthermore, the gene loci undergoing H3K4me1 loss can be divided in two large groups: a) One major group (870 genes) characterized by very low levels of H3K4me3 in both CD133$^+$ and CD36$^+$ cells; b) One minor group (262 genes) characterized by relatively high levels of H3K4me3 only in CD133$^+$ cells (HSCs).

Summary

In this chapter we interrogated the genome wide epigenetic profiles of human and mouse hematopoietic stem cells, early and late erythroid progenitors and megakaryocytes to identify erythroid specific epigenetic signatures. Based on our approach, one of the main epigenetic events characterizing erythroid specification and differentiation processes in mice is the loss of the active promoter state from a large group of genes (~1000) in early (CD71$^+$/Ter119$^-$) and late (KIT$^-$/Ter119$^+$) erythroid committed progenitor cells. Moreover, these genes show significant levels of active promoter associated marks, such as, H3K4me3 and H3K4me1, in both hematopoietic stem cells and in megakaryocytes, thus indicating an erythroid specific

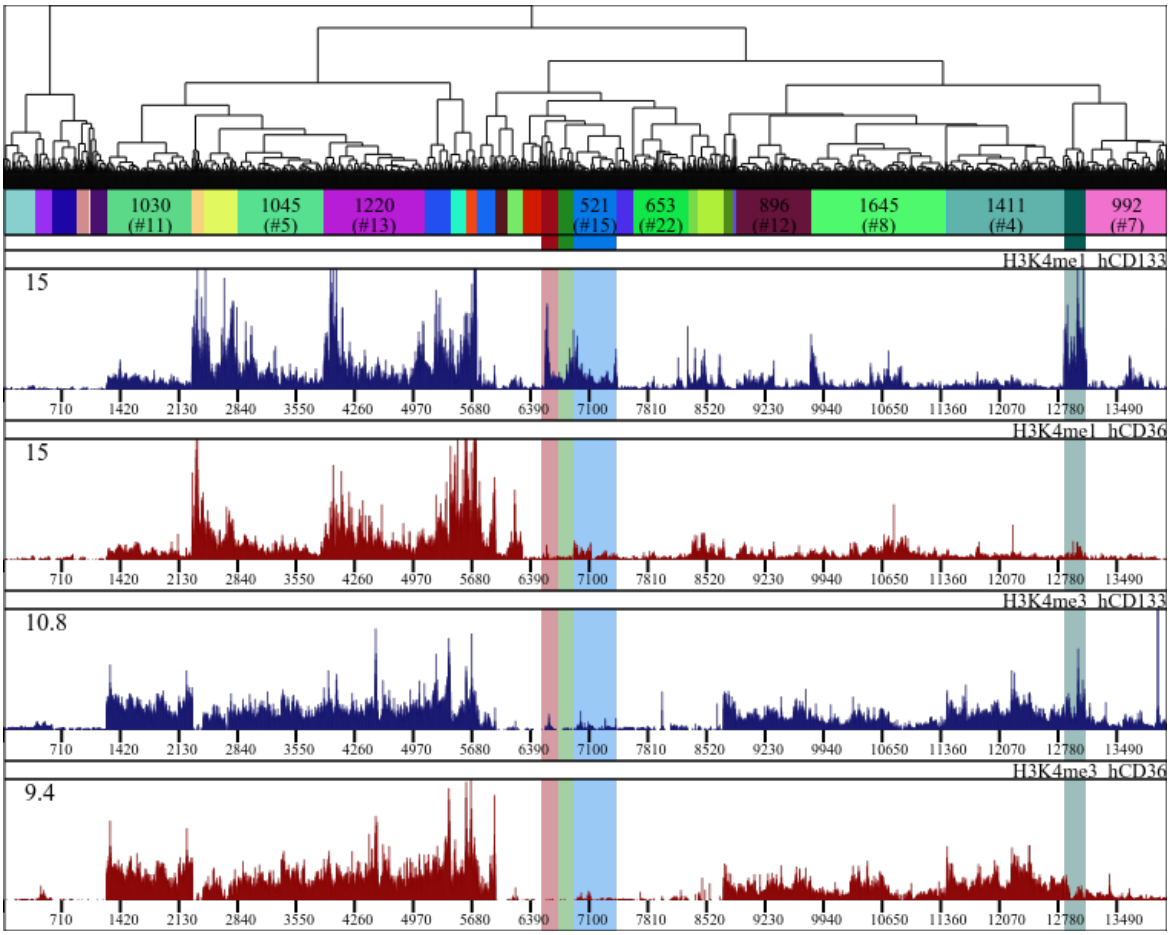


Figure 2.15: Loss of H3K4me1 histone tail modification is evident in human erythroid differentiation

histone demethylation and deacetylation mechanism. In accordance with their epigenetic profiles, functional characterization of these genes showed a highly significant enrichment in genes associated with megakaryocyte functions, such as, blood coagulation and platelet activation.

In order to identify potential erythroid specific factors that could be involved in the dynamic chromatin changes observed we tested a comprehensive set of histone modifiers for erythroid specific expression profiles and epigenetic state of their genomic *loci*. This approach successfully identified epigenetic modifiers with previously described functions in erythroid and megakaryocytic specification and differentiation, such as Mllt3 (Pina et al., 2008; Zini et al., 2012) and Mst1 (Nejigane et al., 2013), as well as, two potential novel factors, Hdac11 and Jhdm1d. Importantly, Jhdm1d and Hdac11, co-segregated with a highly erythroid specific gene subset identified by the differential binding profiles of GATA1, TAL1 and LDB1 in different hematopoietic cell populations (HSCs, erythroid and megakaryocytes), probably accounting for an erythroid specific regulation of these genes.

Comparison of the gene-wide occupancy profiles of hematopoietic associated transcription factors showed clearly distinct occupancy profiles for GATA1 and TAL1 in erythroid and megakaryocytic cells. Further investigation on the molecular mechanisms that could account for these differential binding profiles revealed an erythroid bias of the TAL1 and LDB1 genome wide binding profiles in early, non-committed cells, priming erythroid specific occupancy pattern of GATA1.

Application of our approach on human hematopoiesis revealed a similar, but not identical, pattern. More specifically, erythroid lineage specification (CD133 to CD36 transition) showed a significant loss of H3K4me1 histone mark in a large subset of genes (~1100), but only a small fraction (~240 genes) of these showed significant levels of H3K4me3 in multipotent hematopoietic progenitors. Interestingly, functional characterization of these genes showed a highly significant enrichment in genes associated with lymphoid cell functions, such as, inflammatory response and leukocyte activation and migration.

Chapter 3

Ariadne: Combine and unravel

Abstract

In the previous chapters we described the use of NGS experimental data in identifying common chromatin features characterizing distinct subsets of genes. In order to allow for mining and interpretation of the results obtained using these approaches, we developed Ariadne as a comprehensive tool to compare gene-wide relational profiles of NGS datasets and to visualize primary sequencing data within single gene loci.

Specifically, Ariadne is a web based framework for interactively creating and visualizing chromatin state profiles (aegeas.imbb.forth.gr/Ariadne/). It is composed of two main components:

1. a treeViewer module that allows for interactive visualization and selection of gene subsets based on hierarchical trees calculated using genomic similarity measures;

2. a geneViewer module that allows for interactive visualization of extended gene loci and their comparison with a comprehensive set of regulatory features (histone tail modifications, trancription factor occupancies, DNase hypersensitivity, RNA polymerase II profiles) assayed by NGS approaches (ChIPseq, RNAseq) in primary human and mouse hematopoietic cells.

In this way a gene can be interrogated both as part of a tree branch (cluster) and as single gene locus, taking advantage of the extremely high resolution provided by NGS data.

3.1 Ariadne treeViewer

As described previously the central scope of this study is to assemble single-gene information based on NGS data, in order to identify lineage specific epigenetic patterns which may provide clues as to the regulatory mechanisms underlying the separation of lineage commitment or differentiation processes. As the number of genes assayed by NGS approaches is very large, it becomes critical to identify representative patterns that characterize specific subsets of genes which in turn may hint at common underlying regulatory mechanisms. Our approach produces a hierarchical clustering of all genes under investigation as an end result. One of the advantages provided by hierarchical clustering is that relations between gene similarities can be represented as a dendrogram. A dendrogram is an intuitive graphical way to display the relations between large numbers of samples (i.e. genes). Furthermore, another advantage of hierarchical clustering is the fact that the number of clusters is not predefined, as in k-means approaches, but can be chosen by the observer, further facilitating the identification of specific patterns.

One drawback in applying different parameters on dendrograms is the fact that it requires the use of specialized software (e.g. R, matlab etc etc). Ariadne treeViewer is a web-based interface to graphically interact with dendrograms, thus eliminating the use of any specialized software. More specifically, the structures of the dendrograms, produced in R, are exported as a Json object whereas the visualization part takes advantage of the HTML5 'canvas' element, used to draw graphics on a web page. This allows for a fast and interactive representation of the clustering results produced by our approach in a user-friendly environment, such as a web interface.

Ariadne treeViewer Web Interface (front-end)

The Ariadne treeViewer web interface is divided in four main parts (Figure 3.1).

1. Selection of the clustering approach to interrogate (A);
2. Selection of the NGS datasets' TGS to plot under the dendrogram (B);
3. Functions to be applied on the plots or gene subsets (C);
4. The plots area including the dendrogram and the corresponding TGS profiles (D).

Clustering Selection

First, the user selects the clustering analysis to interrogate. This choice depends on both the predictors and the dependent variable used in the Random Forest regressor. This choice is

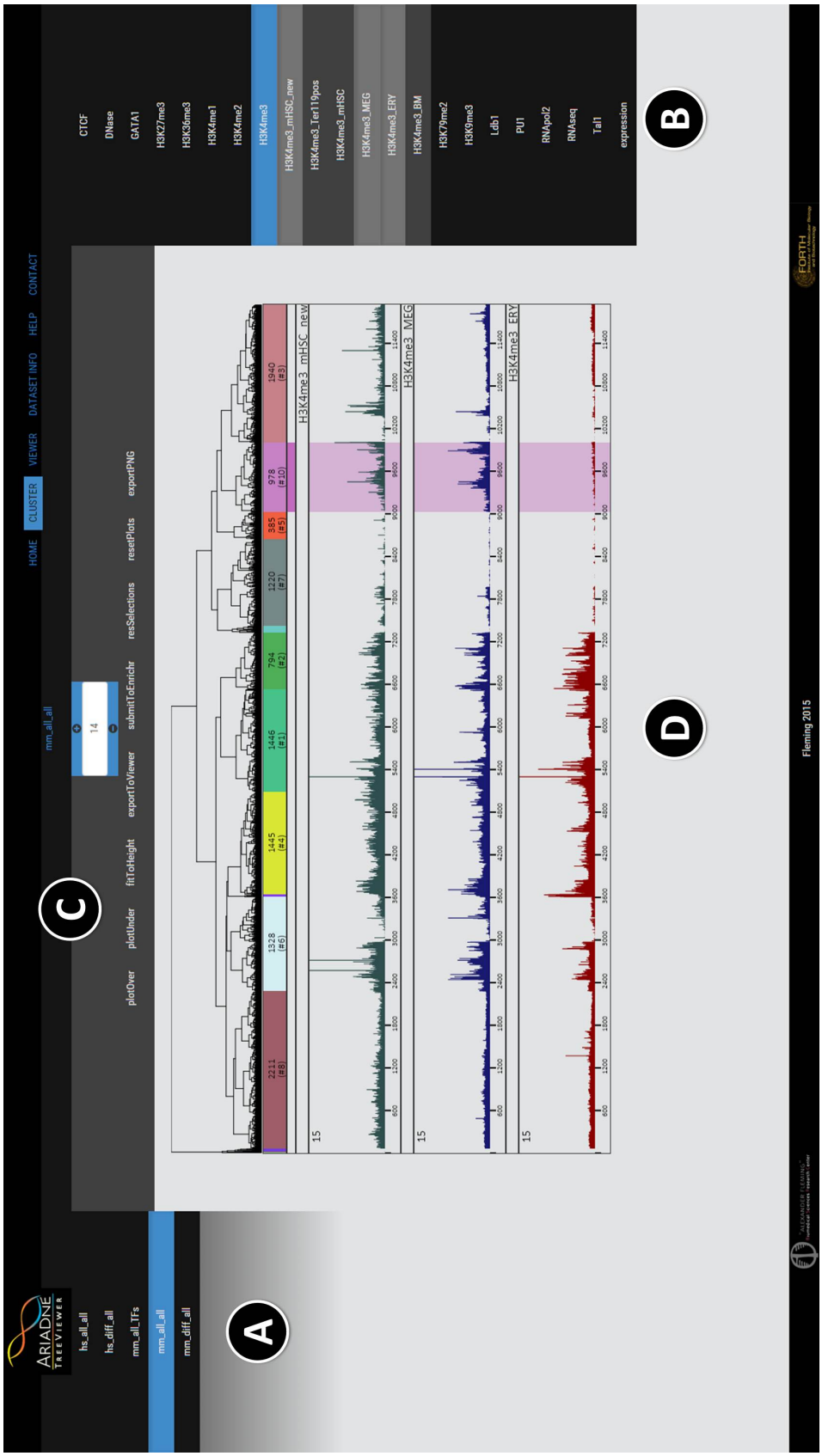


Figure 3.1: Schematic overview of Ariadne treeViewer

critical since the gene similarity values used to identify relationships between genes will be different for different combinations of predictors and/or dependent variables. The Ariadne treeViewer currently includes results based on either differentially expressed genes (1623 genes) or all genes expressed in the erythroid and/or megakaryocytic lineages (11949 genes). Predictors include either epigenetic and TF occupancy or only TF occupancy datasets. In this way epigenetic signatures and TF co-occupancy patterns can be interrogated. Additional models include human differentially expressed genes between $CD133^+$ (hematopoietic stem cells) and $CD37^+$ (erythrocyte precursors) primary cells (1832 genes) or all genes expressed in either $CD133^+$ and/or $CD37^+$ cells (14111 genes). This allows for a direct comparison of chromatin dynamic changes during human and mouse hematopoiesis.

Once the clustering result is selected, the dendrogram showing the relations between the genes included in the selected analysis are plotted in the plots area. The dendrogram plot is divided in two distinct areas: The upper part includes the dendrogram representation. The middle part reports information about the cluster IDs and the number of genes within each cluster whereas the lower part is used to highlight selected clusters.

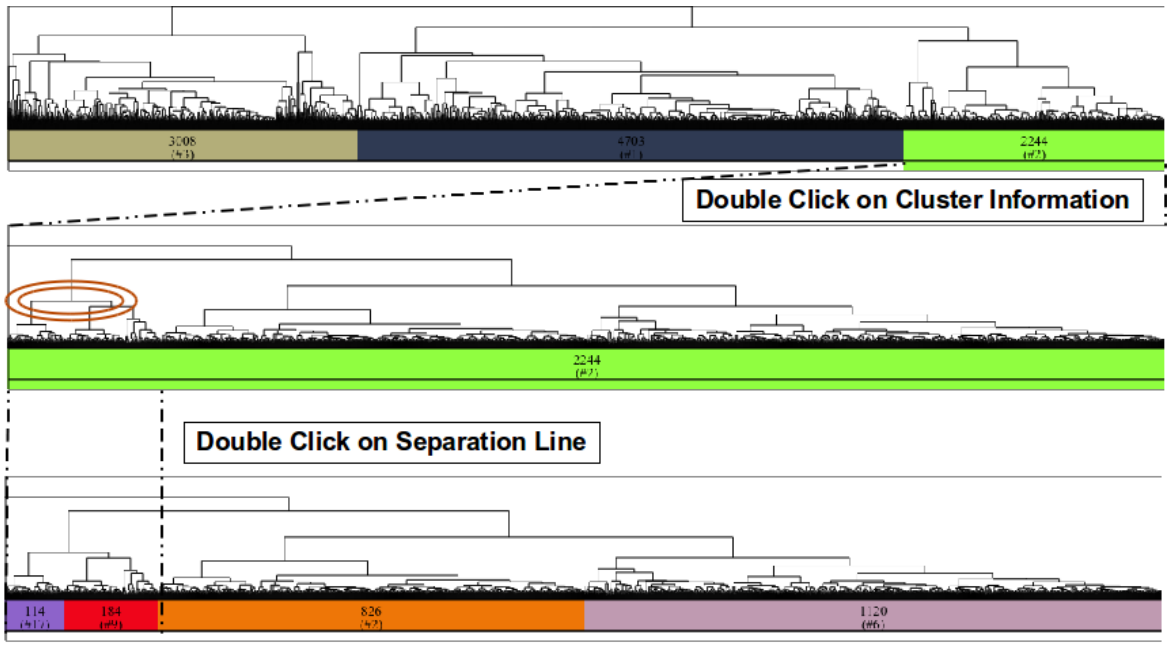


Figure 3.2: Ariadne treeViewer Dendrogram Functions

A number of interactive functions are implemented in the dendrogram plot area. Mouse scrolling triggers a zoom function whereas a drag function allows horizontal movement of the plot in order to facilitate visualization of specific parts of the dendrogram. Clicking on the cluster information area (middle part) selects and highlights the part of the plots within the selected cluster. Double clicking expands the selected cluster in the whole plot area. Double

clicking on a cluster separation line (horizontal lines in the dendrogram) automatically calculates the minimum number of clusters that separate the two branches at the ends of the line and applies it to the current visualization.

