ANALYSIS OF THE HAEMATOPOIETIC TRANSCRIPTOME IN DEVELOPMENT

MYRTO ARETI KOSTADIMA

This thesis is submitted for the degree of Doctor of Philosophy

Jesus College
University of Cambridge

17th October 2014

DECLARATION

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 60000 words as defined by the Biology Degree Committee.

This dissertation has been typeset using LaTeX 2ε in 12 pt Palatino, one and half spaced, according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Cambridge, 17th October 2014

Myrto Areti Kostadima

To those whom I love, who love me back,
and those who challenge me

Of all that is written, I love only
what a person has written with his own blood.

Friedrich Nietzsche

ACKNOWLEDGMENTS

When it comes to long journeys, there is usually a long list of people that have shifted your rota. In my PhD adventure, there have been people that have moved it forward and people that have tried to stall it. Either way, no matter how small or large their contribution was, it has left its mark.

I would like to thank Prof Willem Ouwehand for welcoming me in his group and keeping me sane throughout this journey, and Dr Paul Bertone for giving me the opportunity to do a PhD at the European Bioinformatics Institute, one of the most exciting hubs of bioinformatics.

Sharing my time between these groups, I have met, scientifically interacted and been challenged by numerous people, whose support has been instrumental these last years. Special thanks go out to two outstanding scientists, who have put considerable time and effort in this thesis and whose scientific input has shaped my independent thinking. Remco and Mattia, as I have said before, I cannot thank you enough for everything you have done for me. Remco, above all I am grateful for your friendship, which goes beyond the PhD period. Mattia, who would have thought a few years back, when I met you at the Blood centre library, that we would form such a scientific duo? You have been a great leader and set an example that will be hard to follow.

Special thanks also to other lab members: Heidi, Mara and Pär from EBI and Kate and Sara from the Blood centre, for both their scientific and psychological support.

Many thanks to Vicky Schneider for introducing me into the bioinformatics training world and supporting me when I was trying to juggle between two big commitments. My occupation with training opened a whole new world for me, a really exciting and highly entertaining one, allowing my confidence to grow stronger in this field and helping me get rid of the feeling of loneliness that a PhD can sometimes bring up. It has also resulted in great partnerships and plans to conquer the world, right Gabriella Rustici?

Thanks to everyone whom I have met during my stay in Cambridge and have made it a really pleasant one. So pleasant that even after 4.5 years in this town I still live here despite how much I always complain about its size. Some great friendships have been built throughout these last few years with such a diverse group of people. Angela, Anna, Benedetta, Christine, Evangelos, Filipe, Felix, Maria, Nenad, Nils, Sander, Steve, I am grateful for your friendship, your insights and mainly for giving me the opportunity to fulfil one of my biggest dreams ever, to be part of a highly multicultural and stimulating environment.

It was within this multinational group of people, where I met my partner in crime. Michele, your support has been crucial for the completion of this work. You have managed throughout these years to keep me focused and motivated, but above all happy. I am thankful for the innovative, simplistic, down-to-earth and inspiring approach to life and its problems that you have taught me. I hope I have done my best to bring as much happiness into your life as you have succeeded in bringing into mine. Σ′ αγαπώ!

Moira, Gabry and Jack, molte grazie anche a voi. Grazie per avermi accettato nella vostra famiglia e per tutti i bei momenti che abbiamo avuto finora. I migliori devono ancora venire!

Φυσικά δεν θα ήμουν αυτή που είμαι χωρίς τους ανθρώπους που έχουν σταθεί δίπλα μου σε τεράστιες χαρές, αλλά και σε μεγάλες δοκιμασίες. Ξεκινώ από τον άνθρωπο, που νιώθω σαν οικογένειά

μου πλέον. Δάντουλάκο μου, τα καταφέραμε!! Χωρίς εσένα αυτή η στιγμή θα ήταν μακρινή, σχεδόν απίθανη. Λένα μου, αυτό το πτυχίο μαζί με το δικό σου σηματοδοτεί, ευελπιστώ, το τέλος μιας μακράς και συχνά επώδυνης διαδρομής. Τα καλύτερα έρχονται! Ένα μεγάλο ευχαριστώ στην υπόλοιπη παρέα των 7 για τις αξέχαστες στιγμές που έχουμε περάσει και την απόλυτη κατανόηση κι αγάπη. Αθανασία, Αλεξάνδρα, Αποστόλη, Ζωή, Όλγα σας λατρεύω και σας ευγνομονώ για όλα όσα μου έχετε παρέχει όλα αυτά τα χρόνια. Ακόμη κι αν δεν έχουμε καταφέρει να βρούμε την πόλη και την πολυκατοικία που θα μένουμε όλοι μαζί, η συντροφιά σας με ακολουθεί παντού.

Το μεγαλύτερο ευχαριστώ, βεβαίως, πηγαίνει στην οικογένειά μου, που όλα αυτά τα χρόνια στωικά αναμένει κάθε χρόνο το τέλος των σπουδών μου. Μαμακούλα, Μπαμπακούλα.. Τελείωσα!!! Ναταλίτσα μου, αυτό το διδακτορικό ελπίζω να σημάνει τη λήξη των αγχωτικών μου επεισοδίων. Όσο περνάνε τα χρόνια ερχόμαστε πιο κοντά και με έχεις στηρίξει υπέρμετρα. Σε ευχαριστώ για όλα. Ακόμα και για εκείνες τις μέρες που δεν έπρεπε να κάνεις φασαρία μέσα στο σπίτι και τις πέρασες στην παιδική χαρά από το πρωί μέχρι αργά το βράδυ για να διαβάζει η αδερφή σου για τις εισαγωγικές εξετάσεις. ΣΑΣ ΕΥΧΑΡΙΣΤΩ ΚΑΙ ΣΑΣ ΕΙΜΑΙ ΕΥΓΝΩΜΩΝ!

Overall this was challenging, but I grew stronger.

Cambridge,
17th October 2014

ANALYSIS OF THE HAEMATOPOIETIC TRANSCRIPTOME IN DEVELOPMENT

Myrto Areti Kostadima

All the cells of an individual share the same genetic code. The regulation of gene expression is among the key processes that confers cell identity by ensuring that a specific subset of genes are active in a given cell type. Understanding gene regulation is vital in unravelling the impact of genetic variation on both normal development and disease. This thesis investigates gene expression and its regulation during the development of blood cells, called haematopoiesis, with a focus on platelets. Platelets are small enucleated cells that are important for vascular wound healing after injury and are produced by megakaryocytes, their bone-marrow residing precursors.

MEIS1 is a transcription factor known to be necessary for platelet production. In the first part of this thesis, I analysed gene expression and MEIS1 binding profiles in megakaryocytes. By integrating these data with public DNA-binding profiles of other key regulatory proteins, as well as by computational analysis, I confirmed the functional role of MEIS1 and showed that it frequently co-binds with other key regulatory proteins.

Blood cell production proceeds in a hierarchical fashion from a multipotent stem cell. The formation of mature blood cell types relies on the correct regulation of all intermediate stages. As part of the Blueprint consortium, I studied the transcriptional landscape of rare haematopoietic progenitors. Transcriptional analysis of these cells identified changes in the expression of genes and transcript isoforms between cell types and lineages with much greater depth and resolution than previous studies.

In the last part I integrated the above datasets with the aim to identify potential novel transcriptional regulators in megakaryocytes. This integrative analysis established nuclear factor I/B (NFIB) as a candidate. Using additional evidence from other resources, I identified NFIB as specific to the megakaryocytic lineage.

This thesis analyses whole-genome transcriptome and binding data in relation to critical stages of blood system development. I examined the specific mechanisms of MEIS1-mediated gene regulation in megakaryocytes, as well as applying a global analysis of the full transcriptional landscape of early haematopoietic progenitors. In addition to providing an invaluable resource for the scientific community, the current study provides an example of how such data can be exploited to gain new insights into gene regulatory processes.

PREFACE

"Blood cannot be converted to water" say the Greeks, emphasising the importance of blood ties. The significance of consanguinity, however, is not a social phenomenon of modern societies. Ancient Greeks, Egyptians, Romans and also native Americans were all highly respectful of kinship, forming social units based on these relationships. For example, all Achaeans from Epirus to Crete belonged to one nation, for they shared the same blood (Meletis and Konstantopoulos, 2010).

But why do we use the term "blood ties" to describe kinship? According to Rogers (Rogers, 2010),

> "Aristotle emphasised the importance of blood in heredity. He thought that blood supplied generative material for building all parts of the adult body, and he reasoned that blood was the basis for passing on this generative power to the next generation. In fact, he believed that the male's semen was purified blood and that a woman's menstrual blood was her equivalent of semen. These male and female contributions united in the womb to produce a baby."

Hence, Aristotle's (384 - 322BC) notion of blood as the basis of heredity is still prevalent in everyday expressions that describe ancestry, despite it being far away from the prevailing model of inheritance.

Throughout history, blood held a key role in nutrition and religion as well. Homer, Aristotle, Hippocrates and Galen considered blood as the ultimate food that served as nutrient to the body and its organs. Even nowadays, blood consumption is part of the diet of the Maasai tribe in Africa (Århem, 1989)

and the Hua people in Papua New Guinea. Despite being considered a cannibalistic act, it is culturally justified for the Hua people to consume human blood (Meigs, 1987). Religions, however, have appeared to differ in their beliefs about blood consumption. Judaism, for example, forbade men to consume any blood on the grounds that "Blood belongs to God alone, for life is in blood" (Soler, 1997). Christianity, however, accepts the ritual of "consuming Christ's blood" to achieve the fusion of Man with God. In this case, though, blessed red wine represents the blood of Christ. Blood sacrifices were also popular in various cultures and constituted a gift to the gods in return for fertility, health or good harvest (Dalton, 1974).

In the area of physiology, the importance of blood for the life of *Animalia* has triggered extensive research. Ancient Greeks were the first to identify blood as an indispensable factor for human and animal life (Meletis and Konstantopoulos, 2010). From that point on, blood circulation and the haematopoietic system were extensively studied (Khan et al., 2005) and the discoveries made led to two medical breakthroughs in the 20th century: blood transfusions and bone marrow transplantations.

The first studies on the blood system reported the existence of two types of blood vessels: the veins and the arteries. In 500 BC, Alcmaeon of Croton first distinguished between larger more interior blood vessels and smaller superficial ones. Although some have claimed that he was the first one to distinguish between veins and arteries, this is highly unlikely (Lloyd, 1991), and this is attributed to the Greek physician, Herophilos (335-280 BC). Claudius Galenus (129-c216), known as Galen of Pergamon, identifies that although both veins and arteries carry blood, they have separate distinct functions, based on the discovery that the venous blood is darker, whereas the arterial is brighter and thinner. His discoveries, however, led him to two incorrect theories (Rocca and Galen, 2003). First, he claimed

that the circulatory system consisted of two one-way separate systems, rather than a unified one, and second, he argued that venous blood originates in the liver, while arterial blood from the heart. The circle is completed with the regeneration of the blood in either (Grant, 2000).

Much later, Ibn-al-Nafis (1213-1286), an Egyptian physician and a great admirer of Galen's work, refuted the prevailing theories about a direct pathway in the heart between the left and right chambers. He instead provided a detailed description of the blood circulation to and from the lungs (West, 2008). Similar findings by Spanish Michael Servetus (1511-1553) were not met with enthusiasm in Europe and Serevetus was burned as a heretic. At the same time Andreas Versalius (1514-1564) also critised Galen's work and with his work paved the way for the first drawings of the vein valves published by Hieronymus Fabricius (1537-1619). William Harvey, a pupil of Fabricius, is credited with the discovery of circulation. He published *'Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus'* in 1628, in which he demonstrated that there had to be a direct connection between the venous and arterial systems throughout the body, and not just the lungs (Power, 1897). The capillary system that connects arteries and veins, though, was later discovered by Marcello Malpighi (1628–1694) in 1661.

A few years later, in 1666, the first attempts to transfuse blood were conducted by Richard Lower between dogs (Tubbs et al., 2008; Walton, 1974) and between dog and human (Lower, 2002), while the first attempt for human-to-human transfusion was undertaken by Dr James Blundell in 1818 (1791–1878) (Ellis, 2007; Howell, 1828). Blood transfusions, though, were not a common practice in Britain until 1921, when the British Red Cross organised the first voluntary blood service (Miller, 2010). During the second World War the huge numbers of injured

people and the constant need for blood transfusions led to advances in the way blood is stored and transfused.

In the meantime, Karl Landsteiner (1868–1943), the so-called father of blood transfusion, discovered the main blood groups by mingling serum and blood cells of six individuals (Landsteiner, 1900; Rous, 1947). He found that transfusion between individuals of the same blood group did not lead to destruction of the blood cells, providing a safe way for the blood transfusion between humans. In 1900, Landsteiner discovered three blood groups, which he called A, B and C. The latter one was later renamed to O. A year later, Landsteiner's colleagues, Alfred von Decastello and Andriano Sturti, identified another blood group which they called AB (von Decastello and Sturli, 1902). The work of Landsteiner and his colleagues led to the modern system of classification of blood groups.

In 1985, blood transfusions changed dramatically after a high number of people became infected with HIV. The free love movement of the 1960 had liberated people's sexuality. The consequences of this were first observed among gay people in early 1980; the syndrome was initially called GRID (Gay-Related ImmunoDeficiency). When haemophiliacs also began to develop GRID, it was apparent that the syndrome was not a "privilege" of homosexuals. In the last decades, several tests against infectious elements, such as Hepatitis and HIV, have been introduced in blood transfusions to ensure its safety (Berkman, 1988). According to the World Health Organisation (WHO, 2013), nearly 83 million blood donations were collected globally in 2011, however only sixty countries collect 100% of their blood supply from voluntary donors.