Cluster Characterization

As mentioned above the main goal of the application is to allow for both the identification and the functional characterization of specific gene subsets (clusters). Given the increased complexity resulting from the potentially large number of clusters and the large number of NGS datasets used to define them there is a growing need for an easy way to visualize and infer the importance of each dataset in each cluster. Driven by the fact that a "heatmap" approach not only reflects the mean values for each cluster but is also too complex to interrogate when dealing with the integration of multiple NGS data, we opted for a more human readable visualization format. To do so we plotted the total gene score (TGS; defined as the sum of the ChIPseq peak scores assigned to it) of each gene based on the hierarchical clustering order (Figure 3.3). This approach produces a much higher resolution for each dataset, whereby the ChIPseq signal of each gene is represented by a single vertical line, the height of which corresponds to its TGS.

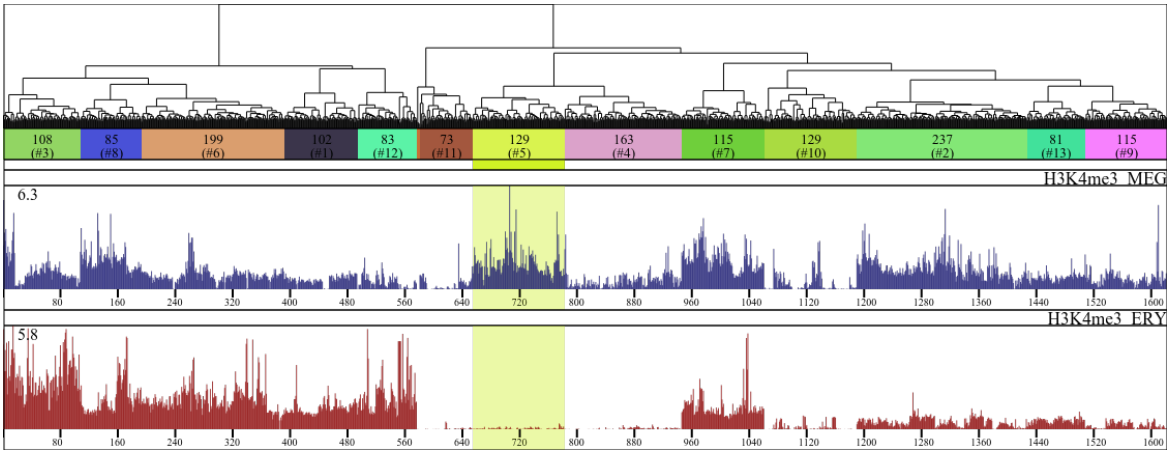


Figure 3.3: Is there a distinct H3K4me3 signature between erythroid and megakaryocytic lineages?

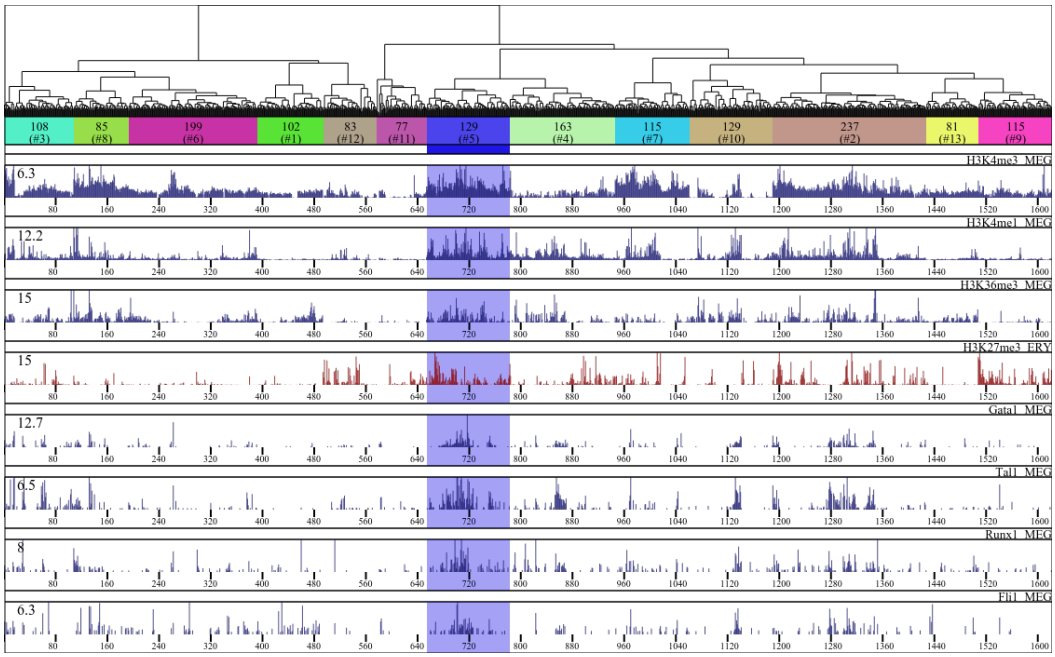
The Ariadne treeViewer allows the user to select which chromatin features to visualize and compare. More specifically, all available NGS datasets are displayed on the right side of the web interface (Figure 3.1, part B). Available datasets are divided by the primary target used in the ChIPseq (e.g. GATA1, H3K4me3 etc.) or RNAseq steps of the experiment. Upon selection of a given target all available cell populations assayed for the selected chromatin feature are shown. This approach allows the user to visualize and compare any combination

of chromatin features and expression profiles in order to identify and/or characterize a given subset of genes.

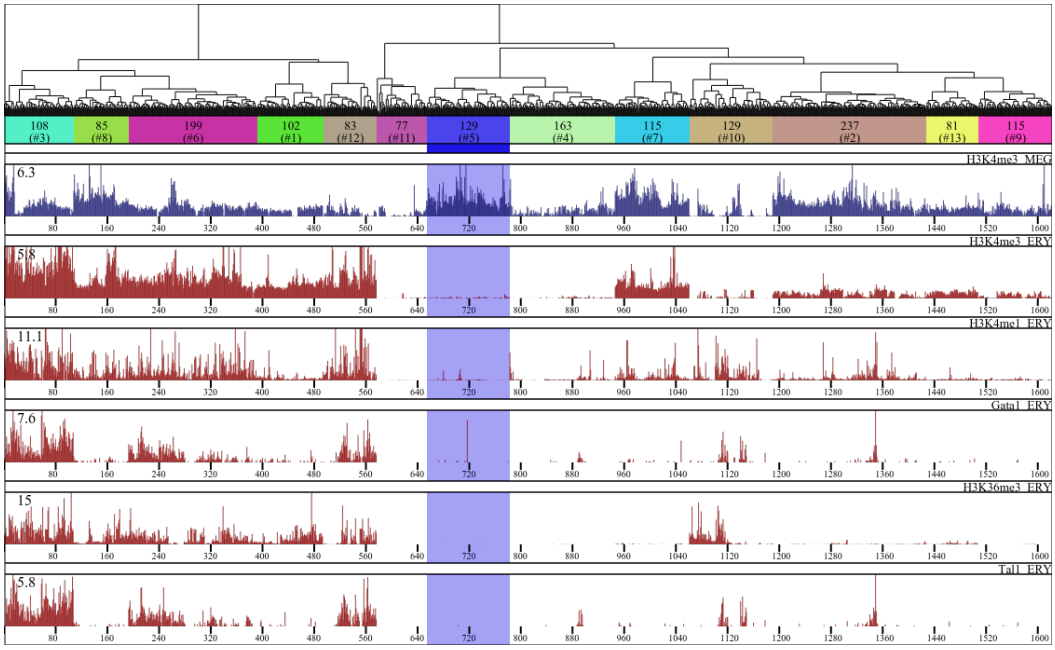
The aim of this functionality is to allow for the development of a pure hypothesis driven approach to be tested easily and interactively (e.g. as described in the previous Chapter: “Is there a distinct H3K4me3 signature between erythroid and megakaryocytic lineages?” or “Is there a distinct GATA1 occupancy profile between erythroid and megakaryocytic lineages?” or “What is the overlap of H3K4me3 and H3K27me3 marked genes in human CD133 and/or CD36 cells?”). Plots representing the gene-wide enrichment of single NGS datasets (Total Gene Scores) based on the clustering order are also based on ’canvas’ elements, thus allowing for several interactive features to be applied. Accessibility features such as the identification of single genes by mousing over, or variation of the overall plot height and direct visualization of the read density profiles simply by double clicking on the selected TGS peak have been implemented in order to further facilitate visualization, interpretation and validation of the results.

To further facilitate the functional characterization of a given cluster, we implemented a function activated by two dedicated buttons in the functions panel to automatically plot NGS datasets over and under-represented within a given cluster, (Figure 3.1, part C). This function greatly simplifies cases where a given cluster is selected based on a given characteristic (e.g. H3K4me3 marked genes specifically in megakaryocytic cells) and the user is interested in additional chromatin features that are enriched or absent within the given gene subset. Statistical significance of under and over-represented datasets is assessed by a Wilcoxon rank-sum test between the distribution of TGS scores within the selected cluster compared with the TGS score distribution of the full dataset. A strict significance threshold of $P<10^{-6}$ is applied for both over and under-representation. Significance is currently calculated in R for 2 to 150 clusters and then exported to the treeViewer database and currently supports only single cluster selection. Future prospects include the implementation of a Javascript based statistical package to allow for the ’on-the-fly’ calculation of significance, as well as merging of multiple clusters.

In the example shown in Figure 3.4a it can be seen that the H3K4me3 megakaryocytic specific cluster is also enriched for H3K4me1 and H3K36me3 in megakaryocytes and for H3K27me3 in erythroid cells. Furthermore, it is enriched for Fli1, Gata1, Tal1 and Runx1 occupancy in megakaryocytes (Figure 3.4a). Not surprisingly, all of these factors are found to be significantly under-represented in erythrocytes (Figure 3.4b).



(a) Over Represented Features



(b) Under Represented Features

Figure 3.4: Over and under represented chromatin features in H3K4me3 megakaryocytic specific gene cluster

Functional characterization of selected clusters

Biological interpretation of the identified epigenetic or transcriptomic signatures is greatly facilitated by the assessment of the functional associations of genes included in the cluster(s).

The Ariadne treeViewer takes advantage of Enrichr, an interactive and collaborative HTML5 gene list enrichment analysis tool (Chen et al., 2013). More specifically, Ariadne treeViewer uses the JavaScript function provided by the EnrichR developers to automatically submit the gene list resulting from the user selected cluster(s). This function is accessible from the functions panel. As an example, if the 129 genes comprising the H3K4me3 megakaryocytic specific signature are submitted to EnrichR to perform a GeneOntology Biological Process enrichment analysis, we find that these genes are highly enriched in blood coagulation and platelet activation related gene ontology terms.

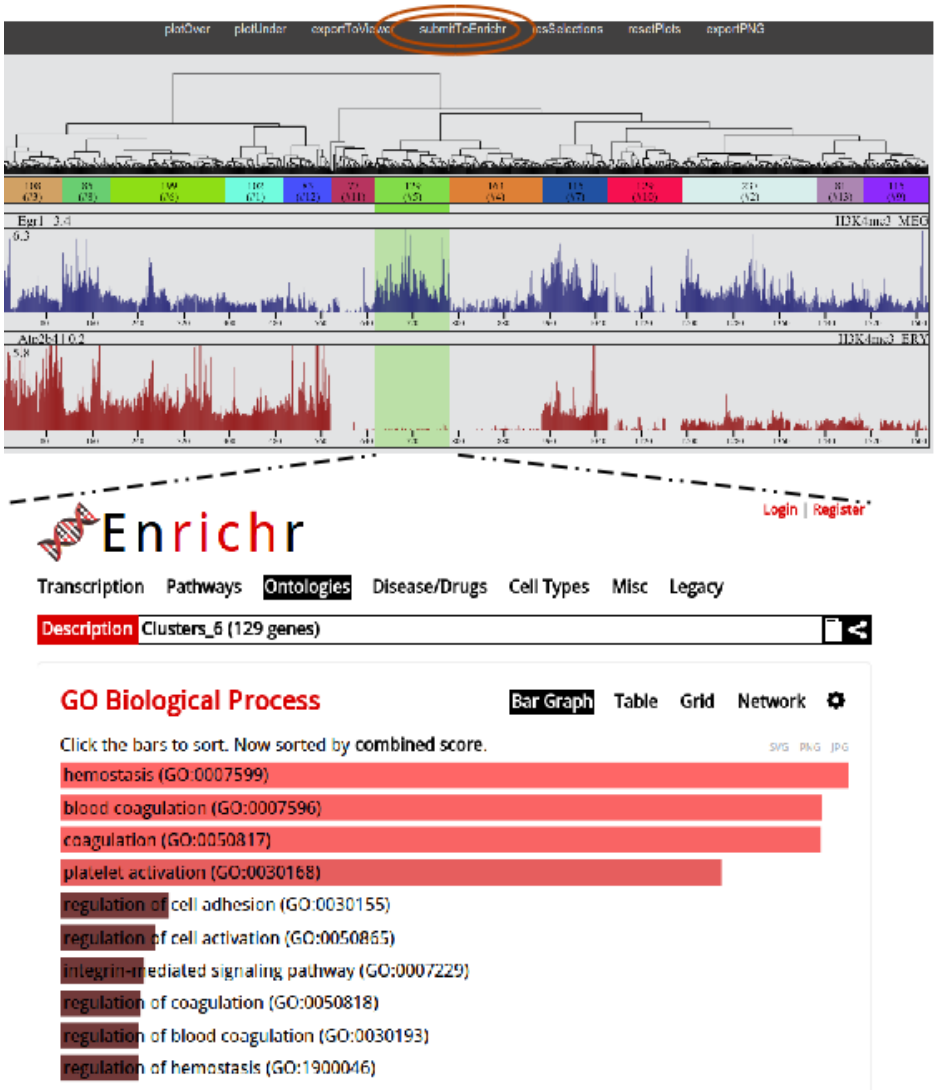


Figure 3.5: Gene lists composing the selected clusters (e.g. MK specific H3K4me3 cluster, upper panel) can be directly submitted to the online functional analysis tool EnrichR. Graphical interface of EnrichR for GO terms enrichment analysis is shown in the lower panel.

Validation of the signature

One major advantage of NGS based technologies is the very high resolution of the results obtained. Until now we presented how the assembly of NGS data into single-gene information (TGS, gene similarity value) can greatly reduce the complexity of the data and facilitate biological interpretation. Nevertheless, this single-gene information measures represent only an approximation of the actual measurements and they don't include any information about the overall gene *locus* structure and/or peak structure. Furthermore, despite the validation and optimization steps applied in our pipeline, there are several statistical thresholds and normalization steps that can introduce false positive and false negative readouts.

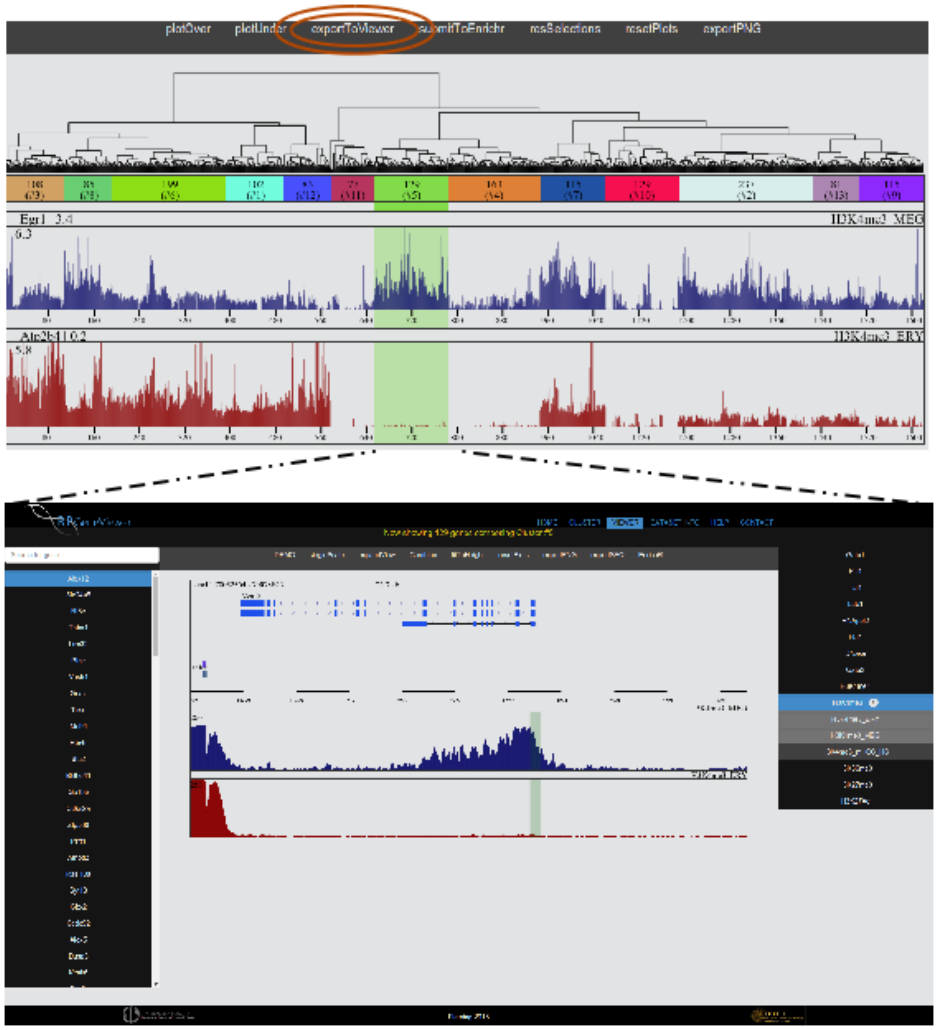


Figure 3.6: Read density profiles of gene *loci* composing the selected clusters (e.g. MK specific H3K4me3 cluster, upper panel) can be directly submitted to the Ariadne geneViewer module.

In order to take advantage of the full resolution provided by NGS based approaches we developed Ariadne geneViewer as a web-based gene browser that fully complements the Ariadne treeViewer single-gene information approach. The main goal of this implementation is that once a specific signature has been identified by the similarity clustering approach of the Ariadne treeViewer, the user is able to access the primary sequencing data (read density profiles) that produced the single-gene information. In this way the user can directly validate the clustering results, interrogate the overall locus chromatin structure, visualize different peak structures (e.g. H3K4me3 and H3K36me3 peaks), identify peak locations (e.g. enhancer or promoter associated binding), identify differential binding events (e.g. loss or gain of TF binding events or alternative promoter usage) within a systematically defined subset of genes. This function is accessed from the 'exportToViewer' button in the functions panel and automatically loads a customized session of the Ariadne geneViewer module composed by the genes included in the selected cluster(s), maintaining their hierarchical order.

3.2 Ariadne geneViewer

Investigation of the genomic profiles of single gene loci based on NGS data can provide invaluable information about the overall locus organization and differential occupancy patterns of differentiating hematopoietic cells. Unfortunately, web based visualization tools of read density profiles (e.g. the UCSC Genome Browser) are presently too slow and complex to use due to the enormous amount of data they need to process, store and display.

The Ariadne geneViewer is a web based gene browser that comprises a comprehensive set of manually selected and curated NGS datasets including ChIPseq, RNAseq and DNAseSeq profiles assayed in hematopoietic cells. More specifically, the portal allows for the interactive visualization of multiple genomic features of extended genomic regions in human and mouse cells The highly targeted content of the portal and the intuitive user interface greatly accelerates and simplifies the comparison of different binding profiles as well as the comparison of distinct binding profiles along different gene loci.

Ariadne geneViewer Web Interface (front-end)

The Ariadne geneViewer web interface is divided in four main parts, similar to the previously described Ariadne treeViewer web interface (Figure 3.7). These parts allow:

1. Selection of the gene (genomic region) to visualize (A);

2. Selection of the NGS datasets read density profiles to plot (B);

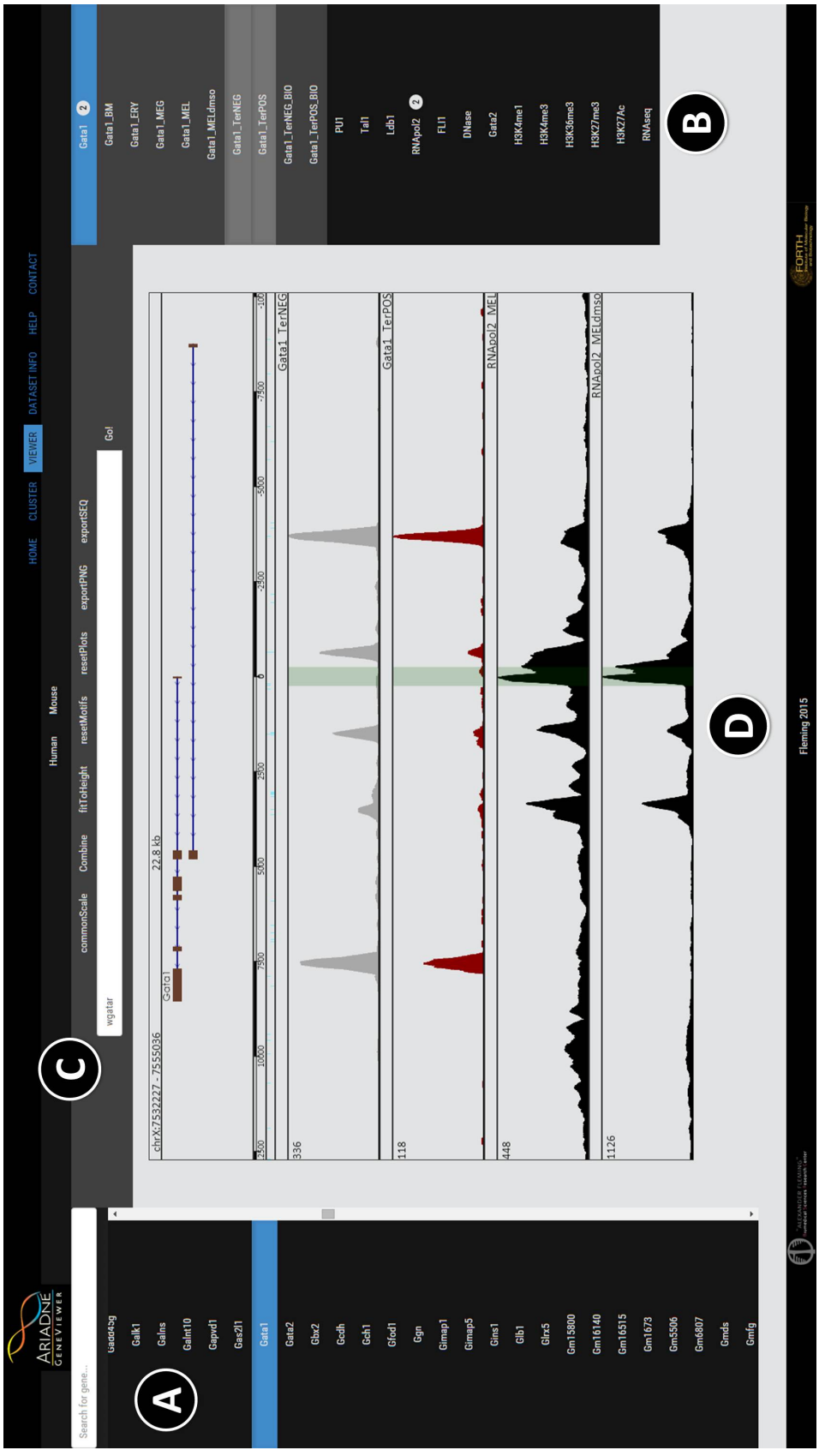


Figure 3.7: Ariadne geneViewer Web Interface

3. Functions to be applied on the plots (C);

4. The plots area including the gene(s) and read density profiles (D).

Gene and Species Selection

First, the user selects the species and the genomic region (gene) to visualize. The Ariadne geneViewer currently supports mouse (mm9) and human (hg19) data. By default, the gene selection panel includes all the genes included in the database. Alternatively, and in order to facilitate the comparison of specific subsets of genes, the gene selection panel can be modified to include either a user defined gene list or, in combination with the treeViewer module, a specific cluster submitted (see above).

To allow for fast access to specific genes we implemented a search panel (accessible at the top of the gene selection panel) that interactively scans the current contents of the selection panel and adjusts the panels' visible area to the search string. The visualization of the gene is triggered by clicking on the gene name button. The currently visualized gene will be highlighted.

NGS Read Density Profile Selection

The NGS dataset profile selection panel maintains an identical layout as in the Ariadne treeViewer module described above. One of the main reasons for this is to allow for a direct switch between the gene-wide view provided by the treeViewer and the actual primary sequencing data used to identify systemic relations between subsets of genes. In fact, in the case of a direct submission from the treeViewer module (either single gene or cluster submission) the geneViewer session will automatically plot the primary sequencing data of the datasets plotted in the treeViewer session.

Ariadne geneViewer Implemented Functions

Several features have been implemented to allow for an interactive overview of the genomic and regulatory patterns of the visualized gene *locus* and read density profiles. Main features include an interactive zoom in and out function from the current visualization, accessible by simply mouse scrolling over the plots area. This allows interrogation of both long range and overall locus organization (over 100kb) down to single transcription factor binding sites. Moreover, the read density profiles can be visualized in either a single plot for each dataset or in combined manner in a single plot (Figure 3.8). This allows for a direct comparison of the height and localization of peaks between different chromatin features and/or TF binding

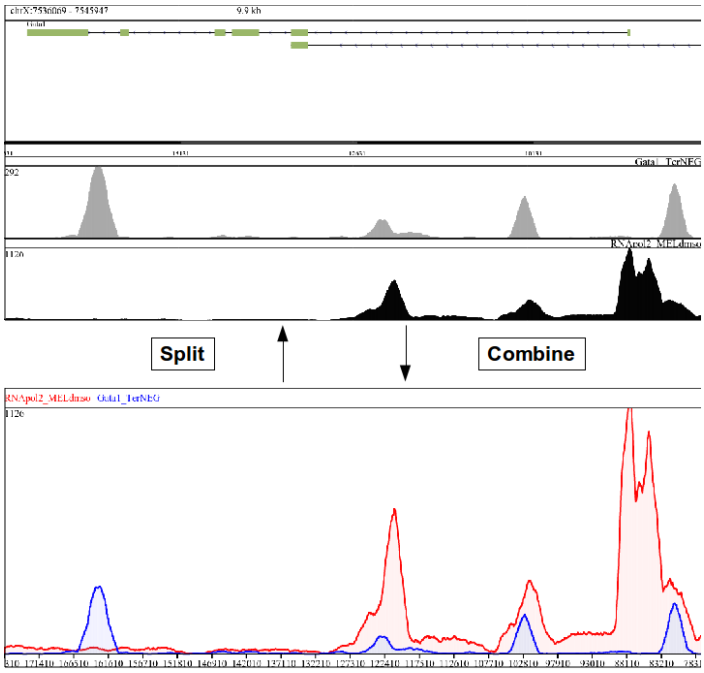


Figure 3.8: Visualization of read density data in combined (lower panel) and separate plots (upper panel)

events and is accessible from the Combine/Split button in the functions panel. Alternatively, the user can set the same y-scale limit (number of reads, visualized at the upper left corner of each plot) to all plots by clicking the singleScale/commonScale button. The y-scale limit of each plot can be manipulated by the + and - buttons appearing at the right side of each plot when the mouse cursor is over the plot area.

To further expand the genomic and genetic information accessible by the Ariadne geneViewer module we implemented a sequence mapping and a sequence export function. To do so we included the primary DNA sequence under the genomic regions available in the Ariadne database. The sequence mapping function is accessible by the text area present in the functions panel. The user is allowed to enter an extended IUPAC nucleotide string that is then mapped to the underlying DNA sequence of the visualized genomic *locus*. All positions that contained the input DNA string are then highlighted in the dedicated area under the genomic *locus* plot. This is particularly useful in cases where one or multiple TF binding motif(s), PCR primer(s), restriction enzyme target sequence(s) or any other sequence based characteristic needs to be related to NGS datasets. Furthermore, the user can directly export the DNA sequence under the visualized part of the genome in order to design qPCR primers to validate ChIPseq readout.

Finally, the current visualization can be exported as a PNG file by clicking the exportPNG button located at the functions panel.

Summary

Overall, the Ariadne geneViewer can be used to visualize multiple read density profiles, assayed in human and mouse hematopoietic cells, mine the underlying sequence and produce high quality images. For example, the Ariadne treeViewer module was used to examine the regulatory regions of the myeloid specific TF PU.1 (Sfpi1) in multiple hematopoietic cell populations (Burda et al, 2015, under review, Figure 3.9).

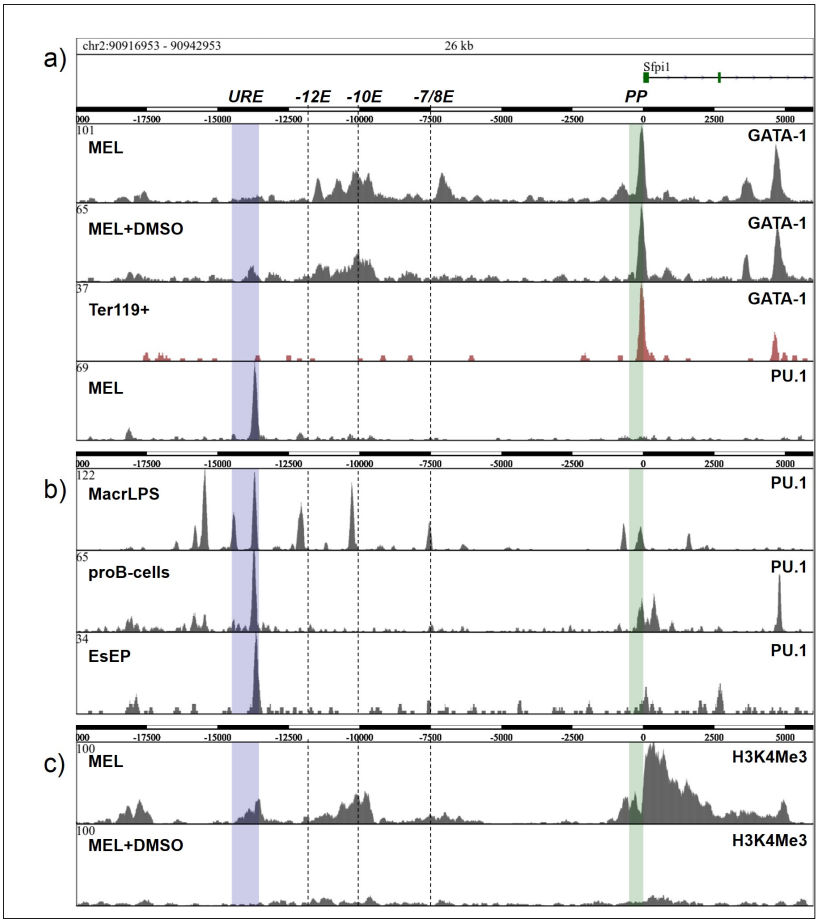


Figure 3.9: ChIP-seq read density profiles of GATA1, PU.1 and H3K4me3 datasets at the PU.1 (Sfpi1) gene locus (-20kb upstream to 6kb downstream of PU.1 TSS). The proximal promoter (PP, 0 to 500nt upstream the TSS) and the Upstream Response Element (URE, -13548bp; -14505bp) are highlighted. A) Occupancy of GATA-1 in MEL cells or MEL induce to differentiate by 2% DMSO for 5 days, or in Ter119+ Fetal liver (E14.5-derived) erythroblasts. The lowest panel shows occupancy by PU.1 in MEL cells. B) Binding profiles of PU.1 and GATA1 in murine primary macrophages stimulated by lipopolysaccharide for 24hrs (MacrLPS), pro-B cells (38B9), and EsEP: differentiating murine ES cell-derived erythroid progenitors. C) H3K4me3 occupancy in MEL and MEL+DMSO cells.

Ariadne Overview

The Ariadne treeViewer allows the user to perform a series of operations on the hierarchical clustering relational tree, as follows:

1. Select the number of clusters to apply
2. Select a particular height to cut the tree in order to isolate a specific branch
3. Select one or multiple clusters
4. Plot the corresponding linear TGS profiles of any NGS dataset based on the clustering order
5. Automatically plot NGS datasets over and under-represented in the selected clusters
6. Export genes composing the selected clusters to the geneViewer module
7. Export genes composing the selected clusters to EnrichR for GeneOntology analysis
8. Export high quality images of the current output

The Ariadne geneViewer allows the user to visualize multiple normalized read density profiles of user selected NGS datasets around UCSC annotated genes (from 25kb upstream the TSS to 25kb downstream the TES) and to perform a series of operations, as follows:

1. Zoom in or out of the selected gene region (from 500nt up to over 100kbs)
2. Select between individual or combined view of the density profiles
3. Export the underlying DNA sequence of the visualized genomic region
4. Map DNA sequences (motifs, primers etc.) in the visualized genomic region
5. Interrogate specific gene subsets submitted by treeViewer
6. Submit user defined gene sets for interrogation
7. Highlight significantly enriched regions (peaks) within the visualized density profiles
8. Export high quality images of the current output

Discussion

Computational analysis and biological interpretation of next generation sequencing (NGS) data are becoming powerful and indispensable tools in biological research and have made a real impact in the genomics and gene regulation fields. NGS applications, such as ChIPseq and RNAseq, produce a highly accurate representation of the epigenetic and transcriptional states of any given cell population or state and provide an excellent means for the comparative characterization of different populations of cells and/or states e.g. differentiation or developmental stages, disease versus physiological states etc. Our main research focus has been the characterization of the epigenetic mechanisms involved in erythroid and megakaryocytic cell fate specification processes and, to these ends, we developed a computational approach to analyze, integrate and visualize NGS data at a genome wide level which can be applied essentially with any NGS collection of datasets. More specifically, our approach involves the use of an extensive range of genomic and transcriptomic datasets (histone tail modifications, transcription factor occupancies, DNA methylation etc.) derived from distinct hematopoietic cell populations and a combination of supervised (RandomForest regression) and unsupervised (hierarchical clustering) machine learning approaches to produce a highly structured, easy to visualize gene clustering pattern for each (epi)genomic feature.