Following World War II, the threat of nuclear bombs was high and scientists began examining ways to save individuals exposed to radiation. Their observation was that irradiation

destroys the bone marrow; hence its replacement by a healthy bone marrow sample might be the solution. Unfortunately, for the individuals exposed to such high doses of irradiation bone marrow failure was only part of the symptoms, as they also suffered from multiple organ failure making their recovery impossible. However, this idea prompted further experiments for patients with bone marrow diseases. In 1956, Dr E. Donnall Thomas (1920-2012) demonstrated that normal blood production could be restored by transplanting bone marrow-derived cells from one man into his identical twin brother, who was suffering from advanced leukemia, after total body irradiation (Thomas et al., 1957). Bone marrow transplantations were not performed on a large scale, though, until the discovery of the human histocompatibility antigens by Jean Dausset (1916-2009) (Dausset, 1958; Dausset and Nenna, 1952). The development of histocompatibility test between patient and donors resulted in bone marrow transplantations between siblings in 1968 and between unrelated individuals in 1973. Still today bone-marrow transplantations are the only stem cell therapy routinely performed. Such successful medical inventions have drawn significant interest on the haematopoietic system, more commonly known as blood system, which has become one of the most studied and characterised systems. As such, it serves as a model system for biological investigation and clinical applications.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

INTRODUCTION

1

INTRODUCTION

1.1 INTRODUCTION TO HAEMATOPOIESIS

The word haematopoiesis comes from the ancient Greek words: αίμα, which means "blood", and ποιείν, which means "to make".

The number of blood cells being formed per day varies among individuals, but on average a healthy adult produces approximately 10^{11}-10^{12} new blood cells (Stites et al., 1997). This high demand of cells in the peripheral circulation requires an efficient and tightly regulated process of blood formation, called haematopoiesis.

There are several sites of haematopoiesis during development and depending on the mammalian developmental stage, haematopoiesis can be classified as either primitive or definitive. In the early stages of mammalian development, primitive haematopoiesis takes place in the yolk sac producing red blood cells (Ueno et al., 2006). Subsequently definitive haematopoiesis occurs first in the aorta gonad mesonephros (AGM) shortly after the first wave of primitive haematopoiesis and later in the foetal liver and then in the bone marrow. Additional sites of haematopoiesis are the thymus and the spleen (reviewed in (Dzierzak and Speck, 2008)).

The final products of haematopoiesis can be split in three main groups: red blood cells or erythrocytes, platelets and thrombocytes, and white blood cells or leukocytes. The daily blood cell production is divided in $2x10^{11}$ red cells, $1x10^{10}$ white cells and $4x10^{11}$ platelets (ENCYCLOPAEDIA BRITANNICA).

1.1.1 *The haematopoietic system*

All different blood cell types are derived from a multipotent undifferentiated cell known as the haematopoietic stem cell (HSC). Extensive analysis using fluorescence activating cell sorting and monoclonal antibodies against cluster differentiation antigens (CD) suggest that haematopoiesis occurs in a hierarchical fashion (reviewed in (Orkin and Zon, 2008; Seita and Weissman, 2010)). The production of any of these cells consists of several differentiation steps, where undifferentiated cells, which are called precursors or progenitors, progressively commit to lineages of increasingly restricted differentiating potential. The haematopoietic system can be represented as an inverted tree structure, with the haematopoietic stem cell sitting at the root of the tree and its leaf nodes representing the mature blood cells (see Figure 1). The sequence of differentiation steps required to produce a mature blood cell from a haematopoietic stem cell is called a lineage differentiation.

Haematopoietic stem cells sit at the top of the haematopoietic system and possess two essential characteristics for haematopoiesis. They can differentiate into progenitor cells that in turn commit to lineages of increasingly restricted potential to produce mature blood cell types. Their differentiation ability is, therefore, essential to replenish circulating blood cells, some of which are short-lived. Exclusive differentiation of the haematopoietic stem cells would deplete the stem cell pool. To avoid this, haematopoietic stem cells can also self-renew, maintaining the stem cell population intact. Haematopoietic stem cells are, therefore, required to maintain a balance between self-renewal and differentiation to meet the high demands for both stem cell and mature cells. This dual nature has been made possible by asymmetric cell divisions (Giebel et al., 2006).

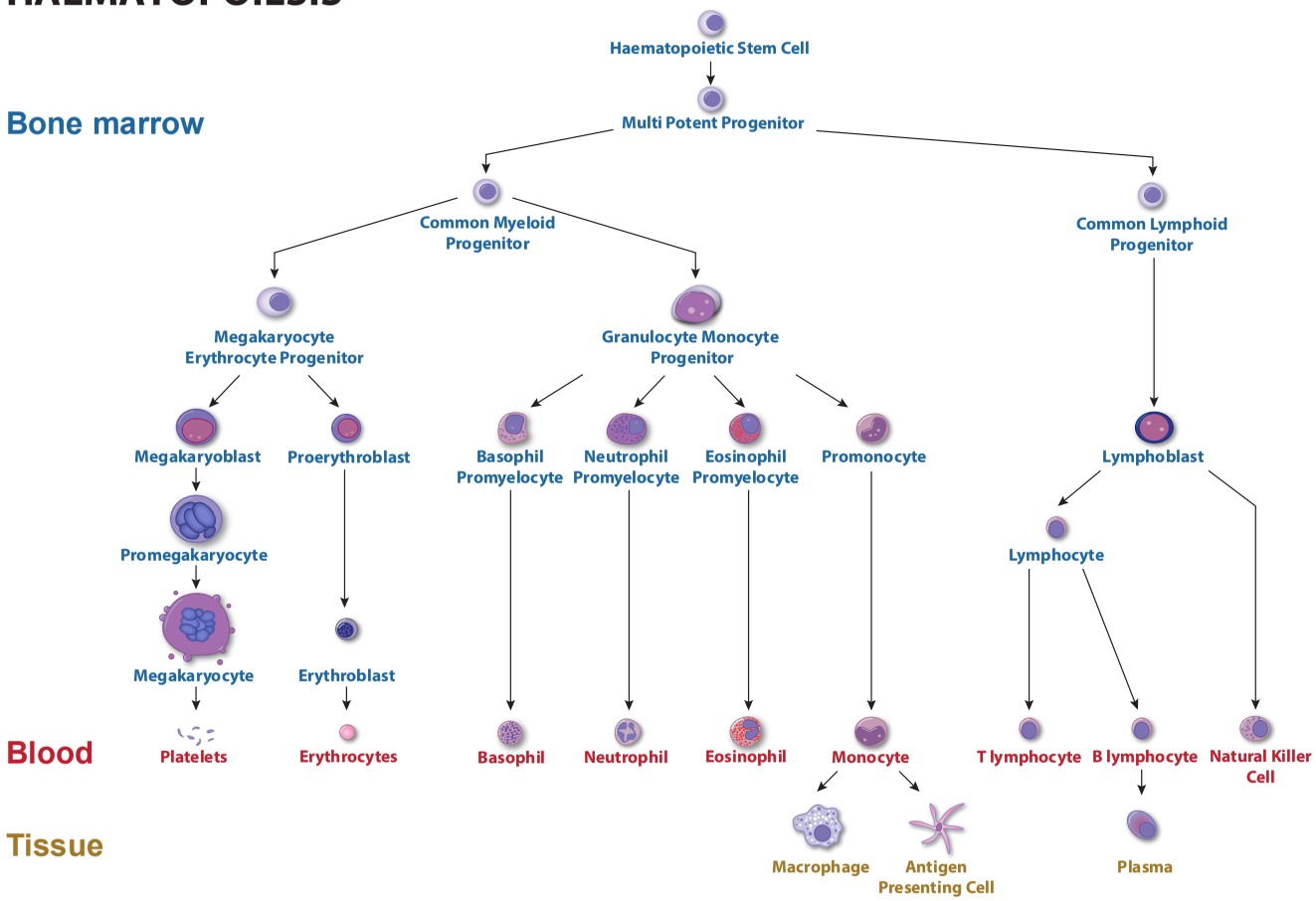


Figure 1: The classical model of haematopoiesis.

In normal circumstances, haematopoietic stem cells are mostly in a quiescent state (G_o phase of the cell cycle) (Bradford et al., 1997; Cheshier et al., 1999; Yoshihara et al., 2007). They only undergo cell division to self-renew or differentiate, a decision that is determined by the microenvironment in which they reside, called the niche. Schofield (1978) first introduced the term 'niche' and according to Wang and Wagers (2011) the niche: ".. includes specific cell types, anatomical locations, soluble molecules, signalling cascades and gradients, as well as physical factors, such as shear stress, oxygen tension and temperature".

Haematopoietic stem cells reside mainly in the bone marrow, the only site of haematopoiesis in adults. Wang et al. (1997) showed that they are rare, with only one cell in $3x10^6$ bone marrow cells having the regenerative properties of a haematopoietic stem cell. Haematopoietic stem cells can also be found in the mobilised peripheral blood (McCredie et al., 1971), where their frequency is one in 10^6 cells (Wang et al., 1997).

Once a haematopoietic stem cell commits to differentiation, it loses its self-renewal ability, while initially maintaining multipotency. The resulting multipotent cells are called multi potent progenitors (MPP). At this point the haematopoietic system has to make the first commitment choice between two main branches; the lymphoid and the myeloid. MPPs can progress into one of these branches by differentiating into one of the two oligopotent progenitors called common lymphoid progenitor (CLP) and common myeloid progenitor (CMP), respectively.

One of the three main blood products, leukocytes, are responsible for the defense of our bodies against disease or infection. The term 'leukocytes' describes a range of blood cells: granulocytes, monocytes and lymphocytes. The lymphoid branch, which stems from the CLPs, produces three different types of lymphocytes: T cells and B cells, which form the adaptive im-

mune system, and natural killer (NK) cells that are part of the innate immune system. The role of lymphocytes is to attack pathogens or tumour cells. In the case of T and B cells, this is achieved through antigen-specific receptors. The lack of such receptors on NK cells classifies them as components of the less specialised immune response of the innate immune system. However, the NK cells are considered to be lymphocytes because of their morphology, their expression of many lymphoid markers, and their origin from the common lymphoid progenitor cell in the bone marrow (Vivier et al., 2011).

Lymphocyte precursors reside in the bone marrow, from where they migrate when they are still immature to different peripheral organs. T cell precursors migrate to the thymus to undergo maturation and at the end of the maturation process they migrate to other lymphoid organs, such as the spleen (Starr et al., 2003). B cells leave the bone marrow to migrate to the peripheral lymphoid tissue in order to mature (Bonilla and Oettgen, 2010). NK cells can mature both in the bone marrow and the peripheral lymphoid organs before entering the blood circulation.

While B cells are responsible for the production of antibodies and NK cells for destroying damaged or cancerous cells, T cells show remarkable diversity of functions depending on their subtype. T cells are further split into:

1. T helper cells, whose role is to help B and T cells and macrophages (mature monocytes),
2. Regulatory T cells that modulate the immune responses in order to reduce any damage caused by them and maintain tolerance against our own antigens,
3. Cytotoxic T cells that kill virally infected or cancerous cells.

The role of communicator between the adaptive and immune system falls on the dendritic cells (reviewed in (Merad et al., 2013)). These cells can derive from both myeloid and lymphoid branches of the haematopoietic system (Manz et al., 2001a,b; Shigematsu et al., 2004; Traver et al., 2000), though the functional differences between these two types are not yet clear.

The other main branch of haematopoiesis, the myeloid, is further split into two sub-branches: the granulocyte/monocyte and the megakaryocyte/erythrocyte, which stem from the granulocyte/monocyte progenitors (GMP) and the megakaryocyte/erythrocyte progenitors (MEP), respectively. The GMP can give rise to granulocytes and monocytes, which are the remaining components of the immune system, while the MEP lineage produces the erythrocytes (red blood cells) and the thrombocytes (platelets), which are the two more abundant cells in the blood.

Granulocytes and monocytes form the innate immune system, which is the first line of defense against antigens, along with NK cells. In contrast to T and B cells, this subgroup of leukocytes has a limited function because they identify only the common features of pathogens. Depending on their role, granulocytes and monocytes can be grouped into the cell eaters, called phagocytes and include neutrophils, macrophages and dendritic cells, and the non-eaters that include basophils and eosinophils.

The term 'granulocytes' describes a subgroup of leukocytes produced by the myeloid branch that are characterised by the presence of granules in their cytoplasm. The white blood cells sharing this feature are:

1. Basophils, which participate in allergic reactions by producing histamines,

2. Eosinophils, cells that attack bacteria and parasites by secreting toxic proteins, and

3. Neutrophils, which are the only cell eaters among the granulocytes. Neutrophils reside in the bone marrow until maturity and are the most abundant phagocytes in the circulating blood. Their role is to attack and engulf invading microorganisms.

The role of monocytes is to produce the other two types of phagocytes; macrophages and dendritic cells. Under normal conditions, monocytes differentiate to replenish the current pool of macrophages and dendritic cells, while during infection monocytes circulate into the blood and travel to the sites of infection to produce new macrophages or dendritic cells. Macrophages are large eaters of cellular debris and pathogens.

Finally, the megakaryocytic erythroid branch of the haematopoietic system gives rise to erythrocytes through a process called erythropoiesis, or thrombocytes through a process called megakaryopoiesis, via their megakaryocytic erythroid progenitors. Both platelets and red blood cells are small enucleated cells with a discoid-shaped cytoplasm, although they perform completely different functions. Erythrocytes are responsible for carrying oxygen and carbon dioxide between the lungs and the tissues, while thrombocytes maintain haemostasis by surveying the vessel wall for damage and if damage is identified, thrombocytes initiate a multi-step repair process that leads to the formation of a platelet clot.