Computational approaches have been extensively used in the integrative analysis of NGS data. More specifically, Random Forests have been used to model the influence of TF binding (Cheng et al., 2012) and histone modifications (Dong et al., 2012) on gene expression using a large set of NGS datasets available for K562 cells (human erythroleukemia cell line) produced by the ENCODE project. Alternative approaches derived quantitative models to predict gene expression using histone modifications and showed that they are well-correlated (Karlic et al., 2010), whereas Cheng et al (2011) derived a support vector machine model from the modENCODE worm data and applied it to human K562 cells and mouse embryonic stem cells with good performance. In all these studies the use of machine learning approaches were successfully employed in quantifying the relationships between histone modifications and TF binding with gene expression levels. One main difference of the approach used in our study is the fact that the modeling step of our pipeline models gene expression variation

between different cell populations, rather than absolute expression levels of a single cell population. Furthermore, the datasets included in the analysis derive from primary cells, thus representing a more physiological condition with respect to the immortalized K562 cell line used in the aforementioned studies. Another important area of application of machine learning is the chromatin-state annotation using combinations of chromatin modification patterns, which represents a powerful approach for discovering important regulatory regions based on genome wide NGS data. Rajagopal et al (2013) developed a Random-Forest based algorithm, RFECS, to integrate histone modification profiles for the identification of enhancers elements in a number of cell-types, whereas Ernst et al (2012) developed ChromHMM, a multivariate Hidden Markov Model based algorithm that explicitly models the observed combination of marks in order to learn and characterize chromatin states across the genome. The importance and 'evolution' of the coordinated application of computational analyses of large scale epigenomic datasets is highlighted in a recent integrative analysis of 111 reference human epigenomes generated as part of the NIH Roadmap Epigenomics Consortium (Kundaje et al., 2015). In this study the authors successfully established global maps of regulatory elements, defined regulatory modules of coordinated activity and their likely activators and repressors, in a wide range of cell types.

Given the rapid development of the field and the establishment of highly accurate global maps of regulatory elements it would be of interest to test how they integrate and compare with our findings and if the inclusion of such single cell chromatin state information can further enhance the predictive power and/or facilitate the interpretation of our approach. In characterizing the genome-wide occupancy profiles of the master erythroid transcription factor GATA1 we identify a highly accurate ensemble of GATA1 target genes in early and late erythroid progenitor cells in mouse fetal liver derived cells (2,590 and 2,826 genes, respectively).

In investigating the underlying structure of the epigenetic landscape of differentially expressed GATA1 target genes (2258 genes) we used the complete set of chromatin state observations to model the changes in their expression levels during terminal erythroid differentiation. This approach resulted in a remarkably highly predictive model (R^2=0.62) of differential gene expression levels by the binding signals of the 4 TFs, 9 histone modifications, RNA pol II and DNA methylation levels measured in $Ter119^-$ and $Ter119^+$ cells. Interestingly, the most predictive features were the variation of H3K79me2 and H3K4 methylation levels, whereas GATA1 variation ranked equally with the variation in the H3K9Ac, RNApolII and H4K16Ac levels. It is interesting to note that the most predictive features (H3K79me2 and H3K4 methylation) can be, at least in part, associated with GATA1 itself, as shown by the fact that GATA1 occupancy is a very good predictor of the variation of these epigenetic

marks during terminal erythroid differentiation. This observation further consolidates the notion that part of the GATA1 regulatory function is exerted through the modulation of the epigenetic landscape of its target genes. It is not clear whether the associations between GATA1 occupancies and specific epigenetic marks are a direct or indirect consequence of GATA1 binding to target DNA sites. Previous work has shown GATA1 to interact with the CBP/p300 histone acetyltransferase (Blobel et al., 1998; Boyes et al., 1998), which may account for the association of GATA1 occupancies with the H3K27Ac mark. However, the association of GATA1 with the H4K16Ac, H3K79me2 or the H3K4-methyl marks cannot be accounted for by the currently known GATA1 interaction partners. Recent work in the Strouboulis lab on the proteomic characterization of GATA1 protein complexes in mouse fetal liver cells, led to the identification of novel epigenetic regulators co-purifying with GATA1, such as EHMT histone methyltransferases or MYST type histone acetyltransferases (Karkoulia et al., unpublished), which may be recruited by GATA1 to target gene loci in effecting specific epigenetic modifications. Furthermore, there may be other epigenetic marks that were not included in our initial genomic analysis, such as H4K12acetylation, and which may be closely associated with changes in GATA1 occupancies.

Collectively, our data reinforce previous observations for GATA1 regulating the erythroid differentiation process at multiple levels. Firstly, GATA1 positively regulates the expression of erythroid specific genes and genes involved in the production of mature hemoglobin molecules. Secondly, it negatively regulates the expression of genes involved in early hematopoietic differentiation and alternative myeloid and lymphoid lineages, by completely shutting them down to allow terminal erythroid differentiation to proceed. Thirdly, it is directly involved in the reduced expression of the mRNA maturation and translation machinery, adjusting it to the reduced needs of the enucleated mature erythrocyte. Importantly, our work shows that specific epigenetic signatures are associated with functionally different subsets of GATA1 target genes, thus suggesting a degree of plasticity in the regulatory functions of GATA1.

One of the three classes of GATA1 target genes identified in our study involves housekeeping genes. Interestingly, further investigation of the GATA1 association with the mRNA translation machinery components that belong to this class, showed a previously unappreciated binding and potential regulation of ribosomal protein (RP) genes by the hematopoietic lineage specific transcription factors GATA1 and PU.1. These observations are significant in light of the fact that ribosomopathies arising from mutations in specific RP genes, invariably present with defective erythropoiesis leading to Diamond Blackfan Anemia (DBA). Recent work has suggested that DBA may be due to defective translation of the full length GATA1 protein, leading to production of an N-terminal truncated GATA1 short (or GATA1s) isoform

which cannot fully sustain proper erythropoiesis (Byrska-Bishop et al., 2015). Our observations on GATA1 binding to RP genes raise an alternative possibility as to how GATA1 may be implicated in ribosomopathies through the direct regulation of DBA associated ribosomal protein genes (Amanatiadou E, Papadopoulos GL et al, 2015, under review). Furthermore, we also find differences in GATA1 binding to RP genes in human fetal versus adult erythropoiesis and in the timing of RP gene decline in expression between human and mouse erythropoiesis. The differential fetal and adult GATA1 occupancy levels in RP genes would suggest a developmental aspect to GATA1 RP gene binding and regulation in man that may also be related to the fact that the median age of onset for DBA is two months after birth, at a time when transition to definitive erythropoiesis is completed.

In characterizing the epigenetic mechanisms involved in erythroid and megakaryocytic cell fate specification processes we applied our computational analysis pipeline to a number of comprehensive sets of genomic data derived from hematopoietic stem cells, MEP cells and erythroid and megakaryocytic lineage committed progenitors. Based on this approach we propose that the chromatin structure of a large group (~1000) of genes transitions from an active state in the hematopoietic stem cell (LSK cells) and mature megakaryocytes, to an inactive state in the erythroid cell population. This indicates a process of specific erasure of active chromatin marks in the erythroid commitment and differentiation pathway. Based on the comparison of DNase hypersensitivity profiles throughout erythroid differentiation we propose that the inactivation process of these gene loci is initiated before the stage of early committed erythroid cells ($CD71^{+}/Ter119^{-}$ cells), thus representing an early step of the erythroid specification process. Furthermore, focusing our approach to the gene-wide occupancies of hematopoietic specific transcription factors, we show that the erythroid specific profile of GATA1, a critical TF in both the erythroid and megakaryocytic lineages, is mirrored by the binding profiles of the TAL1 and LDB1 TFs in hematopoietic stem cells (LSK cells), thus suggesting a potential "bookmarking" function for these factors.

Importantly, we also identify a series of erythroid specific epigenetic modifiers with a potentially active role in the chromatin reconfiguration differences between erythroid and megakaryocytic cells identified above. Among these genes we identify proteins with previously described functions in erythroid and megakaryocytic specification and differentiation, such as Mllt3 and Mst1, providing validation to our approach, as well as, a novel set of epigenetic modifiers was identified bearing highly similar epigenetic and expression profiles, which included the lysine demethylase Jhdm1d/Kdm7a, the histone deacetylase Hdac11 and the ATP-dependent chromatin remodeling protein Cecr2. Interestingly, two of these, Jhdm1d and Hdac11, co-segregated with a highly erythroid specific gene subset (118 genes) characterized by the differential binding profiles of GATA1, TAL1 and LDB1 in different

hematopoietic cell populations (HSCs, erythroid and megakaryocytes), probably accounting for an erythroid specific regulation of these genes, and further supporting their possible implication in the erythroid lineage specification process.

A pro-neural differentiation effect has been ascribed to the Kdm7a demethylase (Huang et al., 2010) further supporting its implication in cellular differentiation and commitment processes. On the other hand, Kdm7a enzymatic activity has been associated with the specific demethylation of H3K9me2 and H3K27me2 marks, which are both associated with transcriptional repression, thus 'diminishing' a potential role for this demethylase in the extensive loss of active chromatin marks, such as H3K4 methylation in terminal erythroid differentiation. Importantly, over-expression of Hdac11 has been recently associated with myelo-proliferative diseases, such as Philadelphia-negative chronic myeloproliferative neoplasms (Skov et al., 2012) and the expansion of myeloid-derived suppressor cells (MDSCs), a heterogeneous population of cells capable of suppressing anti-tumor T cell function in the tumor microenvironment, representing an imposing obstacle in the development of cancer immunotherapeutics (Sahakian et al., 2015). By comparing the immature myeloid cell transition to MDSCs of HDAC11-KO and wild type mice the authors postulate that Hdac11 acts as a gate-keeper of myeloid differentiation. Furthermore, HDAC11 depletion has been shown to be sufficient in causing cell death and inhibiting metabolic activity in HCT-116 colon, PC-3 prostate, MCF-7 breast and SK-OV-3 ovarian cancer cell lines, establishing a more general and critical role in cancer cell survival, thus representing a novel drug target in oncology (Deubzer et al., 2013). The erythroid specific expression pattern observed in this study represents an intriguing gateway as to the elucidation and further characterization of its functions in physiological conditions. Experimental validation of the erythroid specificity of these epigenetic modifiers will further validate their role in the erythroid specification process and will further elucidate their molecular mechanism of action. For example validation of the involvement of Hdac11 and Kdm7a could be assessed by the use of shRNA silencing and/or by the use of specific inhibitors, such as mocetinostat and daminozide, respectively, on human multipotent hematopoietic progenitor cells ($CD34^{+}$) and monitoring for impaired erythroid differentiation potential using in vitro differentiation assays/protocols. Furthermore, if we apply a 'Guilt by association' principle, co-segregation of previously uncharacterized genes with highly specific erythroid related genes, such as the one described above, could be a strong indication of functional specificity. As an example, the erythroid specific gene subset identified above composed by important erythroid specific TFs such as Lmo2 and Eklf, heme biosynthesis related enzymes such as Fech, Alad, Uros and Urod and the potential erythroid specific modifiers Hdac11 and Kdm7a also include 6030468B19Rik, a protein coding uncharacterized gene. Given the fact that the isolation of this sub-cluster

from the original full gene dataset (over 10000 genes) depends on the combinatorial genomic similarity (TF occupancy data) of these genes, rather than an a priori defined hypothesis (e.g. genes bound by GATA1 only in erythroid cells), provides an intriguing indication of an erythroid function for this gene, due to its highly erythroid specific regulatory pattern. The guilt by association principle also highlights one of the main characteristics of systemic approaches that do not rely on previous knowledge to define relational networks, which is their hypothesis generating potential. Importantly, the type of association identified with our approach will largely depend on the choice of the initial dependent variable and predictors included in the modeling step, thus allowing for a high degree of flexibility in the final interpretation of the results.

Due to the very high resolution of the genomic patterns produced by NGS approaches, data visualization is an active component of NGS analyses. Furthermore, due to fact that NGS data rely on complex computational analysis approaches, tight integration between visualizations, analyses and data will transform readers' abilities to evaluate and to inspect results. To facilitate the interrogation and interpretation of the results produced by the computational analysis proposed in this study, we developed Ariadne (aegeas.imbb.forth.gr/Ariadne/), a web based portal that allows the user to navigate through the gene wide distribution patterns of chromatin features in different hematopoietic cell populations. In addition, the portal includes an interactive graphical display of the read density profiles (primary data) of single gene loci, allowing for the generation of hypotheses about the mechanisms of differential gene regulation of specific genes, as well as for the design of validation experiments. Within Ariadne, a gene can be interrogated both as part of a tree branch, derived from the similarity of its genomic context (TreeViewer module) and as single gene locus to allow access to primary sequencing data, as well as to primary DNA sequence information and the overall locus organization (alternative transcript and neighboring genes information), (GeneViewer module). To facilitate the integration between the computational analysis results (clustering) and the visualization of their primary data several functions were implemented that allow manipulation and selection of specific gene subsets to be exported to the GeneViewer. Web based visualization tools of read density profiles (e.g. UCSC Genome Browser) are currently slow and too complex to use because of the enormous amount of data they need to process and store. The highly targeted content of Ariadne and the intuitive user friendly interface greatly accelerate and simplify the comparison of different binding profiles, as well as the comparison of distinct binding profiles along different gene loci. Furthermore, the integration between the computational analysis results (clustering) and the visualization of their primary data greatly facilitates the interrogation of subsets of genes with similar genomic and regulatory patterns. On the other hand, Ariadne is limited in the genomic regions included

in the database (more similar to a Gene Browser, rather than a full Genome Browser, like UCSC), the amount of primary data included and the lack of user data upload functionality. However, to further extend the user's ability to manipulate the results, future plans include a direct submission of selected gene subsets into the computational analysis pipeline, as well as the selection of the genomic predictors to include in the secondary clustering approach. Furthermore, direct submission of user primary data (bigwig format), significantly enriched genomic regions (bed format) and submission of user defined gene lists will be implemented in Ariadne. Finally, further extension of gene loci to include, possibly, the whole genome, as well as the incorporation of more primary sequencing datasets, beyond hematopoietic cells, will enhance Ariadne's functionality.

An important resource of hematopoietic specific NGS data is currently available by the Haemcode repository, part of the Codex initiative (Sanchez-Castillo et al., 2015). Even though Haemcode is currently far more comprehensive than Ariadne in the number of datasets included, it does not provide a combinatorial analysis of the data or a primary data visualization tool. Importantly, the extensive collection, curation and uniform processing of hematopoietic related NGS datasets provided by the Haemcode database represents an excellent source of primary data for modeling hematopoietic differentiation processes and for systematically exploring causal relationships between gene expression and histone modifications, TF binding and histone modifications and sequential TF complex formation from genome-wide data. In fact, one of the main future prospects of this study is the expansion of the model to include non myeloid lineages, thus allowing for a global comparison of the epigenetic events that determine hematopoietic stem cell commitment and differentiation. Further challenges include the incorporation of genome organization information derived from Hi-C and 5C approaches, as and when they will become available. In fact, recent studies have used Random Forest modeling to associate chromatin modifications with changes in interaction frequency between genomic loci and also describe extensive chromatin reorganization during lineage specification of human embryonic stem cells and four human ES-cell-derived lineages (Dixon et al., 2015). Thus, inclusion of chromatin dynamics data will add to our current understanding of blood lineage specification and differentiation processes. Finally, an important landmark for molecular mechanisms identified by cell population based data, such as current NGS datasets, will be their comparison and validation with the currently developing technologies of single-cell genomics and proteomics.

Materials And Methods

Materials And Methods

GATA1 ChIP sequencing and data analysis

Formaldehyde-crosslinked chromatin from 10^7 Ter119$^+$ or Ter119$^-$ cells was prepared as previously described (Schuh et al., 2005). Pilot experiments showed that rabbit polyclonal antibody Ab11835 (Abcam) gave the highest enrichment for GATA1 binding to the -3.5kb HS1 of the GATA1 gene locus. Anti-GATA1 ChIPed DNA from Ter119$^-$ and Ter119$^+$ chromatin and from a 'no antibody' control (input DNA) were processed for deep sequencing using the Illumina Genome Analyzer II platform according to Illumina protocols (www.illumina.com). Deep sequencing was carried out in duplicate for each ChIP sample and once for input DNA controls. All 51-nucleotide sequence reads thus produced were mapped to the NCBI37/mm9 Mouse Genome Assembly using the Eland software (Illumina). Sequence reads with multiple genome alignments and/or more than 2 nucleotide mismatches were excluded. Peak calling was performed using the QuEST algorithm (Valouev et al., 2008). Each sample was analyzed versus the control dataset using a strict Fold Change (Sample/Control > 50), a False Discovery Rate threshold of 0.001 and a peak score threshold of 70. Sequencing data will be deposited in EBI's European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena/) prior to publication.

GATA1 Target Gene identification

Gene location data used in the mapping analyses were extracted from BioMart (Guberman et al., 2011) using the Ensembl transcript database (build 59). Gene mapping analyses were carried out by custom in-house Perl and R scripts. RNA-sequencing gene expression data for Ter119$^-$ and Ter119$^+$ erythroid cells were downloaded from the GEO database (Barrett et al., 2013) using accession number GSE32110. Total Gene Score (TGS) was calculated for each expressed gene separately and corresponds to the sum of the enrichment values of GATA1

peaks that overlap the gene's Transcription Start Site (TSS) within a given distance window. Target genes with TGS scores lower than 100 were discarded from subsequent analyses.