1.1.1.1 *Megakaryopoiesis and platelet production*

Platelets are short-lived cells that circulate in the blood surveying the vascular wall for injury. On average they live for 7-10 days and need to be replenished continuously. An adult produces about $4x10^{11}$platelets daily. Platelets are produced by their bone-marrow residing precursors, called megakaryocytes.

Megakaryocytes are the largest type (averaging 50-100 μm in diameter) of precursor cells that lie in the bone marrow. To produce platelets, megakaryocytes go through a maturation process, called megakaryopoiesis (reviewed in (Italiano, 2013; Patel et al., 2005)). During their maturation, megakaryocytes stop dividing, but continue to replicate their DNA in a series of endomitotic cell cycles (Martin et al., 2012; Ravid et al., 2002). At the end of the maturation process, megakaryocytes are polyploid containing from 16 to 256 copies of their genome (Odell et al., 1965) and with a large cytoplasm that is full of platelet-specific granules (Handagama et al., 1987).

Platelet production occurs through long, thin extensions of the megakaryocyte cytoplasm (Italiano et al., 1999). Such cytoplasmic remodelling and maturation is achieved by microtubules that extend throughout the cytoplasm creating the proplatelets (Schwer et al., 2001). The last essential step in platelet biogenesis is the filling of nascent platelets with organelles and alpha and dense granules. Motor proteins attached to the microtubules carry the necessary components for platelet function from the cytoplasm to the proplatelets. These organelles then become trapped in the proplatelet ends and the microtubules elongate and separate the proplatelet from the megakaryocytic cytoplasm (Richardson et al., 2005). At the end of thrombopoiesis small anucleated cell fragments are released into the blood stream.

1.1.2 *Alternative views to the "classical" model of haematopoiesis*

Linear representations are widely used to analyse normal haematopoiesis. Although these representations oversimplify the haematopoietic development, it provides a useful framework in which to consider the properties of the intermediate cell populations. However, numerous studies have provided

evidence that this over-simplified version does not depict reality in all aspects (Orkin, 2000).

One of the most recent findings suggests the existence of another type of progenitors that is not included in the classical haematopoietic ontogeny described above. These progenitors, termed early lymphoid progenitors (ELP) or lymphoid-primed multipotent progenitors (LMPP) have lost their megakaryocytic and erythrocytic potential (Igarashi et al., 2002). Using an additional cell surface marker, Flt3, researchers have isolated this subset of multipotent progenitors (Adolfsson et al., 2005; Christensen and Weissman, 2001) and suggest that the loss of the potential to produce granulocytes is the last stage before the transition to lymphoid restricted progenitors (Adolfsson et al., 2001; Akashi et al., 2005). According to the "revised" model, the LMPPs can be added on top of the CLPs in the hierarchy tree with the ability to produce neutrophilic granulocytes only, while common myeloid progenitors can give rise to eosinophilic and basophilic granulocytes (Gorgens et al., 2013).

Another theory that has emerged in recent years concerns the haematopoietic stem cell pool. Comparative studies between haematopoietic stem cells from young and aged mice identified that these two populations are different (Geiger and Van Zant, 2002; Marley et al., 1999) in many features including their ability to home to the bone marrow (Morrison et al., 1996), in their surface immune-phenotype (Liang et al., 2005; Marley et al., 1999), cell cycle properties, gene expression and lineage bias. In aging mice the haematopoietic stem cells produce more myeloid cells rather than lymphoid, which is consistent with the age-related impairment of the immune system. Similar studies in human bone marrow haematopoietic stem cells confirmed these findings. The haematopoietic stem cells in aged individuals were reported to be more frequent and less

quiescent. In addition, both *in vitro* and *in vivo* differentiation showed a myeloid-bias of the aged haematopoietic stem cells compared to their young counterparts. The same bias was also found in transplantation experiments, where the lymphoid progeny was not as efficiently generated (Pang et al., 2011).

Despite this observed lineage bias, the belief was that these cells still emerge from a homogeneous stem cell population. However, recent studies have shown that the stem cell pool is heterogeneous and contains subsets of cells that differ in their life span, cycling status and lineage bias even within young mice. The haematopoietic stem cells are now thought to be split into myeloid-biased, lymphoid-biased and unbiased and their lineage bias is maintained even after serial transplantations (Beerman et al., 2010; Challen et al., 2010; Dykstra et al., 2007; Muller-Sieburg et al., 2002). Recently, Sanjuan-Pla et al. (2013) identified and isolated a subset of HSCs primed towards the megakaryocytic lineage, with a propensity for short- and long-term reconstitution of platelets. This HSC subtype can also give rise to myeloid- and lymphoid-biased HSCs, placing it at the top of the hierarchy of all HSC subtypes.

The above mentioned studies have been performed in mice and there is no evidence that these findings are true for human haematopoiesis. It is, therefore, essential to examine the human haematopoietic system for such characteristics.

1.1.3 *Isolation of haematopoietic cells using cell surface antigens*

Over the last thirty years extensive studies on haematopoiesis have focused on the identification and characterisation of homogeneous cell populations at various stages along the differentiation of multipotent haematopoietic stem cells to mature blood cells. The introduction of new technologies such as multicolor fluorescent-activated cell sorting and monoclonal anti-

bodies greatly contributed to these efforts. The haematopoietic tree described in the section above summarises three decades of findings by researchers all over the world.

The self-renewal and differentiation properties of the haematopoietic stem cells were well known before the isolation of these cells due to bone marrow transplantations (Till and McCulloch, 1961). Historically haematopoietic stem cells have been defined based on their competence to reconstitute the haematopoietic system in lethally irradiated mice (Spangrude et al., 1988) and primates (Berenson et al., 1988, 1991; Smith et al., 1991). The regenerative properties of stem cells marked a whole new era in medicine for their clinical use in transplantations and gene therapies. Despite their importance HSC molecular characterisation has lagged behind, also due to the low number of cells present in any sample.

The first attempts to distinguish the bone marrow compartment with self-renewing ability reported that CD34 is a glycoprotein present in the surface of haematopoietic progenitor cells (Andrews et al., 1986; Brown et al., 1991; Civin et al., 1987; Katz et al., 1985). Little is still known about the functions of CD34; Krause et al. (1996) presented a comprehensive review of its expression, structure and clinical use. Despite the CD34+ fraction defining stem cells in clinical settings, CD34 on its own is not sufficient as an HSC marker, due to the fact that the CD34+ fraction of the bone marrow contains cells that are already committed to either the monomyeloid, lymphoid or erythroid lineage (Andrews et al., 1989; Loken et al., 1987; Mayani et al., 1989; Sutherland et al., 1989).

Subsequent studies aimed at identifying cell populations within the $CD34^+$ that did not bear any lineage-associated markers (Lin^-). In 1991, Terstappen et al. (1991) show that CD38 is only expressed in the fraction of $CD34^+$ bone marrow cells

that also express other lineage specific markers. Hence, CD38 is identified as a negative marker of multipotency and an increase of CD38 indicates differentiating $CD34^+$ cells. Apart from forming long term cultures, $CD34^+CD38^-$ cells are also able to reconstitute haematopoiesis in irradiated mice (Bhatia et al., 1997; Conneally et al., 1997). In parallel other groups identified enriched haematopoietic stem cells using antigens against CD90 (also called Thy1) (Baum et al., 1992; Craig et al., 1993) or a specific CD45 protein isoform, termed CD45RA (Mayani et al., 1993). CD45 is a transmembrane protein that has five different protein isoforms that differ in their extracellular domain (Lynch, 2004). Out of the five CD45 isoforms, the two longest ones are the ones termed CD45RA. Mayani et al. (1993) reported that the $CD34^+$ $CD45RA^{low}CD71^{low}$ cell population contains multiprogenitor cells, while the expression of CD90 in the $CD34^+$ cell population establishes long term cell cultures and has reconstitutive abilities when transplanted in lethally irradiated mice.

To separate haematopoietic stem cells from multipotent progenitors, loss of CD90 is sufficient in the $CD34^+CD38^-CD45RA^-$ compartment (Majeti et al., 2007). Notta et al. (2011) went further to show that, within the $CD90^+$ compartment integrin ITGA6 (CD49f) identifies the HSC able to long term repopulate.

The finding that the $CD34^+$ cells form a heterogeneous population of progenitors, sparked extensive research to identify of what other more committed progenitors consist. In 1993, Fritsch et al. (1993) reported that a fraction of the $CD34^+$ cells that does not express CD45RA comprises early myeloid precursors. It was not until 2002 that the separation of three different myeloid progenitors was achieved within the $Lin^-CD34^+CD38^+$ compartment (Manz et al., 2002). Common myeloid progenitors that can give rise to both erythroid and myeloid population have a IL-3$Ra^{low}CD45RA^-$ phenotype. More committed progenitors are identified through IL-3$Ra^+CD45RA^+$ for the

granulocytic/monocytic branch and IL-3Ra^{-}CD45RA^{-} for the megakaryocytic/erythroid one.

For the isolation of early lymphoid progenitors there were two studies reporting either CD10 (Galy et al., 1995) or CD7 (Hao et al., 2001). Within the Lin^{-}CD34^{+}CD38^{+} subpopulation, Galy et al. (1995) identified a fraction of cells that express CD10 and CD45RA on their surface. Later study, though, examined the Lin^{-}CD34^{+}CD38^{-} compartment for a subpopulation that expressed CD7. Both these populations are devoid of any myeloid differentiating potential, while producing all of the lymphoid and dendritic cells.

1.2 GENE EXPRESSION AND ITS REGULATION

Despite the functional and phenotypic differences between cells, such as the different cell types of the haematopoietic system described above, in any given individual all cells share the same genetic code. The diversity of the cells is achieved through a well-orchestrated process of decoding the common genetic information in a cell-type specific manner. This process is termed gene expression. Understanding gene expression under normal conditions improves our understanding of gene expression in disease, which is in turn invaluable for medical advances that aim to restore normal function of the affected cells.

1.2.1 *From DNA to RNA and proteins, or simply gene expression*

Proteins are the key components of cells, carrying out diverse functions that are essential for the maintenance and normal function of a cell. In all living cells the genetic information is encoded in the DNA, which are long double-stranded molecules (Watson and Crick, 1953) formed by four types of repeating monomers (A, C, G and T). The flow of genetic information follows the path from DNA to RNA to protein; a prin-

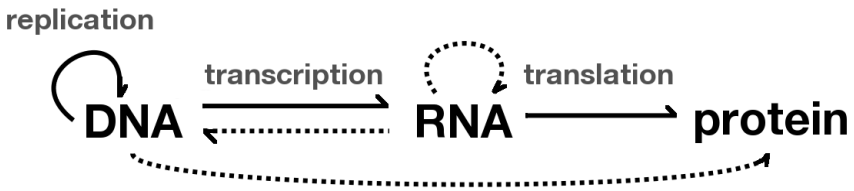


Figure 2: Central dogma of Molecular Biology. The dogma classifies the transfers of information between the three macromolecules, DNA, RNA and protein into: general transfers (straight line), special transfers (dotted line), and unknown transfers.

ciple holding true in all cells from bacteria to humans, termed the central dogma of molecular biology that was formulated by Fancis Crick in 1958 (Crick, 1958) (see Figure 2).

During transcription, only certain segments of DNA, the genes, are used as a template for the synthesis of RNA molecules, called precursor mRNA (pre-mRNA). The products of transcription go through several processing steps, including 5′ capping, poly-A tail addition and splicing, which turn them into processed messenger RNA (mRNA). The mRNA is then exported from the nucleus, where it was synthesised, to the cytoplasm, where it will be translated into a protein. Although the general rule dictates that proteins are the final products of gene expression, sometimes the process terminates with the production of RNAs. In these cases the gene transcribed does not encode for a protein, but for certain types of RNA that along with proteins serve various cellular functions. The most common types of non-protein-coding RNAs along with their functions are summarised in Table 1.

Table 1: Types of non-protein-coding RNAs (reviewed in (Eddy, 2001)).

Name	Abbreviation	Function	Reference
ribosomal RNA	rRNA	form the basic structure of the ribosome and catalyse protein synthesis	(Attardi and Amaldi, 1970)
transfer RNA	tRNA	central to protein synthesis as adaptors between mRNAs and amino acids	(Lengyel and Soll, 1969)
small nuclear RNA	snRNA	function in a variety of nuclear processes, including the splicing of pre-mRNAs	(Busch et al., 1982)
small nucleolar RNA	snoRNA	used to process and chemically modify rRNAs	(Eliceiri, 1999)
microRNA	miRNA	regulate gene expression typically by blocking translation of selective mRNAs	(Lau et al., 2001; Lee and Ambros, 2001)
small interfering RNA	siRNA	turn off gene expression by directing degradation of selective mRNAs	(Hamilton and Baulcombe, 1999)
long non-coding RNA	lincRNA	function in diverse cell processes, including telomere synthesis, X-chromosome inactivation	(Ulitsky and Bartel, 2013)

In bacteria, genes are located in uninterrupted stretches of DNA. In contrast, genes in higher eukaryotes contain both coding (exons) and non-coding (introns) regions (Berget et al., 1977; Chow et al., 1977). Both exons and introns are transcribed into pre-mRNA. However, introns are spliced out by a ribonucleoprotein complex, called spliceosome, during the pre-mRNA processing step (Jurica et al., 2004; Matlin and Moore, 2007). The splicing step offers the possibility to generate greater protein complexity in eukaryotes, as alternative splicing, which is the removal of introns and different stitching of the remaining exons, can generate various protein isoforms from the same DNA locus (Matlin et al., 2005).

1.2.2 *Regulation of gene expression*

Cell identity and normal cell functions are maintained through tight spatio-temporal regulation of gene expression. Not all genes are expressed at a given time and in any given cell. Even those that are expressed, though, are controlled as to how much proteins they produce depending on the cell's needs and external stimuli. The control of gene expression can be exercised at any stage of the transition from DNA to proteins. Hence, we can identify six levels of regulation during this process (Alberts, 2008) :

1. Transcriptional control: regulation of the transcription of a gene. This is the most prevalent mechanism of gene ex-

pression regulation, and special mention is made to it in the following section.

2. RNA processing control: regulation of processing and splicing of pre-mRNAs. For example, alternative splicing, described above, determines which protein isoforms are expressed.

3. RNA transport and localisation control: only fully processed mRNAs are transported out of the nucleus. Otherwise, they are degraded.

4. mRNA degradation control: performed by siRNAs through the RNA-induced silencing complex (RISC) and sequence-specific binding onto the target mRNA.

5. Translational control: performed either through the mRNA surveillance system, such as the nonsense mediated mRNA decay mechanism, which eliminates defective mRNAs prior to translation, or through miRNAs, which bind in a sequence-specific way onto mRNAs and regulate their stability and translation.

6. Post-translational control: in order to become functional certain synthesised proteins may require additional activation or deactivation or sub-cellular localisation.

1.2.3 *Elements of transcriptional regulation*

In marked contrast to the expectation that the human genome contains over a hundred thousand genes, the completion of the Human Genome Project (described in Section 1.3) revealed that there are only 35,000, an estimation that was later dropped to approximately 21,000 (International Human Genome Sequencing Consortium, 2001, 2004). A subset of these genes encode for proteins that regulate the transcription of genes, including some times their own. They are called transcription

factors and constitute the trans-acting regulatory elements because their function extends beyond their own genomic region. In the human genome it is estimated that at around 10% of the annotated protein coding genes are transcription factors (Vaquerizas et al., 2009). To regulate transcription, transcription factors bind to DNA regulatory sites. The genomic loci where the transcription factors are bound are called transcription factor binding sites (TFBSs) and constitute the cis-regulatory elements. These cis-regulatory elements can be classified into promoters, enhancers, silencers, and insulators among others (Maston et al., 2006b). Both promoters and enhancers are short genomic loci that serve as platforms for the binding of transcription factors. While promoters are located either in the immediate vicinity upstream of genes, enhancers can be many kilobases away from the gene they regulate (Spitz and Furlong, 2012).

The next level of transcriptional control involves the packaging and the accessibility of DNA and it is referred to as epigenetic. DNA packaging and compaction in the nucleus is achieved through chromatin, a structure combining DNA and proteins. The fundamental repeated subunit of chromatin is the nucleosome (Kornberg, 1974), around which 146 bp of DNA are wrapped (Luger et al., 1997). The nucleosome is a protein octamer composed of four histones (H2A, H2B, H3 and H4), present in two copies each that heterodimerise in H2A/H2B and H3/H4. The chromatin structure is dynamic and subject to continuous modifications (Bell et al., 2011).

The set of these inheritable modifications that do not alter the DNA sequence itself are called the epigenome. Epigenetic modifications include methylation of the cytosine DNA nucleotides (Bird, 2002), nucleosome re-positioning (Whitehouse et al., 1999) and enzymatic modifications on the histone tails (Allfrey and Mirsky, 1964; Kouzarides, 2007; Zhang and Reinberg, 2001). In the last few years, great advances in the field of genom-

ics (reviewed in Section 1.3.1) have facilitated the completion of genome-wide studies of the epigenome and how it affects transcription. These studies have identified chromatin states that are associated with specific cis-regulatory elements that confer a certain transcriptional output; either enhancement or inhibition (reviewed in (Bell et al., 2011; Zhou et al., 2011)).

In those studies, different chromatin signatures have been used to identify promoters and enhancers. Promoters are in general defined by a H3K4 tri-methylation (H3K4me3) and a H3K27 acetylation (H3K27ac) signal when active, while they bear a repressive histone mark, H3K27 tri-methylation (H3K27me3) when inactive (reviewed in (Lenhard et al., 2012)). Absence of H3K4me3 and enrichment of H3K4 monomethylation (H3K4me1) are indicative of enhancer regions (ENCODE Project Consortium, 2007; Heintzman et al., 2007). When these are present along with H3K27 acetylation (H3K27ac), this chromatin state strongly correlates with enhancers that are in proximity of active genes (Bonn et al., 2012; Creyghton et al., 2010; Rada-Iglesias et al., 2011), while lack of H3K27ac is associated with repressed enhancers. Active enhancers are also nucleosome free, as revealed by DNase I hypersensitivity studies, and show low methylation of cytosines (Xu et al., 2007).

1.2.4 *Key regulators of haematopoiesis*

Haematopoietic multipotent cells exhibit lineage priming, which is the low expression of lineage specific genes (Hu et al., 1997). In undifferentiated cells, these lineage regulators are marked by 'bivalent' chromatin states at their promoter and are subject to complex regulation including the interplay of enhancers and promoter elements and are also associated with an increased likelihood of transcriptional induction during differentiation (Adli et al., 2010). Hence, lineage commitment and subsequent differentiation of multipotent cells involves activa-

tion of lineage-specific and inhibition of 'lineage-inappropriate' genes. These lineage-specific programs are orchestrated by transcription factors that affect differentiation through their expression levels (Iwasaki et al., 2003), timing (Iwasaki et al., 2006) and co-operative or antagonistic interplay with other transcription factors (Dahl et al., 2003; Iwasaki et al., 2006). In this section, I review a set of transcription factors whose key regulatory role in haematopoiesis has been established through knockout or over-expression studies in model organisms (summarised in Table 2).

1.2.4.1 *Regulators of HSC induction and maintenance*

Specification of haematopoiesis is dependent on the right set of cues that promote differentiation of a haematopoietic-endothelial precursor, called haemangioblast, in to the haematopoietic lineage (Lacaud et al., 2001). Different studies have pointed out transcription factors that are important for either primitive or definitive haematopoiesis, or both. For example, SCL/TAL1 and its associated partner, LMO2, are individually crucial for both processes (Patterson et al., 2007; Robb et al., 1995; Shivdasani et al., 1995a). RUNX1 (Okuda et al., 1996; Wang et al., 1996) and MLL (Ernst et al., 2004), however, are essential only for the definitive haematopoiesis (Ernst et al., 2004), while SOX17 knockout has major effect in primitive haematopoiesis only (Kim et al., 2007).

The transcription factors that are essential for the induction of haematopoeisis are not always required for the normal function of the haematopoietic stem cells. Conditional inactivation of SCL and RUNX1, for example, does not affect adult haematopoietic stem cell function (Mikkola et al., 2003). However, these two transcription factors were identified to act synergistically as gatekeepers of blood stem cell development, along with GATA2 (Wilson et al., 2010). Haematopoietic stem cells also express transcription factors that are important for their

Table 2: Altered haematopoietic phenotypes from knockout (-) or over-expression (+) of transcription factors. The altered phenotypes were either a functional defect (func), a decreased (decr) or increased (incr) number of cells, a complete lack of cells (lack), or a block in maturation (matur). Table adapted from (Laiosa et al., 2006).

TF	TF family	HSC	B	T	NK	EOS	BAS	GRAN	MACRO	NEUTR	ERY	MK	References
RUNX1	RUNT	induc^{-}	matur^{-}	matur^{-}						incr^{-}		matur^{-}	(Growney et al., 2005; Ichikawa et al., 2004)
SCL/TAL1	bHLH	induc$^{-/+}$									lack^{-}	lack^{-}	(Gering et al., 1998; Mikkola et al., 2003; Robb et al., 1995; Wilson et al., 2010)
MLL	SET	induc^{-}/func^{-}	lack^{-}										(Ernst et al., 2004; McMahon et al., 2007)
GATA1	GATA zinc-finger					matur^{-}					matur^{-}	matur^{-}	(Fujiwara et al., 1996; Shivdasani et al., 1997; Yu et al., 2002)
GATA2	GATA zinc-finger	func$^{-/+}$									matur^{+}	incr^{+}	(Ikonomi et al., 2000; Persons et al., 1999; Tsai et al., 1994)
LMO2	SET	induc^{-}/func^{-}									matur^{-}		(Patterson et al., 2007; Warren et al., 1994; Yamada et al., 1998)
FLI1	ETS		incr^{+}								func^{-}	matur^{-}/incr^{+}	(Hart et al., 2000; Spyropoulos et al., 2000; Zhang et al., 1995)
PBX1	TALE homeodomain	func^{-}	lack^{-}		lack^{-}								(Ficara et al., 2008; Sanyal et al., 2007)
ETS1	ETS		matur^{-}	func^{-}	func^{-}								(Barton et al., 1998; Bories et al., 1995; Eyquem et al., 2004)
GABPA	ETS	func^{-}		func^{-}				matur^{-}					(Yang et al., 2011; Yu et al., 2010, 2011)
NFE2	bZIP											matur^{-}	(Shivdasani et al., 1995b)
KLF1	Kruppel C2H2-type zinc-finger										matur^{-}		(Nuez et al., 1995)
MEIS1	TALE homeodomain	func^{-}/incr^{+}									func^{-}	lack^{-}	(Azcoitia et al., 2005; Cvejic et al., 2011; Hisa et al., 2004; Wang et al., 2006)
HOXA9	Homeodomain	induc^{-}/func^{-}					decr^{-}						(Lawrence et al., 1997)
HOXB4	Homeodomain	induc^{-}/func^{-}/incr^{+}											(Brun et al., 2004; Sauvageau et al., 1995)

specification and/or maintenance, which in addition have crucial roles in lineage specification in the later stages of haematopoiesis. These factors include RUNX1 (Ichikawa et al., 2004) and MEIS1 (Hisa et al., 2004) with a megakaryocytic role, SCL with an involvement in erythroid and megakaryocytic differentiation (Mikkola et al., 2003) and PU.1 and CEBPA with a role in myeloid priming (Reddy et al., 2002).

Of particular interest is the switch of expression between two members of the GATA family, GATA1 and GATA2, early in haematopoietic differentiation. GATA2 is highly expressed in haematopoietic stem cells and indispensable for the induction and maintenance of these cells (Tsai et al., 1994). Its overexpression, though, downstream of the stem cell compartment blocks any lineage differentiation (Persons et al., 1999), suggesting that decrease of GATA2 expression is essential for commitment in differentiation. Expression of GATA2, and the resulting self-renewal capacities of haematopoietic stem cells are disrupted by GATA1 that antagonises GATA2 by competitively binding to the upstream regulatory element of GATA2. When expressed, GATA1 binds to this site, blocking the positive feedback loop of GATA2 and prompting haematopoietic stem cells to commit to differentiation (Grass et al., 2003).

1.2.4.2 *Regulation of megakaryopoiesis and erythropoiesis*

GATA1 and GATA2 are also involved in the regulation of gene expression in the megakaryocytic-erythroid branch, with GATA1 widely expressed in the myeloid lineage and indispensable for eosinophil development (Yu et al., 2002). GATA1 deficient mice die between E10.5 and E11.5 because of severe anemia (Fujiwara et al., 1996; Pevny et al., 1991) and later studies pointed out that GATA1 regulates the switch of fetal to adult haemoglobin (Bacon et al., 1995). Its essential role in megakaryocytes was not identified until after a conditional knockout in these cells, which caused impaired megakaryocytic maturation

and reduced platelet production (Shivdasani et al., 1997). However, the ablation of GATA1 does not prevent haematopoietic progenitors to commit to the megakaryotic erythroid lineage, suggesting a role in late maturation of these cells (Laiosa et al., 2006). As for GATA2, it was shown to re-program an erythroleukemic cell line (K562) into the megakaryocytic lineage (Ikonomi et al., 2000).

GATA transactivation is achieved through the recruitment of co-factors of the FOG (friend of GATA) transcription factor family (Fox et al., 1999). Mice deficient in a member of this family, FOG1, exhibited a similar erythroid phenotype to that of GATA1 null mice (Tsang et al., 1998), suggesting synergistic regulation between FOG1 and GATA1 in erythroid cells. However, the lack of FOG1 caused a more severe impairment in the megakaryocytes than the one observed in GATA1 deficient cells (Tsang et al., 1998). A similarly severe megakaryocytic phenotype was only observed after double knockout of both GATA1 and GATA2, implying a GATA1-independent role for FOG1 in these cells (Tsang et al., 1998), likely mediated through GATA2 (Chang et al., 2002).

Despite their importance in these cells, none of the transcription factors mentioned so far is responsible for lineage-specification in the megakaryocytic or erythroid lineage. This is achieved through an antagonistic interplay between transcription factors, a frequently employed means of reinforcing lineage choices in the haematopoietic system (Graf and Enver, 2009). In this case the duet of transcription factors includes FLI1 and KLF1, which drive progenitors to the megakaryocytic and erythroid lineage, respectively. FLI1, however, has a dual role in megakaryopoiesis, as it also regulates late stages of megakaryocytic maturation as well (Pang et al., 2006). FLI1, along with GABPA, belong to the ETS transcription factor family, whose members are key regulators throughout haematopoiesis. Their

function is usually induced by their interaction with proteins of other families (Verger and Duterque-Coquillaud, 2002). In contrast to FLI1, GABPA regulates early megakaryocytic regulation (reviewed in (Tijssen and Ghevaert, 2013)).

Finally, NFE2 is a regulator of proplatelet formation. The NFE2 protein is a heterodimer consisting of a large p45 subunit and a small p18 subunit (Igarashi et al., 1994). It is expressed in both megakaryocytes and erythroid cells. NFE2 deficient mice die of hemorrhage and exhibit a complete lack of circulating platelets caused by a blockage in the proplatelet formation (Lecine et al., 1998). In erythroid cells, there are several NFE2 binding sites in the genomic locus controlling the expression of adult globins (Talbot and Grosveld, 1991). The effect of knockout of either NFE2 or either of its subunits independently, however, has only a mild effect in erythroid development (Martin et al., 1998).