Random Forest Regressors

Random Forest (RF) (Breiman, 2001) is a type of non-linear regression implemented by the 'randomForest' R package. The optimal mtry value (number of variables randomly sampled as candidates at each split) was selected separately for each RF model by growing a small number of decision trees (ntree=50) for 100 independent runs. We selected the mtry value that returned the highest mean R^2 value, measured on the out-of-bag (OOB) data, amongst all test runs. A minimum number of 1,000 and up to 2,000 decision trees (ntree) where grown for all the reported RF models, using the optimal mtry value. The RF parameters used in each case are reported in 3.10.

RF regression models for the epigenetic marks were calculated on the subset of expressed genes that presented a mean normalized mark TGS score that was greater than 1. Modelling of the gene-expression fold-change model based on all the occupancy profiles available for $Ter119^-$ and or $Ter119^+$ cells was performed on the differentially regulated subset of GATA1 target genes using a 2-fold gene expression change cut-off. In order to calculate the coefficient of determination (R^2) of each regressor in an unbiased way, each model was trained on the 85% of the data (training set); the coefficient of determination was then calculated on the remaining 15% test-set (out-of-sample R^2). In order to avoid any bias introduced by the random choice of the test set, sampling was repeated 100 times. The final R^2 value reported for each model represents the mean performance of the 100 sampling, training and testing events. The R^2 value was calculated as $1 - E_{model} / E_{mean}$, where E_{model} is the sum of squared residuals of the fitted RF regression model on the test set, and E_{mean} is the sum of squared residuals of the trivial regression model. The latter is computed as the mean value of the dependent variable on the training set, always predicting this mean value. Importantly, we observe very little difference between the R^2 values calculated on our hold out set and the R^2 values reported by the internal estimation of the model's performance calculated on the out-of-bag (OOB) data in all of the proposed regressors. Based on this observation, the proximity values and variable importance measures reported here where calculated by fitting a single RF non-linear regression model without splitting the sample dataset. R^2 values calculated on the training (85%), testing (15%) and full datasets for all the proposed regressors are reported in Figure 3.10.

Dependent Variable	Predictors	# of Predictors	# of Samples	ntree	mtry	Training (85%)	Testing (15%)	Sampling	Training R^2	Testing R^2	Full Dataset R^2
Gene Expression Change (log2(FC))	GATA-1 (1kb TSS)	11	1414	1000	1	1202	212	100	0.06	0.06	0.06
Gene Expression Change (log2(FC))	GATA-1 (2kb TSS)	11	1926	1000	1	1637	289	100	0.07	0.07	0.08
Gene Expression Change (log2(FC))	GATA-1 (5kb TSS)	11	2740	1000	1	2329	411	100	0.12	0.12	0.12
Gene Expression Change (log2(FC))	GATA-1 (10kb TSS)	11	3651	1000	1	3103	548	100	0.13	0.14	0.14
Gene Expression Change (log2(FC))	GATA-1 (20kb TSS)	11	5098	1000	1	4333	765	100	0.13	0.13	0.13
Gene Expression Change (log2(FC))	GATA-1 (20kb TSS 10kb TES)	11	6131	1000	1	5211	920	100	0.12	0.12	0.12
Gene Expression Change (log2(FC))	GATA-1 (NEAREST GENE)	11	3926	1000	1	3337	589	100	0.13	0.13	0.13
Gene Expression Change (log2(FC))	4 TFs, RNApol2, 9 Hist. Mod., DNA m	61	2261	2000	30	1922	339	100	0.617	0.616	0.620
GSM688817_H3K4me2 (ΔTGS)	GATA-1	5	6282	1000	2	5340	942	100	0.36	0.35	0.37
GSM688818_H3K4me3 (ΔTGS)	GATA-1	5	5795	1000	2	4926	869	100	0.29	0.29	0.31
GSM688819_H3K9Ac (ΔTGS)	GATA-1	5	4742	1000	1	4031	711	100	0.22	0.20	0.23
GSM688823_H4K16Ac (ΔTGS)	GATA-1	5	4889	1000	1	4156	733	100	0.29	0.28	0.30
GSM688821_H3K36me3 (ΔTGS)	GATA-1	5	4041	1000	1	3435	606	100	0.18	0.19	0.21
GSM688822_H3K79me2 (ΔTGS)	GATA-1	5	5031	1000	1	4276	755	100	0.31	0.31	0.32
GSM688820_H3K27me3 (ΔTGS)	GATA-1	5	3905	1000	1	3319	586	100	0.06	0.07	0.07
GSM688824_RNAPolII (ΔTGS)	GATA-1	5	3588	1000	1	3050	538	100	0.23	0.24	0.26
GSM688817_H3K4me2 (ΔTGS)	GATA-1, TAL-1, EKLF	11	6282	1000	3	5340	942	100	0.56	0.55	0.56
GSM688818_H3K4me3 (ΔTGS)	GATA-1, TAL-1, EKLF	11	5795	1000	4	4926	869	100	0.47	0.47	0.48

Figure 3.10: Random Forest parameters and results

Mouse Fetal Liver Genomic Occupancy Database

The genome-wide TF occupancy and histone modification profiles presently available for Ter119^{-} and Ter119^{+} fetal liver cells were downloaded from the GEO database using accession numbers GSE27893 (H3K4me2, H3K4me3, H3K9Ac, H3K27me3, H3K36me3, H3K79me2, H4K16Ac, RNApolII), GSE27918 (H3K4me1, H3K27Ac), GSE18720 (TAL1 Ter119^{-}), GSE30142 (SCL/TAL1 Ter119^{+}) and GSE21950 (PU.1). For the GSE27893 datasets, genomic occupancy profiles (wig files) were provided in 25nt genomic bins whereas MACS (Zhang et al., 2008) with default parameters was used to create density profiles for the GSE27918, GSE18720, GSE30142 and GSE21950 datasets. TGS score for each genomic feature was calculated as the sum of the background corrected number of reads present in the density profiles that mapped within a 10kb window upstream or downstream a gene's TSS. Subsequently, TGS scores were mean normalized in all datasets.

SRA Data Processing and Ariadne Database

The Ariadne database currently includes a total of 52 mouse and 56 human NGS datasets (Tables 3.2 and 3.3, respectively). All datasets are directly retrieved in a Sequence Read Archive (SRA) through Gene Expression Omnibus (GEO) (Barrett et al., 2013), European Nucleotide Archive (ENA) (Leinonen et al., 2011) and Encode (ENCODE, 2012) databases. SRA files are converted to raw sequence reads (fastq format) using the SRA toolkit (fastq-dump, (Leinonen et al., 2011)). Fastq files are then aligned to the mm9 (mouse) or hg19 (human) genome assembly using the bowtie.2 algorithm (Langmead and Salzberg, 2012). Normalized read density profiles (wig files) are produced with MACS (Zhang et al., 2008) using default parameters. Read density measures of specific genomic regions (25kb upstream the annotated TSS to 25kb downstream the annotated TES of UCSC annotated genes) are exported in tabular format using bwtool (Pohl and Beato, 2014) 'matrix' function for all available NGS datasets. Each table produced in this way is then exported to Ariadnes' mySQL database. Each genomic region (gene) is stored in a separate table in the database and named based on the gene symbol on which the genomic region is centered. Each read density profile is stored as a TEXT type entry using the NGS datasets' name as key. This configuration allows for new datasets to be added to the current database in a time efficient manner. The actual read density profiles are then retrieved by the server in a JSON format using an asynchronous Ajax call. In this way all subsequent functions are then performed on the client side.

Genomic coordinates included in the database are stored in a separate database that uses the gene symbol as a key. Gene and transcript coordinates are based in the UCSC

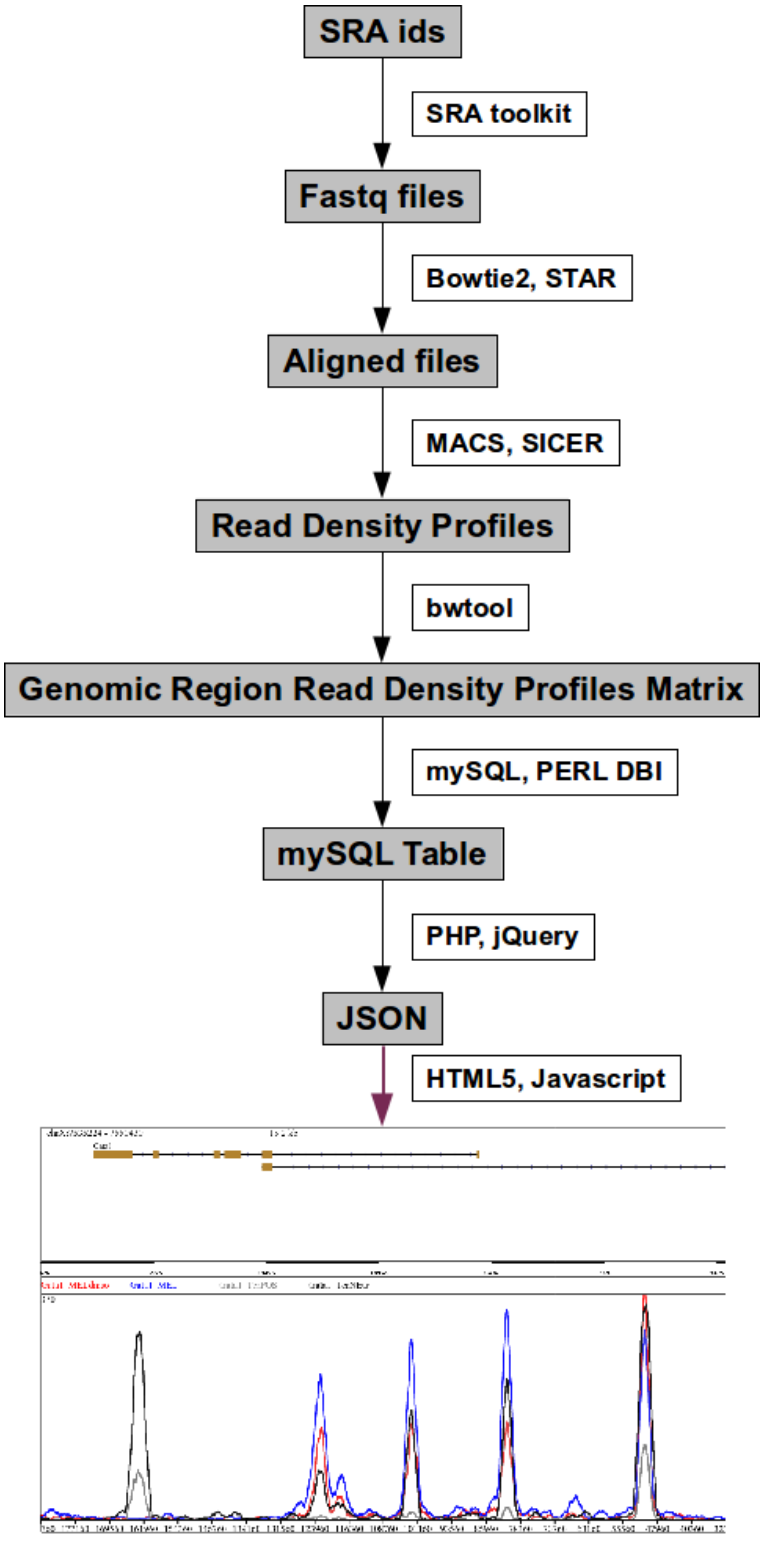


Figure 3.11: Ariadne Data Processing, Storage and Visualization Pipeline

mm9 and hg19 assemblies and have been downloaded from the iGenomes collection (http://support.illumina.com/sequencing/sequencing_software/igenome.html).

Sequence data for each genomic *locus* are retrieved using Bioconductor (Gentleman et al., 2004) BSgenome package for human and mouse sequences. Each sequence is stored in the mm9 or hg19 sequence table as a TEXT type entry using the gene symbol as key.

Datasets Used in This Study

Table 3.1: Chromatin Features Used in RNAseq evaluation. TGS, maximum peak score and TSS distance measures were included in the training datasets

(a) Histone Tail Modifications

Target	Cells
H3K27me3	ERY
H3K27me3	MEG
H3K36me3	ERY
H3K36me3	MEG
H3K4me1	ERY
H3K4me1	MEG
H3K4me3	ERY
H3K4me3	MEG
H3K9me3	ERY
H3K9me3	MEG

(b) Transcription Factors

Target	Cells
Fli1	MEG
Gata1	BoneMarrow
Gata1	ERY
Gata1	MEG
Gata1	Ter119neg
Gata1	Ter119pos
Gata2	mHSC
Ldb1	mHSC
Ldb1	CFCs
Ldb1	BoneMarrow
Tal1	mHSC
Tal1	Ter119neg
Tal1	ERY
Tal1	MEG
Tal1	BoneMarrow

Table 3.2: Chromatin Features Used in mouse RF model training. **a)** Training sets:
i) Differentially expressed genes, 25 predictors, 1623 samples, ntree=2000, mtry=8, R^2=81.91%,
ii) Expressed genes, 25 predictors, 11949 samples, ntree=2000, mtry=8, R^2=54.55%
b) Training set: Expressed genes, 27 predictors, 9955 samples, ntree=2000, R^2=24.06%

(a) All Datasets

Target	Cells
H3K27me3	ERY
H3K27me3	MEG
H3K27me3	mHSC
H3K36me3	ERY
H3K36me3	MEG
H3K4me1	ERY
H3K4me1	MEG
H3K4me3	ERY
H3K4me3	MEG
H3K4me3	mHSC
H3K79me2	mHSC
DNAmeth	CMP
DNAmeth	ERY
DNAmeth	mHSC
Fli1	MEG
Gata1	ERY
Gata1	MEG
Gata2	mHSC
Ldb1	CFCs
Ldb1	mHSC
p300	MEG
Runx1	MEG
Tal1	ERY
Tal1	MEG
Tal1	mHSC

(b) Transcription Factors

Target	Cells
CEBPa	GMP
CEBPa	LSK
Ctcf	BoneMarrow
Ctcf	Ter119pos
DNAmeth	CMP
DNAmeth	ERY
DNAmeth	mHSC
Eklf	Ter119neg
Eklf	Ter119pos
Fli1	MEG
Gata1	BoneMarrow
Gata1	ERY
Gata1	MEG
Gata1	Ter119neg
Gata1	Ter119pos
Gata2	mHSC
Ldb1	BoneMarrow
Ldb1	CFCs
Ldb1	mHSC
Nfe2	Ter119pos
p300	MEG
Runx1	MEG
Tal1	BoneMarrow
Tal1	ERY
Tal1	MEG
Tal1	mHSC
Tal1	Ter119neg

Table 3.3: Chromatin Features Used in Human RF model Training. Training sets:
i) Differentially expressed genes, 58 predictors, 1832 samples, ntree=2000, mtry=19, R^2=73.31%,
ii) Expressed genes, 58 predictors, 14111 samples, ntree=2000, mtry=19, R^2=49.84%

Target	Cells
Ctcf	ProEs Adult
Ctcf	ProEs Fetal
Gata1	ProEs Adult
Gata1	ProEs Fetal
Nfe2	ProEs Adult
Nfe2	ProEs Fetal
Rad21	ProEs Adult
Rad21	ProEs Fetal
RNApol2	ProEs Adult
RNApol2	ProEs Fetal
RNApol2	ProEs Adult HS
RNApol2	ProEs Fetal HS
Tal1	ProEs Adult
Tal1	ProEs Fetal
H3K4me1	ProEs Fetal
H3K4me1	ProEs Adult
H3K4me2	ProEs Fetal
H3K4me2	ProEs Adult
H3K4me3	ProEs Fetal
H3K4me3	ProEs Adult
H3K9me3	ProEs Fetal
H3K9me3	ProEs Adult
H3K27me3	ProEs Fetal
H3K27me3	ProEs Adult
H3K36me2	ProEs Fetal
H3K36me2	ProEs Adult
H3K36me3	ProEs Fetal
H3K36me3	ProEs Adult
Fli1	hMEG
Gata1	hMEG
Gata2	hMEG

Table 3.3: Chromatin Features Used in Human RF model Training. Training sets:
i) Differentially expressed genes, 58 predictors, 1832 samples, ntree=2000, mtry=19, R^2=73.31%,
ii) Expressed genes, 58 predictors, 14111 samples, ntree=2000, mtry=19, R^2=49.84%

Target	Cells
Runx1	hMEG
Tal1	hMEG
H3K4me3	CD34 1
H3K4me3	CD34 2
H3K4me3	CD34 cult
H2AZ	hCD133
H2AZ	hCD36
H3K27me1	hCD133
H3K27me1	hCD36
H3K27me3	hCD133
H3K27me3	hCD36
H3K36me3	hCD133
H3K36me3	hCD36
H3K4me1	hCD133
H3K4me1	hCD36
H3K4me3	hCD133
H3K4me3	hCD36
H3K9me1	hCD133
H3K9me1	hCD36
H3K9me3	hCD133
H3K9me3	hCD36
H4K20me1	hCD133
H4K20me1	hCD36
Input	hCD36
RNApol2	hCD133
RNApol2	hCD36
Input	hCD133

Table 3.4: SRA datasets

Human			Mouse		
SRA ID	Target	Cell	SRA ID	Target	Cell
SRR016977	H2AZ	hCD133	SRR1106099	RNAseq	Proerythroblast
SRR016978	H2AZ	hCD133	SRR1106100	RNAseq	Proerythroblast
SRR016979	H3K27me1	hCD133	SRR1106101	RNAseq	Proerythroblast
SRR016980	H3K27me1	hCD133	SRR1106102	RNAseq	Basophilic
SRR016981	H3K27me3	hCD133	SRR1106103	RNAseq	Basophilic
SRR016982	H3K27me3	hCD133	SRR1106104	RNAseq	Basophilic
SRR016983	H3K27me3	hCD133	SRR1106105	RNAseq	Polychromatic
SRR016984	H3K36me3	hCD133	SRR1106106	RNAseq	Polychromatic
SRR016985	H3K36me3	hCD133	SRR1106107	RNAseq	Polychromatic
SRR016986	H3K36me3	hCD133	SRR1106108	RNAseq	Orthochromatic
SRR016987	H3K4me1	hCD133	SRR1106109	RNAseq	Orthochromatic
SRR016988	H3K4me1	hCD133	SRR1106110	RNAseq	Orthochromatic
SRR016989	H3K4me3	hCD133	SRR1157326	H3K4me2	$Ter119^{+}$
SRR016990	H3K9me1	hCD133	SRR1157327	H3K4me3	$Ter119^{+}$
SRR016991	H3K9me1	hCD133	SRR1157328	H3K9Ac	$Ter119^{+}$
SRR016992	H3K9me1	hCD133	SRR1157329	H3K27me3	$Ter119^{+}$
SRR016993	H3K9me3	hCD133	SRR1157330	H3K36me3	$Ter119^{+}$
SRR016994	H3K9me3	hCD133	SRR1157331	H3K79me2	$Ter119^{+}$
SRR016995	H3K9me3	hCD133	SRR1157332	H4K16Ac	$Ter119^{+}$
SRR016996	H4K20me1	hCD133	SRR1157333	RNApol2	$Ter119^{+}$
SRR016997	H4K20me1	hCD133	SRR1157334	H3K4me2	$Ter119^{-}$
SRR016998	H4K20me1	hCD133	SRR1157335	H3K4me3	$Ter119^{-}$
SRR016999	RNApol2	hCD133	SRR1157336	H3K9Ac	$Ter119^{-}$
SRR017000	H2AZ	hCD36	SRR1157337	H3K27me3	$Ter119^{-}$
SRR017001	H2AZ	hCD36	SRR1157338	H3K36me3	$Ter119^{-}$
SRR017002	H3K27me1	hCD36	SRR1157339	H3K79me2	$Ter119^{-}$
SRR017003	H3K27me1	hCD36	SRR1157340	H4K16Ac	$Ter119^{-}$
SRR017004	H3K27me1	hCD36	SRR1157341	RNApol2	$Ter119^{-}$
SRR017005	H3K27me3	hCD36	SRR1209627	H3K4me1	MES
SRR017006	H3K27me3	hCD36	SRR1209629	H3K27ac	MES
SRR017007	H3K27me3	hCD36	SRR1209631	Ezh2	MES
SRR017008	H3K36me3	hCD36	SRR1209633	Input	MES

Table 3.4: SRA datasets

Human			Mouse		
SRA ID	Target	Cell	SRA ID	Target	Cell
SRR017009	H3K36me3	hCD36	SRR1209635	Lsd1	MES
SRR017010	H3K36me3	hCD36	SRR1209645	RRBS	mESC
SRR017011	H3K4me1	hCD36	SRR1209646	RRBS	MES
SRR017012	H3K4me1	hCD36	SRR1209648	RRBS	MEL
SRR017013	H3K4me3	hCD36	SRR1632492	RNAseq	erythroid R2
SRR017014	H3K9me1	hCD36	SRR1632493	RNAseq	erythroid R3
SRR017015	H3K9me1	hCD36	SRR1632494	RNAseq	erythroid R4
SRR017016	H3K9me1	hCD36	SRR1632495	RNAseq	erythroid R5
SRR017017	H3K9me3	hCD36	SRR1652695	Hand1	MES
SRR017018	H3K9me3	hCD36	SRR330833	PU1	FLDN1 Tcells
SRR017019	H3K9me3	hCD36	SRR330834	PU1	FLDN2a Tcells
SRR017020	H4K20me1	hCD36	SRR330835	PU1	FLDN2b Tcells
SRR017021	H4K20me1	hCD36	SRR330840	Input	FLDN1 Tcells
SRR017022	H4K20me1	hCD36	SRR330841	Input	FLDN2a Tcells
SRR017023	RNApol2	hCD36	SRR330842	Input	FLDN2b Tcells
SRR054910	Fli1	HPC7	SRR363837	Pu1	Spleen
SRR054911	Gata2	HPC7	SRR363838	Pu1	Spleen
SRR054912	Gfi1b	HPC7	SRR363841	Input	Spleen
SRR054913	Input	HPC7	SRR363842	Input	Spleen
SRR054914	Lmo2	HPC7	SRR414936	H3K27Ac	mESC
SRR054915	Lyl1	HPC7	SRR414937	H3K27me3	mESC
SRR054916	Meis1	HPC7	SRR414938	H3K36me3	mESC
SRR054917	Pu1	HPC7	SRR414939	H3K4me1	mESC
SRR054918	Runx1	HPC7	SRR414940	H3K4me2	mESC
SRR054919	Tal1	HPC7	SRR414941	H3K4me3	mESC
SRR070375	Gata1	hMEG	SRR475675	Pu1	Spleen
SRR070376	Gata2	hMEG	SRR475676	Input	Spleen
SRR070377	Runx1	hMEG	SRR549333	mRNA	MEP
SRR070378	Fli1	hMEG	SRR549334	mRNA	MEP
SRR070379	Tal1	hMEG	SRR549349	mRNA	ERY
SRR070380	Input	hMEG	SRR549350	mRNA	ERY
SRR091684	Input	hCD36	SRR549357	mRNA	MEG

Table 3.4: SRA datasets

Human			Mouse		
SRA ID	Target	Cell	SRA ID	Target	Cell
SRR091685	Input	hCD36	SRR549358	mRNA	MEG
SRR094805	Input	HSPC	SRR578999	DNase	$KIT^{+}CD71^{-}TER119^{-}$
SRR094806	Tal1	HSPC	SRR579000	DNase	$KIT^{+}CD71^{-}TER119^{-}$
SRR094807	Tal1	HSPC	SRR579001	DNase	$KIT^{+}CD71^{-}TER119^{-}$
SRR094808	Pu1	HSPC	SRR579002	DNase	$KIT^{-}CD71^{+}TER119^{+}$
SRR094809	Pu1	HSPC	SRR579003	DNase	$KIT^{-}CD71^{+}TER119^{+}$
SRR1041830	H3K27me3	hERY	SRR579004	DNase	$KIT^{-}CD71^{+}TER119^{+}$
SRR1041831	H3K27Ac	hERY	SRR579006	DNase	$KIT^{-}CD71^{+}TER119^{+}$
SRR1041832	Gata1	hERY	SRR579007	DNase	$KIT^{+}CD71^{+}TER119^{+}$
SRR1041833	Tal1	hERY	SRR579008	DNase	$KIT^{+}CD71^{+}TER119^{+}$
SRR1041834	Gfi1b	hERY	SRR579009	DNase	$KIT^{+}CD71^{+}TER119^{+}$
SRR1041835	Input	hERY	SRR579010	DNase	$KIT^{+}CD71^{+}TER119^{+}$
SRR1106084	RNAseq	Proerythroblast	SRR579011	DNase	$KIT^{+}CD71^{+}TER119^{+}$
SRR1106085	RNAseq	Proerythroblast	SRR579012	DNase	$KIT^{+}CD71^{+}TER119^{-}$
SRR1106086	RNAseq	Proerythroblast	SRR579013	DNase	$KIT^{+}CD71^{+}TER119^{-}$
SRR1106087	RNAseq	BasophilicEarly	SRR579014	DNase	$KIT^{+}CD71^{+}TER119^{-}$
SRR1106088	RNAseq	BasophilicEarly	SRR579136	DNase	Total Fetal Liver E14.5
SRR1106089	RNAseq	BasophilicEarly	SRR579137	DNase	Total Fetal Liver E14.5
SRR1106090	RNAseq	BasophilicLate	SRR579138	DNase	Total Fetal Liver E14.5
SRR1106091	RNAseq	BasophilicLate	SRR579139	DNase	Total Fetal Liver E14.5
SRR1106092	RNAseq	BasophilicLate	SRR579140	DNase	Total Fetal Liver E14.5
SRR1106093	RNAseq	Polychromatic	SRR611721	H3K4me2	mHSC
SRR1106094	RNAseq	Polychromatic	SRR858755	Tal1	MES
SRR1106095	RNAseq	Polychromatic	SRR858756	Tal1	MES
SRR1106096	RNAseq	Orthochromatic	SRR858758	Input	MES
SRR1106097	RNAseq	Orthochromatic	SRR858759	Input	MES
SRR1106098	RNAseq	Orthochromatic	SRR858760	Tal1	MES GataKD
SRR452927	H3K4me1	Fetal	SRR858761	Input	MES GataKD
SRR452928	H3K4me1	Adult	SRR858764	Tal1	MEL
SRR452929	H3K4me2	Fetal	SRR892966	H3K27me3	mHSC
SRR452930	H3K4me2	Adult	SRR892967	H3K27me3	mHSC
SRR452931	H3K4me3	Fetal	SRR892968	H3K27me3	mHSC

Table 3.4: SRA datasets

Human			Mouse		
SRA ID	Target	Cell	SRA ID	Target	Cell
SRR452932	H3K4me3	Adult	SRR892972	H3K36me3	mHSC
SRR452933	H3K9me3	Fetal	SRR892973	H3K36me3	mHSC
SRR452934	H3K9me3	Adult	SRR892974	H3K36me3	mHSC
SRR452935	H3K27me3	Fetal	SRR892978	H3K4me3	mHSC
SRR452936	H3K27me3	Adult	SRR892979	H3K4me3	mHSC
SRR452937	H3K36me2	Fetal	SRR892995	RNAseq	mHSC
SRR452938	H3K36me2	Adult	SRR892996	RNAseq	mHSC
SRR452939	H3K36me3	Fetal	SRR892997	RNAseq	mHSC
SRR452940	H3K36me3	Adult	SRR892998	RNAseq	mHSC
SRR452941	H3K9Ac	Fetal	SRR064917	PU1	proBcells
SRR452942	H3K9Ac	Adult	SRR064918	PU1	proBcells
SRR452943	H3K27Ac	Fetal	SRR064930	Input	proBcells
SRR452944	H3K27Ac	Adult			
SRR452947	Tal1	Fetal			
SRR452948	Tal1	Adult			
SRR452951	Nfe2	Fetal			
SRR452952	Nfe2	Adult			
SRR452957	Ctcf	Fetal			
SRR452958	Ctcf	Adult			
SRR452959	Rad21	Fetal			
SRR452960	Rad21	Adult			
SRR452961	RNApol2	Fetal			
SRR452962	RNApol2	Adult			
SRR452965	Input	Fetal			
SRR452966	Input	Adult			
SRR524934	Gata1	Fetal			
SRR524935	Gata1	Adult			
SRR524936	RNApol2	HS			
SRR524937	RNApol2	HS			
SRR524938	Irf2	Adult			
SRR524939	Input	HS			
SRR524940	Input	HS			

Table 3.4: SRA datasets

Human			Mouse		
SRA ID	Target	Cell	SRA ID	Target	Cell
SRR525259	Brg1	hCD4			
SRR525260	Brg1	hHSC			
SRR525261	p300	hHSC			
SRR525262	Brg1	hCD36			
SRR525263	p300	hCD36			
SRR525264	H2AZ	Bcell			
SRR525265	H2AZ	Bcell			
SRR525266	H3K27me1	Bcell			
SRR525267	H3K27me3	Bcell			
SRR525268	H3K4me1	Bcell			
SRR525269	H3K4me3	Bcell			
SRR525270	Input	Bcell			
SRR525271	Brg1	Bcell			
SRR525272	RNApol2	Bcell			
SRR772106	Erg	hHSC			
SRR772107	Fli1	hHSC			
SRR772108	Tal1	hHSC			
SRR772109	Lyl1	hHSC			
SRR772110	Gata2	hHSC			
SRR772111	Runx1	hHSC			
SRR772112	Lmo2	hHSC			
SRR772113	Input	hHSC			

Acknowledgements

I would like to thank my scientific supervisor, Dr. John Strouboulis, and my academic supervisor, professor George Garinis, for all their guidance, support and understanding during this whole project. I would also like to thank professor Ioannis Tsamardinos for introducing me to the critical aspects of machine learning techniques, professor Charalampos Spilianakis for critical discussions and comments on progress reports and meetings and all the members of the examination committee for their constructive comments.

Particular thanks to Dr Elena Karkoulia for performing all the wet lab experiments included in this study, as well as, all current and past members of the JS310 lab for critical scientific and less scientific discussions had in the past years. I would also like to thank Carmelo Papadopoulos and Jonathan Panagiotides from quietroom.eu for their help in setting up Ariadne's graphical interface and for developing the 'DendroApp' function.

This project was funded by the 'InteGer' FP7 Marie Curie Initial Training Network [PITN-GA-2008-214902] and by a grant from 'Fondazione Cariplo'.

References

J. Adolfsson, R. Mansson, N. Buza-Vidas, A. Hultquist, K. Liuba, C. Jensen, D. Bryder, L. Yang, O. Borge, L. Thoren, K. Anderson, E. Sitnicka, Y. Sasaki, M. Sigvardsson, and S. Jacobsen. Identification of flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic commitment. *Cell, PMID: 15851035*, 121(2):295–306, 2005.

K. Akashi, D. Traver, T. Miyamoto, and I. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature, PMID: 10724173*, 404(6774): 193–7, 2000.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol, PMID: 20979621*, 11(10):R106, 2010.

P. Aplan, K. Nakahara, S. Orkin, and I. Kirsch. The scl gene product: a positive regulator of erythroid differentiation. *EMBO J, PMID: 1396592*, 11(11):4073–81, 1992.

T. Barrett, S. Wilhite, P. Ledoux, C. Evangelista, I. Kim, M. Tomashevsky, K. Marshall, K. Phillippy, P. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. Robertson, N. Serova, S. Davis, and A. Soboleva. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Res, PMID: 23193258*, 41(Database issue):D991–5, 2013.

T. Barth and A. Imhof. Fast signals and slow marks: the dynamics of histone modifications. *Trends Biochem Sci, PMID: 20685123*, 35(11):618–26, 2010.

B. Bernstein, T. Mikkelsen, X. Xie, M. Kamal, D. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. Schreiber, and E. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell, PMID: 16630819*, 125(2):315–26, 2006.

G. Blobel, T. Nakajima, R. Eckner, M. Montminy, and S. Orkin. Creb-binding protein cooperates with transcription factor gata-1 and is required differentiation. *Proc Natl Acad Sci U S A, PMID: 9482838*, 95(5):2061–6, 1998.

J. Boyes, P. Byfield, Y. Nakatani, and V. Ogryzko. Regulation of activity of the transcription factor gata-1 by acetylation. *Nature, PMID: 9859997*, 396(6711):594–8, 1998.

L. Breiman. Random forests. *Machine Learning, October 2001*, 45(1):5–32, 2001.

E. Bresnick, H. Lee, T. Fujiwara, K. Johnson, and S. Keles. Gata switches as developmental drivers. *J Biol Chem, PMID: 20670937*, 285(41):31087–93, 2010.

M. Byrska-Bishop, D. VanDorn, A. Campbell, M. Betensky, P. Arca, Y. Yao, P. Gadue, F. Costa, R. Nemiroff, G. Blobel, D. French, R. Hardison, M. Weiss, and S. Chou. Pluripotent stem cells reveal erythroid-specific activities of the gata1 n-terminus. *J Clin Invest, PMID: 25621499*, 125(3):993–1005, 2015.

A. Cantor and S. Orkin. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene, PMID: 12032775*, 21(21):3368–76, 2002.

E. Chen, C. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics, PMID: 23586463*, 14:128, 2013.

C. Cheng, K. Yan, K. Yip, J. Rozowsky, R. Alexander, C. Shou, and M. Gerstein. A statistical framework for modeling gene expression using chromatin features and datasets. *Genome Biol, PMID: 21324173*, 12(2):R15, 2011.

C. Cheng, R. Alexander, R. Min, J. Leng, K. Yip, J. Rozowsky, K. Yan, X. Dong, S. Djebali, Y. Ruan, C. Davis, P. Carninci, T. Lassman, T. Gingeras, R. Guigo, E. Birney, Z. Weng, M. Snyder, and M. Gerstein. Understanding transcriptional regulation by integrative analysis of transcription data. *Genome Res, PMID: 22955978*, 22(9):1658–67, 2012.

Y. Cheng, W. Wu, S. Kumar, D. Yu, W. Deng, T. Tripic, D. King, K. Chen, Y. Zhang, D. Drautz, B. Giardine, S. Schuster, W. Miller, F. Chiaromonte, Y. Zhang, G. Blobel, M. Weiss, and R. Hardison. Erythroid gata1 function revealed by genome-wide analysis of transcription factor expression. *Genome Res, PMID: 19887574*, 19(12):2172–84, 2009.