1.2.4.3 *HOX genes and their co-factors, PBX1 and MEIS1*

Of special interest in development in general, but also for its role in haematopoiesis is the HOX transcription factor family. It consists of 39 genes organised in four genomic clusters (A - D) and spread over four different chromosomes (reviewed in (Duboule, 2007)). HOX genes, with the exception of those in cluster D, are expressed in various blood cell types. Their expression peaks in haematopoietic stem and progenitor cells and is either absent or significantly down-regulated along commitment (Giampaolo et al., 1994; Moretti et al., 1994; Sauvageau et al., 1994). The importance of HOX genes in the function and maintenance of haematopoietic stem cells is evident from the fact that over-expression of some of these rescues the lack of HSCs observed in $MLL^{-/-}$ mice, suggesting that MLL is regulating their expression in these cells (Ernst et al., 2004).

HOXB4 and HOXA9 are two of the HOX genes extensively studied in haematopoiesis. HOXB4 over-expression is sufficient to promote haematopoietic stem cell expansion (Sauvageau et al., 1995). However, HOXB4 knockout in mice causes a modest reduction in the stem cell numbers (Brun et al., 2004), suggesting a potential redundancy among the HOX genes. In contrast, HOXA9 is the HOX gene causing the most severe haematopoietic phenotype when knocked out, as it impairs the stem cells' repopulating activity and defects in multiple other lineages (Lawrence et al., 1997). For HOX proteins to confer their specificity, however, they require to interact interactions with other transcription factors, such as the TALE homeobox transcription factors (Chang et al., 1996). Through interaction with these co-factors their binding affinity, specificity, and/or stability increase significantly (Abramovich et al., 2005).

MEIS1 is one of the TALE homeobox proteins interacting with HOX genes. It is expressed in the haematopoietic stem cell compartment (Abramovich et al., 2005) and positively regulates the expansion and pool size of the HSCs (Wang et al., 2006). Homozygous MEIS1 knockout mice exhibit an embryonic lethal phenotype characterised by multiple haematopoietic defects such as extensive bleeding due to complete lack of megakaryopoiesis (Azcoitia et al., 2005; Hisa et al., 2004) and severe reduction in the number and colony-forming potential of haematopoietic stem cells (Wang et al., 2006). In-depth studies of MEIS1 in zebrafish defined its role in haematopoiesis and vasculogenesis more precisely. Cvejic et al. (2011) showed a complete ablation of thrombocyte formation (the zebrafish equivalent of megakaryocytes) and substantial effects on erythropoiesis and vasculogenesis. In vitro studies with primary human cells have also shown that MEIS1 can affect myeloid differentiation and proliferation (Abramovich et al., 2005).

Hox proteins form heterodimers and/or trimers with members of the PBX and MEIS gene families at different stages of haematopoiesis in order to modulate their gene function (Pineault et al., 2002). HOX proteins from paralog groups 1-10 can gain specificity through cooperative binding with PBX family members, whereas HOX proteins from paralog groups 9-13 have been shown to cooperatively bind DNA with members of the MEIS family (Abramovich et al., 2005). The interaction of PBX1 with co-repressors, such as histone deacetylases, mediates repression, while MEIS1 contains a C-terminal domain that promotes transcactivation in response to cell signals.

1.3 NEXT-GENERATION SEQUENCING

The Human Genome Project was an ambitious research project that proved to be a milestone in the field of genomics. It was formally launched in 1990 and was aiming to identify all the genes in the human genome and determine all of its three billion nucleotides within 15 years (Collins et al., 2003). Simultaneously, bioinformatic analysis tools and databases were built, greatly expanding the impact of this project in human biomedical research.

When the project was conceived, Sanger sequencing was widely used to decipher genomic sequences. Sanger sequencing, published in 1977 (Sanger et al., 1977), had revolutionised the existing sequencing technologies that were based on cleavage of short sequences of nucleotides and their identification using their migration characteristics on two-dimensional chromatography paper (Gilbert and Maxam, 1973). For 30 years after the first published method by Sanger and Coulson (1975), Sanger sequencing was still extremely popular despite its lack of throughput.

A next revolution in sequencing was achieved with the introduction of the first next-generation sequencing platform, the GS 20 (454 Life Sciences). The first whole genome to be sequenced by this method was the bacteria *Mycoplasma genitalia* at 96% coverage and 99.96% accuracy (Margulies et al., 2005). Similar to Sanger sequencing, next-generation sequencing consists of iterative steps of nucleotide incorporation, followed by a detection step where the output, either light or pH change, is captured and a wash step, where any blocking terminators are removed. However, these steps are performed in parallel on millions of DNA fragments. What made possible this parallelisation of the process was the invention of flow cells, where the different DNA fragments are spatially separated.

Prior to loading the DNA fragments onto the sequencing machine, the DNA is converted into a library of DNA fragments. During this process the DNA is fragmented and a set of platform specific adaptors are ligated on to the DNA fragments. Once the library is loaded, the fragments attach to the complementary adaptors that are on the flow cell. Then the fragments are amplified *in situ* on the flow cell. This amplification step is needed to provide sufficient signal during each DNA reaction step.

Currently, four next-generation sequencing platforms are available: Roche 454, Illumina, SOLiD and Ion Torrent. These platforms differ in their amplification step, sequencing approaches and output (reviewed in (Mardis, 2013; Pareek et al., 2011)), with Illumina being the most widely used platform.

All platforms come with their own base-calling algorithms, that translate images to sequences and their equivalent confidence, called quality scores. The output of a sequencing run, after base-calling, is a FASTQ file (Cock et al., 2010) that con-

tains all the reads that passed each algorithms quality thresholds along with their equivalent quality scores.

Compared to Sanger sequencing, next-generation sequencing produces shorter lengths and slightly less accurate results (Pareek et al., 2011). However, the advantage of massive parallel sequencing producing millions of reads exceeds by far such limitations. Nevertheless, it also introduces several sources of errors and biases, during both library preparation and sequencing. These include PCR amplification errors and biases, adaptor ligation biases and base calling errors due to different phasing (fragments within a cluster of amplified fragments are at different phases of elongation due to inefficient cleavage of prior terminators).

Efforts to eliminate the PCR amplification step needed for signal boosting led to the introduction of new platforms for sequencing, termed third or next-next-generation sequencing machines. Those aim to sequence single DNA molecules. Some of the new commercial platforms in this category are: Oxford Nanopore, Helicos Heliscope and Pacific Biosciences. Single molecule sequencing could eliminate some of the biases and errors introduced by current technologies. However, the lack of abundant signal due to the single molecules poses great challenges in the implementation of these platforms. Despite having being announced already two years ago, few of those platforms have become available to sequencing centres. One of the major drawbacks being their high error rates (Mardis, 2013).

1.3.1 *Next-generation sequencing applications*

The continuously plummeting cost of next-generation sequencing and its accessibility outside of large sequencing centres created a gradual increase in the range of sequencing applications (Shendure and Ji, 2008). The earliest sequencing pro-

jects aimed at creating reference genomes for model organisms. Nowadays, however, whole genome sequencing projects have expanded to various species, even extinct ones (Noonan et al., 2006), making it possible to perform phylogenetic studies. Using the reference genomes as a scaffold we can then re-sequence different individuals to assess genetic variation within the population. Finally, even within an individual we can identify genetic and epigenetic differences between different cells and conditions or study various processes that are taking place within a cell.

Table 3: List of widely-used next-generation sequencing applications (table adapted from (Shendure and Lieberman Aiden, 2012)). ChIP-seq: chromatin immunoprecipitation sequencing, CLIP-Seq: cross-linking immunoprecipitation sequencing, DNA-seq: DNA sequencing, DNase I-seq: DNase digestion sequencing, FAIRE-seq: formaldehyde-assisted isolation of regulatory elements sequencing, MNase-seq: micrococcal nuclease digestion sequencing, Ribo-seq: ribosome profiling sequencing, and RNA-seq: RNA sequencing.

Method	Starting material	Example reference
ChIP-seq	DNA sequence	(Johnson et al., 2007)
CLIP-seq	RNA sequence	(Yeo et al., 2009)
DNA-seq	DNA sequence	(Margulies et al., 2005)
DNase I-seq	DNA sequence	(Margulies et al., 2005)
FAIRE-seq	DNA sequence	(Gaulton et al., 2010)
Hi-C	DNA sequence	(Lieberman-Aiden et al., 2009)
Methyl-seq	DNA sequence	(Brunner et al., 2009)
MNase-seq	DNA sequence	(Margulies et al., 2005)
Ribo-seq	RNA sequence	(Ingolia et al., 2009)
RNA-seq	RNA sequence	(Marioni et al., 2008)

Shendure and Lieberman Aiden (2012) provide a detailed review of all available sequencing applications, some of which are summarised in Table 3, and briefly describe the library preparation steps of each one. In the following sections I will provide an overview of the two sequencing applications used in this thesis: RNA-seq and ChIP-seq.

1.3.2 *ChIP sequencing*

Chromatin immunoprecipitation (ChIP) is a technique to map the protein-DNA interactions and epigenetic marks that regulate gene expression (Solomon et al., 1988). A ChIP protocol consists of the following steps:

1. cross-linking of proteins onto the DNA using formaldehyde,
2. nuclei isolation and lysis,
3. DNA extraction followed by fragmentation,
4. immunoprecipitation of the DNA fragments using a specific antibody against the protein of interest, and
5. reverse cross-linking to dissociate proteins from the DNA fragments.

What remains to be done is identify the genomic loci that the protein was bound to.

The first ChIP-seq experiments were published in 2007 (Johnson et al., 2007; Robertson et al., 2007), and since then thousands have followed making ChIP-seq one of the most widely used and standardised next-generation sequencing applications. With certain modifications, the ChIP protocol described above can be used to study epigenetic modifications such as histone marks (Barski et al., 2007; Mikkelsen et al., 2007), nucleosome positioning (Schones et al., 2008) and DNA methylation (Down et al., 2008). During library preparation for ChIP-seq, the immunoprecipitated DNA needs to be subjected to size selection, typically between 200-300 bp, using gel electrophoresis. The ENCODE and modENCODE consortia have developed a set of working standards and guidelines for ChIP experiments to address such issues (Landt et al., 2012).

Bioinformatics analysis usually consists of an alignment step in the beginning, where the reads are aligned back to the reference genome. Popular aligners include Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009). This step is followed by identification of enriched regions by tools called peak callers (Fejes et al., 2008; Jothi et al., 2008; Xu et al., 2010; Zhang et al., 2008). Downstream analysis depends on the type of questions addressed and can include *de novo* motif analysis and peak annotation in respect to nearest genes among other.

1.3.3 *RNA sequencing*

The transcriptome is the full repertoire of transcripts expressed in a cell at a given developmental stage and condition. Identifying the transcriptome is essential for understanding the cellular mechanisms governing both developmental and disease states. The study of the transcriptome aims to catalogue and quantify the full set of transcripts, identify their splicing patterns, accurately define their 5′ and 3′ ends and identify even subtle differences between different conditions (Ruan et al., 2004).

Sanger sequencing has proven useful also in identifying the nucleotide sequence of RNAs. Using cDNA or expressed sequence tags (ESTs), researchers were able to identify transcripts in a high-cost, low-throughput and not quantitative way (Adams et al., 1991; Gerhard et al., 2004). One of the most commonly used tools for the quantification of transcript expression is qPCR, which has a limited throughput however, making it impractical for genome-wide studies.

The invention of microarrays revolutionised transcriptomics, as they offer genome-wide profiling of transcript expression. Microarray platforms contain thousands of DNA probes (approximately 50 bp long) targeting specific regions of genes.

The transcriptome in question is fragmented, converted into cDNA strands, labeled with fluorescent dyes and finally incubated along with the probes. The cDNA fragments are hybridised onto the complementary probes. Light excitation of the fluorescent dyes follows. A camera captures the light emitted and specific algorithms quantify the expression based on the probe intensity (Augenlicht and Kobrin, 1982).

Throughout the years more specialised microarray platforms were designed to tackle specific tasks, including exon arrays for splice junction detection and quantification (Clark et al., 2002), and genomic tiling arrays for higher genome resolution (Bertone et al., 2004; Cheng et al., 2005; Yamada et al., 2003). Microarrays are straightforward to use, rapid and cost-effective, but have several limitations, such as low reproducibility, high background and cross-hybridisation levels, limited dynamic range and intensity saturation points, they measure relative transcript abundance and are problematic for repeated sequences (Shendure, 2008).

Microarrays, in addition, require prior knowledge of the sequences, which is not always available, especially for non-model organisms. To overcome the limitations posed by microarrays, tag-based techniques were developed (reviewed in (Harbers and Carninci, 2005; Ruan et al., 2004)), such as serial analysis of gene expression (SAGE) (Velculescu et al., 1995), cap analysis of gene expression (CAGE) (Shiraki et al., 2003) and massive parallel signature sequencing (MPSS) (Brenner et al., 2000a,b). However, these approaches are only based on a short signature sequence, not allowing for transcript isoform discovery.

Soon after the development of next-generation sequencing platforms, these were introduced in the field of transcriptomics. Next-generation sequencing replaced previous technologies in

identifying and quantifying transcripts at genome-wide level (Morin et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm et al., 2008). This application is termed RNA-seq (reviewed in (Costa et al., 2010; Wang et al., 2009)).