J. Crispino. Gata1 in normal and malignant hematopoiesis. *Semin Cell Dev Biol, PMID: 15659348*, 16(1):137–47, 2005.

K. Cui, C. Zang, T. Roh, D. Schones, R. Childs, W. Peng, and K. Zhao. Chromatin signatures in multipotent human hematopoietic stem cells indicate the differentiation. *Cell Stem Cell, PMID: 19128795*, 4(1):80–93, 2009.

da Huang, B. Sherman, and R. Lempicki. Systematic and integrative analysis of large gene lists using david resources. *Nat Protoc, PMID: 19131956*, 4(1):44–57, 2009.

H. Deubzer, M. Schier, I. Oehme, M. Lodrini, B. Haendler, A. Sommer, and O. Witt. Hdac11 is a novel drug target in carcinomas. *Int J Cancer, PMID: 23024001*, 132(9):2200–8, 2013.

J. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. Antosiewicz-Bourget, A. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. Lobanenkov, J. Ecker, J. Thomson, and B. Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature, PMID: 25693564*, 518(7539):331–6, 2015.

X. Dong, M. Greven, A. Kundaje, S. Djebali, J. Brown, C. Cheng, T. Gingeras, M. Gerstein, R. Guigo, E. Birney, and Z. Weng. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol, PMID: 22950368*, 13(9):R53, 2012.

D. Donze, T. Townes, and J. Bieker. to beta-globin gene switching. *J Biol Chem, PMID: 7829533*, 270(4):1955–9, 1995.

E. Dzierzak and S. Philipsen. Erythropoiesis: development and differentiation. *Cold Spring Harb Perspect Med, PMID: 23545573*, 3(4):a011601, 2013.

D. Emery, G. Gavriilidis, H. Asano, and G. Stamatoyannopoulos. The transcription factor klf11 can induce gamma-globin gene expression in the erythropoiesis. *J Cell Biochem, PMID: 17131378*, 100(4):1045–55, 2007.

P. ENCODE. An integrated encyclopedia of dna elements in the human genome. *Nature, PMID: 22955616*, 489(7414):57–74, 2012.

J. Ernst and M. Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nat Methods, PMID: 22373907*, 9(3):215–6, 2012.

R. Ferreira, K. Ohneda, M. Yamamoto, and S. Philipsen. Gata1 function, a paradigm for transcription factors in hematopoiesis. *Mol Cell Biol, PMID: 15684376*, 25(4):1215–27, 2005.

P. Frontelo, D. Manwani, M. Galdass, H. Karsunky, F. Lohmann, P. Gallagher, and J. Bieker. Novel role for eklf in megakaryocyte lineage commitment. *Blood, PMID: 17715392*, 110 (12):3871–80, 2007.

A. Fujieda, N. Katayama, K. Ohishi, K. Yamamura, T. Shibasaki, Y. Sugimoto, E. Miyata, K. Nishi, M. Masuya, H. Ueda, H. Nakajima, and H. Shiku. A putative role for histone deacetylase in the differentiation of human erythroid cells. *Int J Oncol, PMID: 16077924*, 27(3):743–8, 2005.

T. Fujiwara, H. O'Geen, S. Keles, K. Blahnik, A. Linnemann, Y. Kang, K. Choi, P. Farnham, and E. Bresnick. Discovering hematopoietic mechanisms through genome-wide analysis of gata factor occupancy. *Mol Cell, PMID: 19941826*, 36(4):667–81, 2009.

J. Galloway, R. Wingert, C. Thisse, B. Thisse, and L. Zon. Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. *Dev Cell, PMID: 15621534*, 8(1):109–16, 2005.

A. Galvez, L. Huang, M. Magbanua, K. Dawson, and R. Rodriguez. Differential expression of thrombospondin (thbs1) in tumorigenic and soy peptide. *Nutr Cancer, PMID: 21526452*, 63(4):623–36, 2011.

L. Gay and B. Felding-Habermann. Contribution of platelets to tumour metastasis. *Nat Rev Cancer, PMID: 21258396*, 11(2):123–34, 2011.

R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol, PMID: 15461798*, 5(10):R80, 2004.

N. Goardon, J. Lambert, P. Rodriguez, P. Nissaire, S. Herblot, P. Thibault, D. Dumenil, J. Strouboulis, P. Romeo, and T. Hoang. Eto2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J, PMID: 16407974*, 25(2):357–66, 2006.

T. Graf. Differentiation plasticity of hematopoietic cells. *Blood, PMID: 11964270*, 99(9): 3089–101, 2002.

J. Guberman, J. Ai, O. Arnaiz, J. Baran, A. Blake, R. Baldock, C. Chelala, D. Croft, A. Cros, R. Cutts, G. Di, S. Forbes, T. Fujisawa, E. Gadaleta, D. Goodstein, G. Gundem, B. Haggarty, S. Haider, M. Hall, T. Harris, R. Haw, S. Hu, S. Hubbard, J. Hsu, V. Iyer, P. Jones, T. Katayama, R. Kinsella, L. Kong, D. Lawson, Y. Liang, N. Lopez-Bigas, J. Luo, M. Lush, J. Mason, F. Moreews, N. Ndegwa, D. Oakley, C. Perez-Llamas, M. Primig, E. Rivkin, S. Rosanoff, R. Shepherd, R. Simon, B. Skarnes, D. Smedley, L. Sperling, W. Spooner, P. Stevenson, K. Stone, J. Teague, J. Wang, J. Wang, B. Whitty, D. Wong, M. Wong-Erasmus, L. Yao, K. Youens-Clark, C. Yung, J. Zhang, and A. Kasprzyk. Biomart central portal: an open database network for the biological community. *Database (Oxford), PMID: 21930507*, 2011:bar041, 2011.

L. Gutierrez, F. Lindeboom, R. Ferreira, R. Drissen, F. Grosveld, D. Whyatt, and S. Philipsen. A hanging drop culture method to study terminal erythroid differentiation. *Exp Hematol, PMID: 16219530*, 33(10):1083–91, 2005.

I. Hamlett, J. Draper, J. Strouboulis, F. Iborra, C. Porcher, and P. Vyas. Characterization of megakaryocyte gata1-interacting proteins: the corepressor maturation. *Blood, PMID: 18625887*, 112(7):2738–49, 2008.

A. Hart, F. Melet, P. Grossfeld, K. Chien, C. Jones, A. Tunnacliffe, R. Favier, and A. Bernstein. Fli-1 is required for murine vascular and megakaryocytic development and is thrombocytopenia. *Immunity, PMID: 10981960*, 13(2):167–77, 2000.

S. Hattangadi, P. Wong, L. Zhang, J. Flygare, and H. Lodish. From stem cell to red cell: regulation of erythropoiesis at multiple levels by modifications. *Blood, PMID: 21998215*, 118(24):6258–68, 2011.

C. Heyworth, S. Pearson, G. May, and T. Enver. Transcription factor-mediated lineage switching reveals plasticity in primary cells. *EMBO J, PMID: 12110589*, 21(14):3770–81, 2002.

C. Huang, Y. Xiang, Y. Wang, X. Li, L. Xu, Z. Zhu, T. Zhang, Q. Zhu, K. Zhang, N. Jing, and C. Chen. Dual-specificity histone demethylase kiaa1718 (kdm7a) regulates neural fgf4. *Cell Res, PMID: 20084082*, 20(2):154–65, 2010.

H. Huang, K. Kathrein, A. Barton, Z. Gitlin, Y. Huang, T. Ward, O. Hofmann, A. Dibiase, A. Song, S. Tyekucheva, W. Hide, . ORCID:, Y. Zhou, and L. Zon. A network of epigenetic regulators guides developmental haematopoiesis in vivo. *Nat Cell Biol, PMID: 24240475*, 15(12):1516–25, 2013.

K. Ikuta and I. Weissman. Evidence that hematopoietic stem cells express mouse c-kit but do not depend on generation. *Proc Natl Acad Sci U S A, PMID: 1371359*, 89(4):1502–6, 1992.

H. Iwasaki, S. Mizuno, R. Wells, A. Cantor, S. Watanabe, and K. Akashi. Gata-1 converts lymphoid and myelomonocytic progenitors into the lineages. *Immunity, PMID: 14499119*, 19(3):451–62, 2003.

S. Jayapal, K. Lee, P. Ji, P. Kaldis, B. Lim, and H. Lodish. Down-regulation of myc is essential for terminal erythroid maturation. *J Biol Chem, PMID: 20940306*, 285(51): 40252–65, 2010.

P. Ji, V. Yeh, T. Ramirez, M. Murata-Hori, and H. Lodish. Histone deacetylase 2 is required for chromatin condensation and subsequent erythroblasts. *Haematologica, PMID: 20823130*, 95(12):2013–21, 2010.

Q. Jin, L. Yu, L. Wang, Z. Zhang, L. Kasper, J. Lee, C. Wang, P. Brindle, S. Dent, and K. Ge. Distinct roles of gcn5/pcaf-mediated h3k9ac and cbp/p300-mediated h3k18/27ac in transactivation. *EMBO J, PMID: 21131905*, 30(2):249–62, 2011.

R. Karlic, H. Chung, J. Lasserre, K. Vlahovicek, and M. Vingron. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A, PMID: 20133639*, 107(7):2926–31, 2010.

M. Kassouf, J. Hughes, S. Taylor, S. McGowan, S. Soneji, A. Green, P. Vyas, and C. Porcher. Genome-wide identification of tal1's functional targets: insights into its cells. *Genome Res, PMID: 20566737*, 20(8):1064–83, 2010.

M. Kerenyi and S. Orkin. Networking erythropoiesis. *J Exp Med, PMID: 21098097*, 207 (12):2537–41, 2010.

S. Kim and E. Bresnick. Transcriptional control of erythropoiesis: emerging mechanisms and principles. *Oncogene, PMID: 17934485*, 26(47):6777–94, 2007.

S. Kim, S. Bultman, H. Jing, G. Blobel, and E. Bresnick. Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. *Mol Cell Biol, PMID: 17438135*, 27(12):4551–65, 2007.

T. Kina, K. Ikuta, E. Takayama, K. Wada, A. Majumdar, I. Weissman, and Y. Katsura. The monoclonal antibody ter-119 recognizes a molecule associated with glycophorin lineage. *Br J Haematol, PMID: 10848813*, 109(2):280–7, 2000.

M. Kondo, I. Weissman, and K. Akashi. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell, PMID: 9393859*, 91(5):661–72, 1997.

M. Kowalczyk, J. Hughes, D. Garrick, M. Lynch, J. Sharpe, J. Sloane-Stanley, S. McGowan, G. De, M. Hosseini, D. Vernimmen, J. Brown, N. Gray, L. Collavin, R. Gibbons, J. Flint, S. Taylor, V. Buckle, T. Milne, W. Wood, and D. Higgs. Intragenic enhancers act as alternative promoters. *Mol Cell, PMID: 22264824*, 45(4):447–58, 2012.

A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. Ziller, V. Amin, J. Whitaker, M. Schultz, L. Ward, A. Sarkar, G. Quon, R. Sandstrom, M. Eaton, Y. Wu, A. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. Harris, N. Shoresh, C. Epstein, E. Gjoneska, D. Leung, W. Xie, R. Hawkins, R. Lister, C. Hong, P. Gascard, A. Mungall, R. Moore, E. Chuah, A. Tam, T. Canfield, R. Hansen, R. Kaul, P. Sabo, M. Bansal, A. Carles, J. Dixon, K. Farh, S. Feizi, R. Karlic, A. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. Mercer, S. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. Sallari, K. Siebenthall, N. Sinnott-Armstrong, M. Stevens, R. Thurman, J. Wu, B. Zhang, X. Zhou, A. Beaudet, L. Boyer, J. De, P. Farnham, S. Fisher,

D. Haussler, S. Jones, W. Li, M. Marra, M. McManus, S. Sunyaev, J. Thomson, T. Tlsty, L. Tsai, W. Wang, R. Waterland, M. Zhang, L. Chadwick, B. Bernstein, J. Costello, J. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature, PMID: 25693563*, 518(7539):317–30, 2015.

B. Langmead and S. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods, PMID: 22388286*, 9(4):357–9, 2012.

R. Leinonen, H. Sugawara, and M. Shumway. The sequence read archive. *Nucleic Acids Res, PMID: 21062823*, 39(Database issue):D19–21, 2011.

D. Letting, C. Rakowski, M. Weiss, and G. Blobel. Formation of a tissue-specific histone acetylation pattern by the hematopoietic gata-1. *Mol Cell Biol, PMID: 12556492*, 23(4): 1334–40, 2003.

C. Li and G. Johnson. Murine hematopoietic stem and progenitor cells: I. enrichment and biologic characterization. *Blood, PMID: 7534130*, 85(6):1472–9, 1995.

L. Li, J. Lee, J. Gross, S. Song, A. Dean, and P. Love. A requirement for lim domain binding protein 1 in erythropoiesis. *J Exp Med, PMID: 21041453*, 207(12):2543–50, 2010.

F. Liu, M. Walmsley, A. Rodaway, and R. Patient. Fli1 acts at the top of the transcriptional network driving blood and endothelial development. *Curr Biol, PMID: 18718762*, 18(16): 1234–40, 2008.

F. Lohmann and J. Bieker. Activation of eklf expression during hematopoiesis by gata2 and smad5 prior to commitment. *Development, PMID: 18448565*, 135(12):2071–82, 2008.

J. Lowry and J. Mackay. Gata-1: one protein, many partners. *Int J Biochem Cell Biol, PMID: 16095949*, 38(1):6–11, 2006.

K. MacQuarrie, A. Fong, R. Morse, and S. Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet, PMID: 21295369*, 27(4):141–8, 2011.

M. Martowicz, J. Grass, M. Boyer, H. Guend, and E. Bresnick. Dynamic gata factor interplay at a multicomponent regulatory region of the gata-2 locus. *J Biol Chem, PMID: 15494394*, 280(3):1724–32, 2005.

M. Maslova and T. Tavrovskaia. [the seasonal dynamics of erythropoiesis in the frog rana temporaria]. *Zh Evol Biokhim Fiziol, PMID: 8317184*, 29(2):211–4, 1993.

H. Mikkola, J. Klintman, H. Yang, H. Hock, T. Schlaeger, Y. Fujiwara, and S. Orkin. Haematopoietic stem cells retain long-term repopulating activity and multipotency gene. *Nature, PMID: 12540851*, 421(6922):547–51, 2003.

I. Miller and J. Bieker. A novel, erythroid cell-specific murine transcription factor that binds to the proteins. *Mol Cell Biol, PMID: 7682653*, 13(5):2776–86, 1993.

F. Morceau, M. Schnekenburger, M. Dicato, and M. Diederich. Gata-1: friends, brothers, and coworkers. *Ann N Y Acad Sci, PMID: 15659837*, 1030:537–54, 2004.

M. Mukhopadhyay, A. Teufel, T. Yamashita, A. Agulnick, L. Chen, K. Downs, A. Schindler, A. Grinberg, S. Huang, D. Dorward, and H. Westphal. Functional ablation of the mouse ldb1 gene results in severe patterning defects gastrulation. *Development, PMID: 12490556*, 130(3):495–505, 2003.

H. Nakajima. Role of transcription factors in differentiation and reprogramming of cells. *Keio J Med, PMID: 21720200*, 60(2):47–55, 2011.

S. Nejigane, S. Takahashi, Y. Haramoto, T. Michiue, and M. Asashima. Hippo signaling components, mst1 and mst2, act as a switch between self-renewal progenitors. *Int J Dev Biol, PMID: 23873372*, 57(5):407–14, 2013.

B. Nuez, D. Michalovich, A. Bygrave, R. Ploemacher, and F. Grosveld. Defective haematopoiesis in fetal liver resulting from inactivation of the eklf gene. *Nature, PMID: 7753194*, 375(6529):316–8, 1995.

A. Nurden. Platelets, inflammation and tissue regeneration. *Thromb Haemost, PMID: 21479340*, 105 Suppl 1:S13–33, 2011.

P. Nurden, N. Debili, W. Vainchenker, R. Bobe, R. Bredoux, E. Corvazier, R. Combrie, E. Fressinaud, D. Meyer, A. Nurden, and J. Enouf. Impaired megakaryocytopoiesis in type 2b von willebrand disease with severe thrombocytopenia. *Blood, PMID: 16720832*, 108(8):2587–95, 2006.

M. Ogawa. Differentiation and proliferation of hematopoietic stem cells. *Blood, PMID: 8499622*, 81(11):2844–53, 1993.

K. Orford, P. Kharchenko, W. Lai, M. Dao, D. Worhunsky, A. Ferro, V. Janzen, P. Park, and D. Scadden. Differential h3k4 methylation identifies developmentally poised hematopoietic genes. *Dev Cell, PMID: 18477461*, 14(5):798–809, 2008.

S. Orkin. Diversification of haematopoietic stem cells to specific lineages. *Nat Rev Genet, PMID: 11262875*, 1(1):57–64, 2000.