Cell lysis and RNA extraction are the initial steps of a RNA-seq experiment. The majority of the total RNA is ribosomal RNA (rRNA) (Lindberg and Lundeberg, 2010), and unless a project focuses on studying rRNAs, library preparation protocols include a rRNA removel step of sort. Otherwise the sequencing pool will be dominated by these molecules. There are two ways to remove rRNA: either by enriching for poly(A)$^+$ tails or by depleting rRNA from the total RNA. The former technique captures mostly mature RNA molecules (mRNAs), while the latter one captures the whole transcriptome including non-coding (ncRNA), micro (miRNA) and premature RNAs (pre-mRNA).

Following RNA extraction, single strand RNA molecules are converted into double stranded cDNA(ds cDNA), because current sequencing technologies are unable to sequence RNA directly. For short RNA molecules following ds cDNA synthesis library preparation includes only the adaptor ligation step described above. However, longer molecules require to be fragmented. In this case a ds CDNA library is constructed either by fragmenting ds cDNA using DNAse1 treatment or sonication, or fragmenting RNA by RNA hydrolysis or nebulisation (detailed protocols for these two techniques can be found in (Costa et al., 2010)). Each of these techniques come with different biases in gene coverage with the former one introducing a bias towards the 3′ end of the transcripts, and the latter one towards the very 5′ and 3′ ends (Wang et al., 2009). Another consideration during library preparation is whether to maintain the information about the orientation of transcripts by the use of strand-specific protocols (Cloonan et al., 2008; Lister et al.,

2008). Finally, platform-specific adaptors are ligated onto the cDNA fragments as a last step prior to sequencing.

Bioinformatic analysis of RNA-seq data starts with aligning the reads to the reference sequence, assuming a reference genome or transcriptome is available. Reads can be aligned either to the reference transcriptome using an unspliced aligner, or the reference genome using a splice-aware aligner. The advantage of aligning to the reference genome is the identification of novel features. Engstrom et al. (2013) reported major performance differences among eleven splice-aware aligners based on an assessment of their alignment yield, basewise accuracy, mismatches and exon junction discovery. Based on the benchmarking results the authors concluded that four (GSNAP (Wu and Nacu, 2010), GSTRUCT (unpublished), MapSplice (Wang et al., 2010) and STAR (Dobin et al., 2013)) out of the eleven aligners perform better, with STAR being the fastest among all.

Post-alignment analysis includes the quantification of gene or transcript isoform expression. The RNA-seq data are 'digital' and the quantification of expression is based on the one-to-one relationship between RNA fragments and reads. The quantification process requires a normalisation step that takes into consideration the length of the gene or transcript and the total number of reads mapped. Expression is reported in Fragments Per Kilobase of transcript per Million mapped reads (FPKMs). Downstream analysis can include guided assembly of the transcriptome, where the assembly uses any existing transcript annotation as a guide (Trapnell et al., 2010b) and differential expression analysis for the identification of statistically significant differences in expression levels between different conditions (Anders et al., 2013; Robinson et al., 2010; Trapnell et al., 2012).

RNA-seq is a powerful, multi-purpose technique for the quantification of expression levels, allele-specific expression, iden-

tification of differential splicing, RNA-editing and gene fusions. It provides an estimation of absolute expression levels and a large dynamic range. Both library preparation and bioinformatic analysis steps are however undergoing constant modifications due to the many challenges posed by this complex technique. In addition to the biases and errors of next-generation sequencing described before, RNA-seq introduces more during the library preparation step. Several reactions cause sample loss. Reverse transcription for the cDNA generation may introduce errors. Certain biases are introduced by the use of random hexamers and the fragmentation step as described above.

1.3.4 *Integrative approaches enabled by next-generation sequencing*

The plethora of available next-generation sequencing applications, the increasing accessibility because of cost decrease and of course the availability of public data sets are crucial factors that enable researchers to gain biological insights at unprecedented depth. To this end, the integrative analysis of next-generation sequencing data is of great importance. Unique but complementary data sets can address long-standing and complex biological questions (reviewed in (Hawkins et al., 2010)). Large combinatorial genomics projects include the ENCODE (ENCODE Project Consortium, 2012) and the Cancer Genome Atlas (McLendon et al., 2008).

A lower-scale example of integrative analysis is exome sequencing coupled with RNA sequencing. Exome sequencing is a cost-effective alternative to whole-genome sequencing that enables the identification of variants within the coding regions of the genome and their flanking regions (Bamshad et al., 2011). It is an application widely used to identify causative single nucleotide polymorphisms (SNPs) in the coding sequences. To distinguish SNPs that may have an effect on the transcript and in turn on the protein generated, it is essential to identify those

that are located on regions transcribed in related cells using RNA-seq data. During my PhD I was involved in three collaborative projects aiming at identifying causative genes for two platelet-related diseases; the grey-platelet syndrome (GPS) (Albers et al., 2011) and the thrombocytopenia with absent radius (TAR) syndrome (Albers et al., 2012), and one identifying the underlying gene of the Vel blood group (Cvejic et al., 2013).

Part II

DATA ANALYSIS AND RESULTS

2

TRANSCRIPTIONAL LANDSCAPE OF MEGAKARYOCYTES AND THE MEIS1 TRANSCRIPTION FACTOR

2.1 RNA SEQUENCING OF MEGAKARYOCYTES

2.1.1 *Previous studies of gene expression profiles in blood cell types*

Previous studies of the gene expression profile of megakaryocytes were based on microarray platforms. These include the comparison between normal and disease state megakaryocytes (Tenedini et al., 2004), *in vitro* derived megakaryocytes and uncultured progenitors (Shim et al., 2004) or other blood cell types (Balduini et al., 2006; Fuhrken et al., 2008; Kim et al., 2002; Lim et al., 2008; Macaulay et al., 2007). In addition to the above studies that were conducted at single time points, human megakaryocytes have also been profiled at different stages of maturation in order to identify sets of genes that are continuously up- or down-regulated throughout megakaryopoiesis (Raslova et al., 2007). Therefore, our current knowledge of gene expression in megakaryocytes is relative to other blood cell types.

Following these smaller scale studies, several groups have focused on larger experiments with the aim of characterising the expression profiles of several distinct blood cell types. These gene expression studies generated large sets of data that serve in turn as useful resources for the scientific community. The first dataset, published in 2009 and called HaemAtlas hereafter, is a study performed in Cambridge that focused on identifying cell-type-specific genes of eight mature blood cell types using Il-

lumina bead arrays (Watkins et al., 2009). In 2011, Novershtern et al. (2011) presented the largest gene expression dataset of haematopoietic cells, referred to as DMAP dataset hereafter. A total of 38 purified populations of blood cells were analysed, expanding previous knowledge to include rare blood progenitor cells (Novershtern et al., 2011).

More recently RNA sequencing has also been used for the characterisation of the transcriptome of blood cells. Specifically in platelets, in 2011, Rowley et al. (2011) sequenced poly(A)$^+$ selected RNA from primary human and mouse platelets to identify differences in the transcriptome of the two species. Two years later, Bray et al. (2013) used ribosome-depleted RNA to generate the full transcriptome of human platelets. To date, the complete transcriptional landscape of the megakaryocyte lineage has not been resolved.

In this chapter, I analyse RNA-seq data from human megakaryocytes and integrate these data with ChIP-seq data of known regulators of the blood system, aiming to study certain aspects of the regulation of transcription in the platelet precursors.

2.1.2 *Quality control of megakaryocyte RNA-seq data*

Megakaryocytes were produced *in vitro* by culture from cord blood-derived CD34$^+$ haematopoietic stem cells. The cell population at the completion of the culture contained 70-90% of cells with a megakaryocytic phenotype, with the majority being CD41$^+$CD42b$^+$CD34$^-$. Libraries of non-strand specific poly(A)$^+$ selected RNA were sequenced on the Illumina platform, generating paired-end 76bp reads (see section 2.4.1 for more details). The cell culture and the RNA-seq library were prepared by Dr Pete Smethurst and Dr Katrin Voss, in Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge.

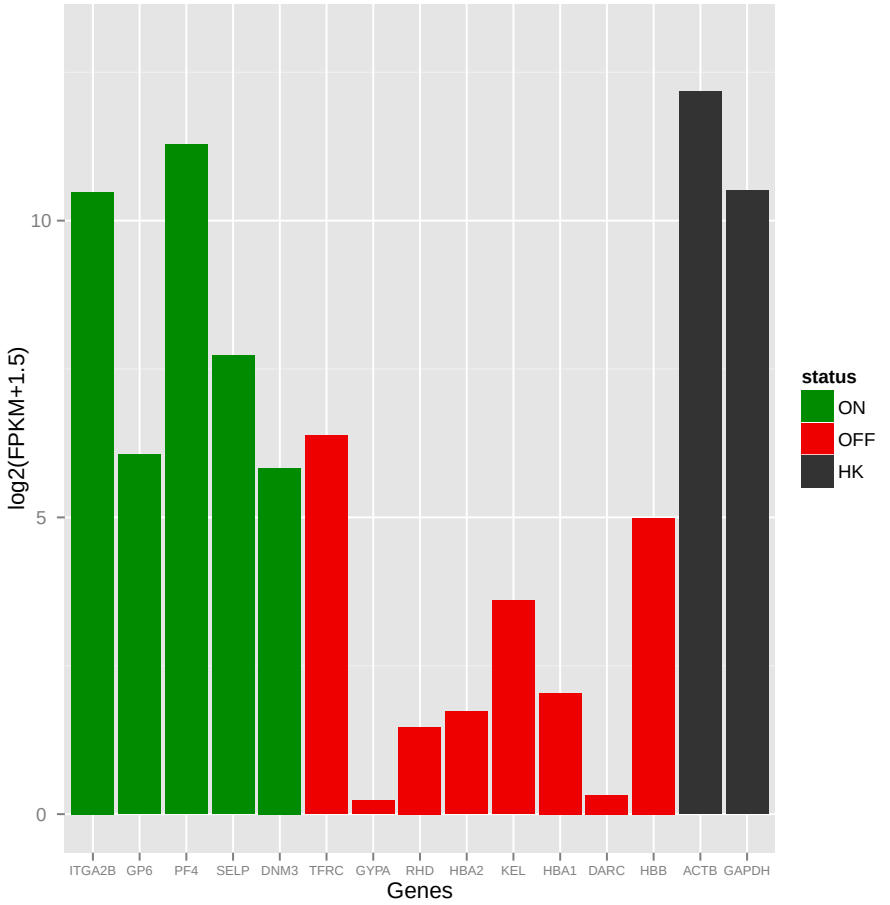


Figure 3: Gene expression of a set of marker genes in our megakaryocyte RNA-seq data. The selected marker genes are: ITGA2B: Integrin alpha-IIb, GP6: Platelet glycoprotein VI, PF4: Platelet factor IV, SELP: P-selectin, DNM3: Dynamin 3, TFRC: Transferrin receptor, GYPA: Glycophorin-A, RHD: Blood group Rh(D) polypeptide, HBA1/2: Haemoglobin subunit alpha, KEL: Kell blood group glycoprotein, DARC: Duffy antigen/chemokine receptor, ACTB: Actin cytoplasmic 1, GAPDH: Glyceraldehyde-3-phosphate dehydrogenase.
These markers that are expected to be expressed in megakaryocytes (green - ON) show high levels of transcription. The set of negative control genes (red - OFF) are expected to be either not transcribed or transcribed at lowel levels in megakaryocytes. Surprisingly, some of these that are characteristic of alternate lineages (e.g. TFRC, which is traditionally associated with nucleated red blood cells) are present in our dataset as well. The two housekeeping genes (black - HK) are also highly transcribed as expected.

To assess the quality of our megakaryocyte RNA-seq dataset, I examined the expression levels of selected lineage-specific erythroid and megakaryocytic genes. Three different gene sets were used. The first comprises genes that are highly and uniquely expressed in megakaryocytes (ITGA2B, SELP, DNM3, GP6 and

PF4) (Gewirtz et al., 1989; Tomer, 2004a), the second includes genes that encode blood group antigens and haemoglobins as negative controls (GYPA, KEL, RHD, DARC, HBAA1, HBAA2, HBB and TFRC) (Anstee, 1995; Tanner, 1993), and the third set contains two ubiquitously expressed genes (ACTB and GAPDH).

In our dataset the megakaryocyte-specific genes are all highly expressed. Unexpectedly, I also observe transcription of some of the negative control genes (see Figure 3). For example, the level of transcription of haemoglobin beta (HBB) is similar to that of dynamin 3 (DNM3), which has been associated with a significant role in late megakaryopoiesis (Nürnberg et al., 2012). There are two possible explanations for this:

1. A contamination of the cell population with erythroid cells. Even less than 1% of red cells could generate a high signal for haemoglobins or other erythroid cell surface proteins, since these are highly expressed in erythroblasts.

2. A significant overlap of the transcriptional programs between megakaryocytes and erythroblasts. However, similar mRNA expression does not always guarantee similar protein expression, or in this case cell surface expression of the antigens in megakaryocytes.

Given that the cell population was 70-90% megakaryocytic, it is highly likely that our population was contaminated with early erythrocytes. To discriminate between these two hypotheses, I also examined the expression level of the transferrin receptor (TFRC or CD71), which is necessary for development of erythrocytes (Levy et al., 1999). TFRC is also highly expressed in our dataset, supporting the contamination explanation. However, this does not exclude the possibility that the gene transcriptional profiles of megakaryocytes and erythrocytes might overlap substantially. To test this hypothesis, one needs to compare RNA-seq data from megakaryocytes and erythrocytes. In

case the expression level of the haemoglobins or other red blood specific genes is similar in the two cell types, then we can conclude that there has been some contamination in our cell population, whereas if the expression level is considerably higher in erythroblasts, then the two cell types exhibit substantial overlap in their transcriptional machinery. These datasets will be available later through the BLUEPRINT consortium (see Chapter 3) and a comparison may then be carried out.