S. Orkin. Priming the hematopoietic pump. *Immunity, PMID: 14614848*, 19(5):633–4, 2003.

S. Orkin and L. Zon. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell, PMID: 18295580*, 132(4):631–44, 2008.

L. Pang, H. Xue, G. Szalai, X. Wang, Y. Wang, D. Watson, W. Leonard, G. Blobel, and M. Poncz. Maturation stage-specific regulation of megakaryopoiesis by pointed-domain ets proteins. *Blood, PMID: 16757682*, 108(7):2198–206, 2006.

G. Papadopoulos, E. Karkoulia, I. Tsamardinos, C. Porcher, J. Ragoussis, J. Bungert, and J. Strouboulis. Gata-1 genome-wide occupancy associates with distinct epigenetic profiles in erythropoiesis. *Nucleic Acids Res, PMID: 23519611*, 41(9):4938–48, 2013.

P. Park. Chip-seq: advantages and challenges of a maturing technology. *Nat Rev Genet, PMID: 19736561*, 10(10):669–80, 2009.

A. Pekowska, T. Benoukraf, P. Ferrier, and S. Spicuglia. A unique h3k4me2 profile marks tissue-specific gene regulation. *Genome Res, PMID: 20841431*, 20(11):1493–502, 2010.

A. Perkins, A. Sharpe, and S. Orkin. Lethal beta-thalassaemia in mice lacking the erythroid caccc-transcription factor eklf. *Nature, PMID: 7753195*, 375(6529):318–22, 1995.

A. Pilon, S. Ajay, S. Kumar, L. Steiner, P. Cherukuri, S. Wincovitch, S. Anderson, J. Mullikin, P. Gallagher, R. Hardison, E. Margulies, and D. Bodine. Genome-wide chip-seq reveals a dramatic shift in the binding of the transcription differentiation. *Blood, PMID: 21900194*, 118(17):e139–48, 2011.

J. Pimanda, K. Ottersbach, K. Knezevic, S. Kinston, W. Chan, N. Wilson, J. Landry, A. Wood, A. Kolb-Kokocinski, A. Green, D. Tannahill, G. Lacaud, V. Kouskoff, and B. Gottgens. Gata2, fli1, and scl form a recursively wired gene-regulatory circuit during development. *Proc Natl Acad Sci U S A, PMID: 17962413*, 104(45):17692–7, 2007.

C. Pina, G. May, S. Soneji, D. Hong, and T. Enver. Mllt3 regulates early human erythroid and megakaryocytic cell fate. *Cell Stem Cell, PMID: 18371451*, 2(3):264–73, 2008.

A. Pohl and M. Beato. bwtool: a tool for bigwig files. *Bioinformatics, PMID: 24489365*, 30 (11):1618–9, 2014.

R. Pop, J. Shearstone, Q. Shen, Y. Liu, K. Hallstrom, M. Koulnis, J. Gribnau, and M. Socolovsky. A key commitment step in erythropoiesis is synchronized with the cell cycle clock progression. *PLoS Biol, PMID: 20877475*, 8(9), 2010.

C. Porcher, W. Swat, K. Rockwell, Y. Fujiwara, F. Alt, and S. Orkin. The t cell leukemia oncoprotein scl/tal-1 is essential for development of all lineages. *Cell, PMID: 8689686*, 86(1):47–57, 1996.

C. Porcher, E. Liao, Y. Fujiwara, L. Zon, and S. Orkin. Specification of hematopoietic and vascular development by the bhlh transcription binding. *Development, PMID: 10498694*, 126(20):4603–15, 1999.

E. Querfurth, M. Schuster, H. Kulessa, J. Crispino, G. Doderlein, S. Orkin, T. Graf, and C. Nerlov. Antagonism between c/ebpbeta and fog in eosinophil lineage commitment of progenitors. *Genes Dev, PMID: 11018018*, 14(19):2515–25, 2000.

N. Rajagopal, W. Xie, Y. Li, U. Wagner, W. Wang, J. Stamatoyannopoulos, J. Ernst, M. Kellis, and B. Ren. Rfecs: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol, PMID: 23526891*, 9(3):e1002968, 2013.

T. Reya, S. Morrison, M. Clarke, and I. Weissman. Stem cells, cancer, and cancer stem cells. *Nature, PMID: 11689955*, 414(6859):105–11, 2001.

J. Rhodes, A. Hagen, K. Hsu, M. Deng, T. Liu, A. Look, and J. Kanki. Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish. *Dev Cell, PMID: 15621533*, 8(1):97–108, 2005.

L. Robb, I. Lyons, R. Li, L. Hartley, F. Kontgen, R. Harvey, D. Metcalf, and C. Begley. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc Natl Acad Sci U S A, PMID: 7624372*, 92(15):7075–9, 1995.

L. Robb, N. Elwood, A. Elefanty, F. Kontgen, R. Li, L. Barnett, and C. Begley. The scl gene product is required for the generation of all hematopoietic lineages mouse. *EMBO J, PMID: 8861941*, 15(16):4123–9, 1996.

M. Robinson, D. McCarthy, and G. Smyth. edger: a bioconductor package for differential expression analysis of digital data. *Bioinformatics, PMID: 19910308*, 26(1):139–40, 2010.

P. Rodriguez, E. Bonte, J. Krijgsveld, K. Kolodziej, B. Guyot, A. Heck, P. Vyas, B. de, F. Grosveld, and J. Strouboulis. Gata-1 forms distinct activating and repressive complexes in erythroid cells. *EMBO J, PMID: 15920471*, 24(13):2354–66, 2005.

M. Rylski, J. Welch, Y. Chen, D. Letting, J. Diehl, L. Chodosh, G. Blobel, and M. Weiss. Gata-1-mediated proliferation arrest during erythroid maturation. *Mol Cell Biol, PMID: 12832487*, 23(14):5031–42, 2003.

E. Sahakian, J. Powers, J. Chen, S. Deng, F. Cheng, A. Distler, D. Woods, J. Rock-Klotz, A. Sodre, J. Youn, K. Woan, A. Villagra, D. Gabrilovich, E. Sotomayor, and J. Pinilla-Ibarz. Histone deacetylase 11: A novel epigenetic regulator of myeloid derived function. *Mol Immunol, PMID: 25155994*, 63(2):579–85, 2015.

M. Sanchez-Castillo, D. Ruau, A. Wilkinson, F. Ng, http://orcid.org/0000 0001-6335-711X ORCID:, R. Hannah, E. Diamanti, P. Lombard, N. Wilson, and B. Gottgens. Codex: a next-generation sequencing experiment database for the haematopoietic communities. *Nucleic Acids Res, PMID: 25270877*, 43(Database issue):D1117–23, 2015.

A. Schuh, A. Tipping, A. Clark, I. Hamlett, B. Guyot, F. Iborra, P. Rodriguez, J. Strouboulis, T. Enver, P. Vyas, and C. Porcher. Eto-2 associates with scl in erythroid cells and megakaryocytes and provides erythropoiesis. *Mol Cell Biol, PMID: 16287841*, 25(23): 10235–50, 2005.

J. Semple, J. Italiano, and J. Freedman. Platelets and the immune continuum. *Nat Rev Immunol, PMID: 21436837*, 11(4):264–74, 2011.

J. Shearstone, R. Pop, C. Bock, P. Boyle, A. Meissner, and M. Socolovsky. Global dna demethylation during mouse erythropoiesis in vivo. *Science, PMID: 22076376*, 334(6057): 799–802, 2011.

Y. Shi, C. Matson, F. Lan, S. Iwase, T. Baba, and Y. Shi. Regulation of lsd1 histone demethylase activity by its associated factors. *Mol Cell, PMID: 16140033*, 19(6):857–64, 2005.

R. Shivdasani, E. Mayer, and S. Orkin. Absence of blood formation in mice lacking the t-cell leukaemia oncoprotein tal-1/scl. *Nature, PMID: 7830794*, 373(6513):432–4, 1995.

V. Skov, T. Larsen, M. Thomassen, C. Riley, M. Jensen, O. Bjerrum, T. Kruse, and H. Hasselbalch. Increased gene expression of histone deacetylases in patients with neoplasms. *Leuk Lymphoma, PMID: 21806350*, 53(1):123–9, 2012.

M. Socolovsky, H. Nam, M. Fleming, V. Haase, C. Brugnara, and H. Lodish. Ineffective erythropoiesis in stat5a(-/-)5b(-/-) mice due to decreased survival erythroblasts. *Blood, PMID: 11719363*, 98(12):3261–73, 2001.

E. Soler, C. Andrieu-Soler, B. de, J. Bryne, S. Thongjuea, R. Stadhouders, R. Palstra, M. Stevens, C. Kockx, I. van, J. Hou, C. Steinhoff, E. Rijkers, B. Lenhard, and F. Grosveld. The genome-wide dynamics of the binding of ldb1 complexes during erythroid differentiation. *Genes Dev, PMID: 20123907*, 24(3):277–89, 2010.

S. Song, C. Hou, and A. Dean. A positive role for nli/ldb1 in long-range beta-globin locus control region function. *Mol Cell, PMID: 18082606*, 28(5):810–22, 2007.

G. Spangrude, S. Heimfeld, and I. Weissman. Purification and characterization of mouse hematopoietic stem cells. *Science, PMID: 2898810*, 241(4861):58–62, 1988.

D. Spyropoulos, P. Pharr, K. Lavenburg, P. Jackers, T. Papas, M. Ogawa, and D. Watson. Hemorrhage, impaired hematopoiesis, and lethality in mouse embryos carrying a factor. *Mol Cell Biol, PMID: 10891501*, 20(15):5643–52, 2000.

J. Starck, N. Cohet, C. Gonnet, S. Sarrazin, Z. Doubeikovskaia, A. Doubeikovski, A. Verger, M. Duterque-Coquillaud, and F. Morle. Functional cross-antagonism between transcription factors fli-1 and eklf. *Mol Cell Biol, PMID: 12556498*, 23(4):1390–402, 2003.

D. Steger, M. Lefterova, L. Ying, A. Stonestrom, M. Schupp, D. Zhuo, A. Vakoc, J. Kim, J. Chen, M. Lazar, G. Blobel, and C. Vakoc. Dot1l/kmt4 recruitment and h3k79 methylation are ubiquitously coupled with gene cells. *Mol Cell Biol, PMID: 18285465*, 28(8):2825–39, 2008.

T. Suganuma and J. Workman. Signals and combinatorial functions of histone modifications. *Annu Rev Biochem, PMID: 21529160*, 80:473–99, 2011.

P. Talbert and S. Henikoff. Histone variants–ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol, PMID: 20197778*, 11(4):264–75, 2010.

M. Tallack, T. Whitington, W. Yuen, E. Wainwright, J. Keys, B. Gardiner, E. Nourbakhsh, N. Cloonan, S. Grimmond, T. Bailey, and A. Perkins. A global role for klf1 in erythropoiesis revealed by chip-seq in primary cells. *Genome Res, PMID: 20508144*, 20(8): 1052–63, 2010.

C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, B. van, S. Salzberg, B. Wold, and L. Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts differentiation. *Nat Biotechnol, PMID: 20436464*, 28(5):511–5, 2010.

T. Tripic, W. Deng, Y. Cheng, Y. Zhang, C. Vakoc, G. Gregory, R. Hardison, and G. Blobel. Scl and associated proteins distinguish active from repressive gata transcription complexes. *Blood, PMID: 19011221*, 113(10):2191–201, 2009.

A. Tsiftsoglou, I. Vizirianakis, and J. Strouboulis. Erythropoiesis: model systems, molecular regulators, and developmental programs. *IUBMB Life, PMID: 19621348*, 61(8):800–30, 2009.

A. Valouev, D. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods, PMID: 19160518*, 5(9):829–34, 2008.

V. Valverde-Garduno, B. Guyot, E. Anguita, I. Hamlett, C. Porcher, and P. Vyas. Differences in the chromatin structure and cis-element organization of the human identification. *Blood, PMID: 15265794*, 104(10):3106–16, 2004.

A. Vannucchi, L. Bianchi, C. Cellai, F. Paoletti, V. Carrai, A. Calzolari, L. Centurione, R. Lorenzini, C. Carta, E. Alfani, M. Sanchez, G. Migliaccio, and A. Migliaccio. Accentuated response to phenylhydrazine and erythropoietin in mice genetically mice). *Blood, PMID: 11342429*, 97(10):3040–50, 2001.

C. Waddington. The strategy of the genes. 1957.

K. Watamoto, M. Towatari, Y. Ozawa, Y. Miyata, M. Okamoto, A. Abe, T. Naoe, and H. Saito. Altered interaction of hdac5 with gata-1 during mel cell differentiation. *Oncogene, PMID: 14668799*, 22(57):9176–84, 2003.

I. Weissman, D. Anderson, and F. Gage. Stem and progenitor cells: origins, phenotypes, lineage commitments, and transdifferentiations. *Annu Rev Cell Dev Biol, PMID: 11687494*, 17:387–403, 2001.

J. Welch, J. Watts, C. Vakoc, Y. Yao, H. Wang, R. Hardison, G. Blobel, L. Chodosh, and M. Weiss. Global regulation of erythroid gene expression by transcription factor gata-1. *Blood, PMID: 15297311*, 104(10):3136–47, 2004.

D. Whyatt, F. Lindeboom, A. Karis, R. Ferreira, E. Milot, R. Hendriks, B. de, A. Langeveld, J. Gribnau, F. Grosveld, and S. Philipsen. An intrinsic but cell-nonautonomous defect in gata-1-overexpressing mouse cells. *Nature, PMID: 10952313*, 406(6795):519–24, 2000.

A. Wickrema and J. Crispino. Erythroid and megakaryocytic transformation. *Oncogene, PMID: 17934487*, 26(47):6803–15, 2007.

N. Wilson, S. Foster, X. Wang, K. Knezevic, J. Schutte, P. Kaimakis, P. Chilarska, S. Kinston, W. Ouwehand, E. Dzierzak, J. Pimanda, B. de, and B. Gottgens. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide regulators. *Cell Stem Cell, PMID: 20887958*, 7(4):532–44, 2010.

P. Wong, S. Hattangadi, A. Cheng, G. Frampton, R. Young, and H. Lodish. Gene induction and repression during terminal erythropoiesis are mediated by changes. *Blood, PMID: 21860024*, 118(16):e128–38, 2011.

S. Wontakal, X. Guo, C. Smith, T. MacCarthy, E. Bresnick, A. Bergman, M. Snyder, S. Weissman, D. Zheng, and A. Skoultchi. A core erythroid transcriptional network is repressed by a master regulator of differentiation. *Proc Natl Acad Sci U S A, PMID: 22357756*, 109 (10):3832–7, 2012.

W. Wu, Y. Cheng, C. Keller, J. Ernst, S. Kumar, T. Mishra, C. Morrissey, C. Dorman, K. Chen, D. Drautz, B. Giardine, Y. Shibata, L. Song, M. Pimkin, G. Crawford, T. Furey, M. Kellis, W. Miller, J. Taylor, S. Schuster, Y. Zhang, F. Chiaromonte, G. Blobel, M. Weiss, and R. Hardison. Dynamics of the epigenetic landscape during erythroid differentiation after gata1 restoration. *Genome Res, PMID: 21795386*, 21(10):1659–71, 2011.

J. Xu, Z. Shao, K. Glass, D. Bauer, L. Pinello, H. Van, S. Hou, J. Stamatoyannopoulos, H. Mikkola, G. Yuan, and S. Orkin. Combinatorial assembly of developmental stage-specific enhancers controls gene erythropoiesis. *Dev Cell, PMID: 23041383*, 23(4): 796–811, 2012.

L. Yang, J. Wan, Y. Ge, Z. Fu, S. Kim, Y. Fujiwara, J. Taub, L. Matherly, J. Eliason, and L. Li. The gata site-dependent hemogen promoter is transcriptionally regulated by gata1 cells. *Leukemia, PMID: 16437149*, 20(3):417–25, 2006.

O. Yousuf and D. Bhatt. The evolution of antiplatelet therapy in cardiovascular disease. *Nat Rev Cardiol, PMID: 21750497*, 8(10):547–59, 2011.

C. Yu, A. Cantor, H. Yang, C. Browne, R. Wells, Y. Fujiwara, and S. Orkin. Targeted deletion of a high-affinity gata-binding site in the gata-1 promoter vivo. *J Exp Med, PMID: 12045237*, 195(11):1387–95, 2002.

M. Yu, L. Riva, H. Xie, Y. Schindler, T. Moran, Y. Cheng, D. Yu, R. Hardison, M. Weiss, S. Orkin, B. Bernstein, E. Fraenkel, and A. Cantor. Insights into gata-1-mediated gene activation versus repression via genome-wide analysis. *Mol Cell, PMID: 19941827*, 36(4): 682–95, 2009.

W. Zhang and J. Bieker. Acetylation and modulation of erythroid kruppel-like factor (eklf) activity by acetyltransferases. *Proc Natl Acad Sci U S A, PMID: 9707565*, 95(17):9855–60, 1998.

Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. Liu. Model-based analysis of chip-seq (macs). *Genome Biol, PMID: 18798982*, 9(9):R137, 2008.

R. Zini, R. Norfo, F. Ferrari, E. Bianchi, S. Salati, V. Pennucci, G. Sacchi, C. Carboni, G. Ceccherelli, E. Tagliafico, S. Ferrari, and R. Manfredini. Valproic acid triggers erythro/megakaryocyte lineage decision through induction expression. *Exp Hematol, PMID: 22885124*, 40(12):1043–1054.e6, 2012.

List of figures