2.1.3 *Comparison of megakaryocyte RNA-seq with microarray datasets*

To test how comparable our megakaryocyte RNA-seq data are to publicly available microarray datasets, I compared gene expression levels between the two studies on a genome-wide scale. To perform this comparison, I first used the HaemAtlas dataset (Watkins et al., 2009), as the megakaryocyte differentiation protocol was identical in both projects, minimising any technical variability due to cell derivation. The Illumina microarray probes were matched to Entrez gene IDs according to the re-annotation of the microarray platform (Barbosa-Morais et al., 2010).

The two different measures of expression in megakaryocytes are positively correlated with a squared Spearman correlation (ρ^2) of 0.637. Nevertheless, some genes show highly inconsistent expression between the HaemAtlas and the RNA-seq experiments (see Figure 4). A manual inspection of some of the disconcordant results between the two platforms revealed the following reasons for this discrepancy:

1. The gene is targeted by more than one microarray probe, and the intensity of those probes vary substantially. In this case, capturing only the maximum value may not be representative. Repeating the comparison using the average probe signals did not improve the correlation observed

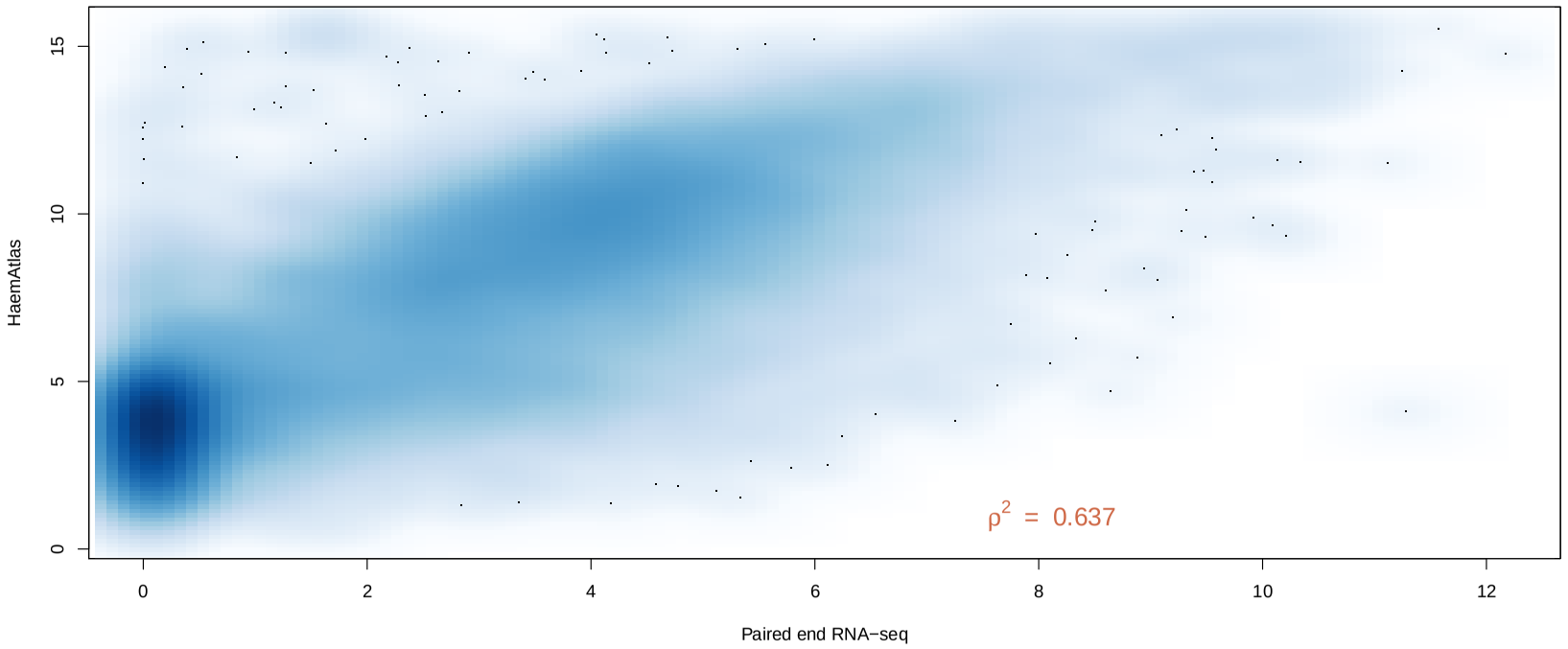


Figure 4: Comparing gene expression levels in megakaryocytes obtained from RNA sequencing and microarrays. The $\log_2$(FPKM + 1) value for each gene (x axis) is plotted against the $\log_2$ transformed normalised probe intensity from the HaemAtlas (Watkins et al., 2009) dataset (y axis). The added unit to the FPKM values was used to avoid negative $\log_2$ values given that the FPKM values can be equal or greater than 0. For genes targeted by more than one probe on the Illumina array, I used the maximum probe intensity reported among the probes.

(data not shown). While this approach amended some of the observed inconsistencies, it also created new ones, reflecting discrepancies inherent to the microarray design that cannot be completely resolved.

2. The probe on the microarray is not correctly targeting the gene to which it is assigned. For example, the probe might target a low-complexity region, an intron or matches to the reverse strand of the transcript; such cross-hybridisation targets confound accurate expression level estimates.

In both cases, the inconsistencies can be attributed to the limited sampling of the gene body intrinsic to the Illumina microarray platform.

One of the megakaryocytic specific genes, platelet factor IV (PF4, see Section 2.1.2), is an example of inconsistent gene expression between the two datasets. In this case, the microarray probe is targeting a part of the 3' UTR that is expressed at much lower level than the rest of the gene.

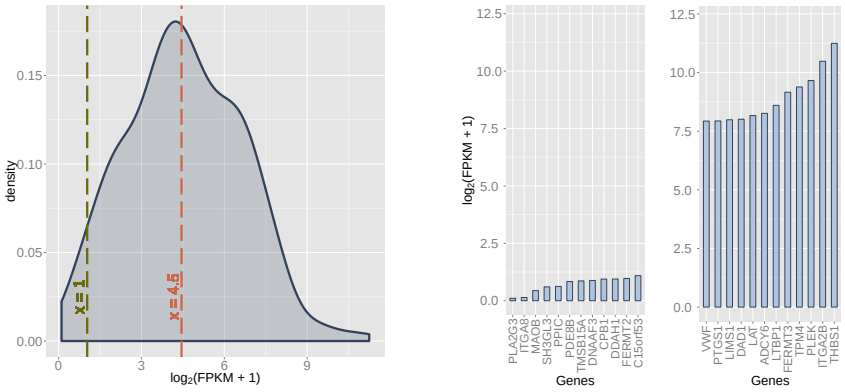
I followed the same approach to compare the megakaryocyte RNA-seq data to the other large haematopoietic study, DMAP, by Novershtern et al. (2011). The two datasets correlate poorly with a squared Spearman correlation (ρ^2) of 0.37. Within the set of highly divergent genes, those such glycoprotein VI (GPVI) and von Willebrand Factor (VWF) are highly expressed in the RNA-seq data, whereas GYPA is highly expressed in the microarray data. Given that GPVI is over-expressed in more mature megakaryocytes (Lagrue-Lak-Hal et al., 2001), the above findings could suggest that the two cell populations profiled are at different stages of megakaryocytic maturation. Similarly, VWF is expressed in early ploidy megakaryocytes (Tomer, 2004b). Indeed, in the DMAP dataset, the cell surface markers used for flow cytometric selection of megakaryocytes are $CD34^-$ $CD41^+$ $CD61^+$ $CD45^-$, excluding CD42 which appears later during megakaryocytic

maturation than CD41 and CD61 (Lepage et al., 2000). Lepage et al. also reported that cell cultures exclusively expressing $CD41^+$ $CD61^+$ generate more primitive erythroid progenitor cells than $CD42^+$ $CD41^+$ $CD61^+$ cultures. This potentially explains the high expression of GYPA (an erythrocytic specific gene, see Section 2.1.2) in the DMAP dataset.

In summary, using the HaemAtlas data, the sequencing and array data on megakaryocytes are highly correlated. This is consistent with previous comparisons between the two different technologies, which reported similar findings (Marioni et al., 2008). In the case of the DMAP dataset the poor correlation can be explained by the differences in the maturation stage of the megakaryocytes.

The HaemAtlas project also reported a set of 256 genes that were identified as megakaryocyte specific. These were determined by comparing megakaryocytes to all other mature blood cell types included in the study. Genes were reported only if they were consistently up-regulated in megakaryocytes with at least a 2-fold change in expression. To further compare our RNA-seq data with the HaemAtlas megakaryocyte data, we examined the gene expression levels of these megakaryocyte specific genes.

Most of the 256 genes are highly expressed in the RNA-seq dataset with a median $\log_2(\text{FPKM} + 1)$ value of 4.5 (see Figure 5(a)). However, twelve out of 256 genes have low expression values ($0 \leqslant \text{FPKM} \leqslant 1$) suggesting that these genes are lowly or not expressed at all in megakaryocytes. The individual expression levels of the top and bottom genes of this set can be seen on Figure 5(b). These results highlight the limitations of the relative measurements made using microarrays and show how RNA sequencing can be used to complement, improve and expand on the results obtained by microarray experiments.



(a) (b)

Figure 5: Gene expression of 256 megakaryocyte unique genes (Watkins et al., 2009) in the megakaryocyte RNA-seq dataset.
(a) The distribution of gene expression values of the 256 megakaryocyte specific genes verifies that the majority is expressed in our dataset (min = 0.2, median = 4.5, max = 11.3). A threshold of 1 FPKM is often used to distinguish between expressed and not expressed genes.
(b) Barplot of the expression levels of the bottom (left panel) and top (right panel) megakaryocyte unique genes as these were ranked based on their FPKM values. The genes depicted in the left panel have a low FPKM value (< 1) despite having been defined as specific to megakaryocytes.

2.1.4 *Splice junction analysis of megakaryocyte RNA-seq dataset*

Alternative splicing of genes gives rise to multiple mRNA products from the same genomic locus. It is one of the biological mechanisms used to create high protein complexity (Lopez, 1998). As a first step towards determining the set of transcript particular to megakaryocytes, I identified the splice junctions in our RNA-seq dataset and then compared them to annonated ones from public databases.

To identify splice junctions, I aligned the RNA-seq reads to the human genome using a splice-aware aligner. I then extracted the splicing information based on spliced read alignments. I only considered unique alignments where at least 10bp of the read had been aligned on either side of the splice junction. Engstrom et al. (2013) showed that splice sites with low read support, tend to be false positives. Therefore, to reduce the number of false positive splice junctions, the above generated set was

Class	#	%
Annotated	165,740	99.02
Novel	439	0.26
Partially novel	736	0.44
Novel within gene	416	0.25
Novel between genes	52	0.03

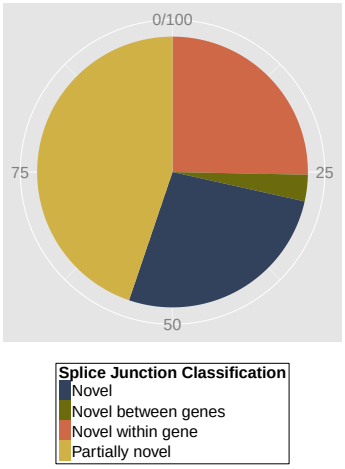


Figure 6: The Table shows the number of splice junctions defined by the spliced reads of the megakaryocyte RNA-seq dataset. In total there are 167,435 splice sites that are spanned by at least 10 reads. The majority of these splice junctions is annotated in Ensembl v70. The novel splice junctions can be classified into four categories: *novel*: none of the ends of the splice junction is annotated, *partially novel*: one of the ends of the splice junction is annotated, *novel within gene*: the splice junction is novel, but the ends are annotated within the same gene, and *novel between genes*: the splice junction is novel, but the ends are annotated in different genes.

further filtered to retain those that were supported by a minimum of ten reads. In total, there were 167,338 splice junctions that fulfilled the above criteria. To determine which of those are novel, I compared them to a set of annotated splice junctions from Ensembl v70. The majority of the splice junctions are already annotated (see Figure 6).

I then classified the novel splicing events based on whether either of the ends of the splice junction were already annotated (partially novel) or not (novel). In case both ends were annotated, I examined whether the ends originated from the same gene (novel within gene), or from different genes (novel between genes). Only few of these belonged to the latter category, while the majority of the novel splice junctions are partially novel and the rest are equally distributed among novel and novel within gene (see Figure 6).

Transcription Factor	# peaks
MEIS1	26,726
GATA1	4,457
GATA2	2,039
FLI1	8,260
RUNX1	6,883
SCL	3,063

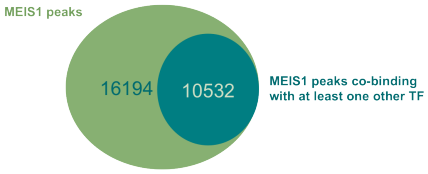


Figure 7: Number of peaks identified per transcription factor in human megakaryocytes. Venn diagram shows the number of MEIS1 peaks that co-localise with at least one other transcription factor from: FLI1, GATA1, GATA2, RUNX1 and SCL.

To conclude, this catalogue of splicing events is the initial step towards elucidating the set of transcripts present in megakaryocytes. In Section 2.3.2, I will focus on the novel splice junctions in order to study novel transcription start sites in megakaryocytes and to examine if these can be linked to the transcription factor MEIS1.

2.2 MEIS1 CHIP SEQUENCING IN MEGAKARYOCYTES

As described in more detail in Section 1.2.4.3, MEIS1 has a dual role in haematopoiesis: promoting cell proliferation of heamatopoietic stem cells and maintaining normal megakaryopoiesis (Cai et al., 2012). Megakaryopoiesis and subsequent platelet formation is blocked in MEIS1 deficient mice (Azcoitia et al., 2005; Hisa et al., 2004) and zebrafish (Cvejic et al., 2011). The severe phenotype observed led us to study the binding activity of MEIS1 in human megakaryocytes, with the aim to determine its regulatory targets.

2.2.1 *MEIS1 binding profile in human megakaryocytes*

To identify potential MEIS1 binding sites, Dr Sylvia Nürnberg (a PhD student at the time in Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge) per-

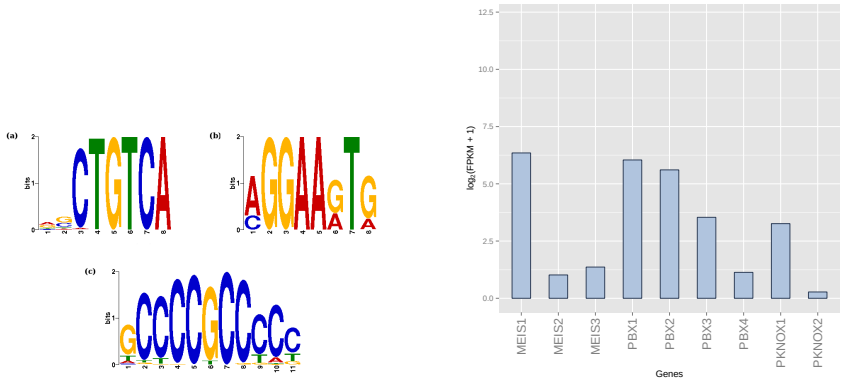
formed chromatin immunoprecipitation using a MEIS1 antibody in human megakaryocytes followed by sequencing. Subsequent bioinformatic analysis of the MEIS1 ChIP-seq data (see Section 2.4.5) revealed 26,726 peaks of signal enrichment (see Figure 7). To examine the relative distribution of MEIS1 binding sites in promoters, intergenic or intragenic regions, I determined the proportion of MEIS1 peaks in each category of genomic loci. A similar number of MEIS1 peaks are located in promoter regions, introns and intergenic regions, and only 2% of peaks are found in exons (see Figure 9).

Of the 244 megakaryocytic specific genes described in Section 2.1.3 (256 in total minus 12 that have low expression in the RNA-seq dataset), 192 are bound by MEIS1 in their promoter region. To further test for enriched biological functions of the MEIS1 bound genes, I then examined the set of genes with a MEIS1 promoter occupancy for over-represented gene ontology terms and pathways. This gene set is highly enriched for platelet specific terms and pathways (see Table 4), supporting the significance of MEIS1 in megakaryopoiesis and platelet function.

Table 4: Enriched Gene Ontology terms, KEGG and Reactome pathways among the genes bound by MEIS1 in the promoter region. Only the top 2 terms for each category are being shown. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg procedure.

Category	Term	*P*-value	q-value
Biological Process	platelet activation	7.95e-18	1.78e-14
Biological Process	blood coagulation	1.07e-16	7.39e-14
Cellular Compartment	platelet alpha granule	2.97e-08	7.93e-06
Cellular Compartment	actin filament bundle	2.89e-05	7.01e-05
KEGG	ECM-receptor interaction	2.18e-06	2.05e-04
KEGG	Focal adhesion	5.11e-06	2.40e-04
Reactome	Hemostasis	1.41e-18	2.61e-16
Reactome	Platelet activation, signaling and aggregation	9.96e-17	9.17e-15

To examine whether a consensus sequence binding motif can be detected among the MEIS1 peaks, I performed *de novo*



(a) (b)

Figure 8: Conserved motifs within MEIS1 peaks and expression of TALE homeodomain genes in megakaryocytes
(a) The conserved motifs within MEIS1 peaks represent sequence-specific binding of transcription factor families that are known regulators of megakaryocytic gene expression; (a) TALE homeodomain, (b) ETS, and (c) SP1 - Zinc finger proteins. *De novo* motif discovery was performed using MEME on 100 bp long sequences around the summit of the peaks. The search was restricted to 3 motifs between 6 and 30 bp in length.
(b) Expression levels of TALE homeodomain genes in megakaryocytes; MEIS1: Homeobox protein Meis1, MEIS2: Homeobox protein Meis2, MEIS3: Homeobox protein Meis3, PBX1: Pre-B-cell leukemia transcription factor 1, PBX2: Pre-B-cell leukemia transcription factor 2, PBX3: Pre-B-cell leukemia transcription factor 3, PBX4: Pre-B-cell leukemia transcription factor 4, PKNOX1: Homeobox protein PKNOX1, and PKNOX2: Homeobox protein PKNOX2.

motif discovery analysis, applying MEME (Bailey and Elkan, 1994) to the set of 100 bp sequences flanking each peak summit. Three motifs were found to be conserved among the MEIS1 peaks, summarised in Figure 8. To test for similarity between these motifs and known transcription factor binding sequences, I used Tomtom (Gupta et al., 2007) to search against the JASPAR (Bryne et al., 2008) and Uniprobe (Robasky and Bulyk, 2011) databases. The three different motifs are representatives of: (a) members of the Three Amino acid Loop Extension (TALE) Homeodomain transcription factor subfamily, (b) members of the E-Twenty-Six (ETS) family, and (c) the SP1 transcription factor. In more detail, motif (a) represents the sequence-specific binding of several TALE homeodomain proteins, the subfamily of transcription factors to which MEIS1 belongs. Motif (b) resembles those associated with members of the ETS transcrip-

tion factor family, whose members have been shown to be key regulators of haematopoiesis (Fisher and Scott, 1998). Finally, motif (c) is similar to the motif for the transcription factor SP1, which has been implicated in the regulation of gene expression in megakaryocytes (Gannon and Kinsella, 2008).

Motif (a), representing members of the TALE Homeodomain family, is similar to the hexameric DNA sequence that MEIS and PKNOX proteins preferentially bind to (Berthelsen et al., 1998; Chang et al., 1997; Ferretti et al., 2000; Shen et al., 1997) according to *in vitro* selection of target sequences. Penkov et al. (2013), however, recently showed that this motif in mice embryos is found in genomic regions bound by MEIS factors, either alone or dimerised with PBX family members.

In humans, there are four PBX (PBX1, PBX2, PBX3 and PBX4) and two PKNOX (PKNOX1 and PKNOX2) genes. According to our RNA-seq data, MEIS1, PBX1 and PBX2 are the most highly expressed genes among the TALE homeodomain members. PBX3 and PKNOX1 are also expressed, but at lower levels (see Figure 8(b)). Multiple TALE homeodomains are expressed in megakaryocytes and each binds similar DNA sequences. Given that they are also frequently found to form dimers or trimers with HOX genes, further ChIP-seq experiments will be needed to infer how the TALE homeodomain members (MEIS, PBX and PKNOX) orchestrate transcription in megakaryocytes.

2.2.2 *Combining MEIS1 binding sites with publicly available ChIP-seq datasets*

Transcriptional regulation is a complex process that involves extensive cooperation of multiple transcription factors (reviewed in Spitz and Furlong (2012)). The motifs of different transcription factors within MEIS1 peaks that I identified in the previous section suggest that MEIS1 can act in tandem with other

regulatory factors in megakaryocytes. To identify potential co-occupancy of MEIS1 with other key regulators of haematopoiesis (previously discussed in Chapter 1), I integrated our MEIS1 dataset with ChIP-seq data of five additional transcription factors profiled in human megakaryocytes (Tijssen et al., 2011). The set of transcription factors included GATA1, GATA2, RUNX1, FLI1 and TAL1/SCL. The number of peaks identified for each transcription factor is summarised in Figure 7.

Intersection of the MEIS1 peaks with those identified for all the other five transcription factors reveal a substantial overlap between MEIS1 and at least one other transcription factor. As shown in Figure 7, 10,532 MEIS1 peaks (39.4%) co-localise within a window of 100 bp with another transcription factor. The much lower number of peaks identified for the transcription factor set from Tijssen et al. (2011) could imply that not all transcription factor binding sites were identified and therefore the numbers of MEIS1 co-localising binding sites may be much higher than observed.

The proximity of transcription factor binding sites with respect to genes can give some indication of the regulatory processes the binding event could be involved in. When binding to the proximal promoter, transcription factors typically directly regulate the expression level of the gene in question, often in conjunction with other factors. Binding in the gene body can regulate the level of expression, as well as affecting the transcriptional and splicing machinery. Intergenic binding sites can correspond to enhancers, or represent non-functional binding (Maston et al., 2006a).

Of the five transcription factors, FLI1 and RUNX1 are predominantly recruited to promoter regions. In the case of RUNX1 the percentage of peaks within promoter regions reaches 50%. In contrast, GATA1, GATA2 and SCL have a stronger binding

preference towards introns and intergenic regions; with GATA2 mostly situated intergenically, and SCL in introns. Compared to these two trends, MEIS1 lies somewhere in the middle with peaks evenly distributed among promoter regions, introns and intergenic regions (see Figure 9).

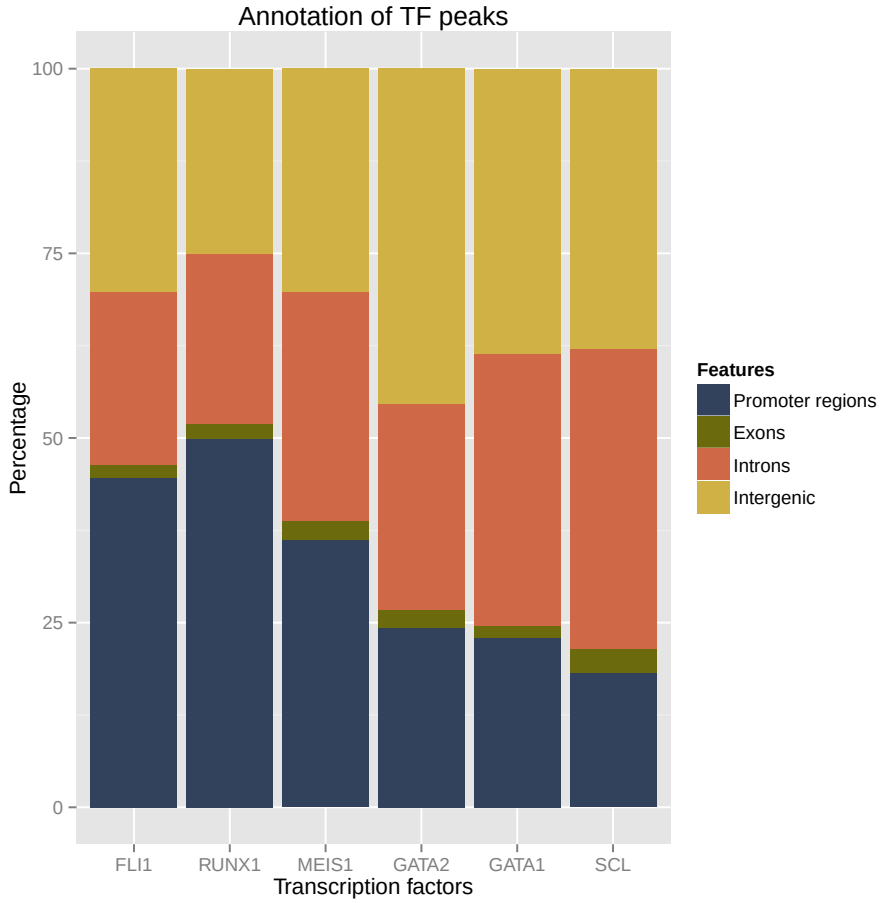


Figure 9: Annotation of FLI1, GATA1/2, MEIS1, RUNX1 and SCL peaks in human megakaryocytes. RUNX1 anf FLI1 show strong preference towards promoter regions, while GATA1/GATA2 and SCL are predominantly located in intergenic or intronic regions. MEIS1 peaks are equally distributed among promoters regions, introns and intergenic regions. All six transcription factors show low numbers of peaks located in exons. The peaks were assigned to promoter regions (from 5000bp upstream the transcription start site to 500 bp downstream), collapsed exons, introns and intergenic regions. Peaks were classified in these four groups based on the Ensembl v70 annotation.

I then examined if the binding of MEIS1 ia altered depending on if these sites co-localise with another transcription factor

Region	# MEIS1 only peaks	# co-binding peaks	Total
Promoters	4,821	4,869	9,690
Introns	5,579	2,707	8,286
Exons	439	227	666
Intergenic	5,355	2,729	8,084
Total	16,194	10,532	26,726

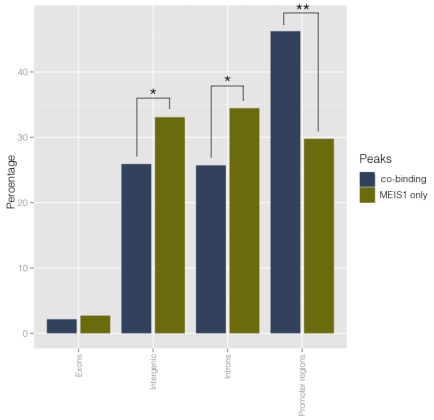


Figure 10: Annotation of two groups of MEIS1 binding sites; those that co-localise with at least another transcription factor and those that do not. MEIS1 peaks co-localising with other transcription factors are more often found in promoter regions rather than MEIS1 peaks only (**P-value < $2e^{-30}$). A different trend is observed though in the intergenic and intronic regions, where MEIS1 peaks are not found to overlap with other transcription factors (*P-value < $2e^{-10}$).

binding site. The most striking change is to promoter regions, where MEIS1 peaks are more often found to co-localise with another transcription factor (see Figure 10). This finding is supported by the fact that promoters are crucial for gene regulation and usually bound by a large number of regulatory elements that tightly orchestrate gene expression (Farnham, 2009).

A different trend was observed for intronic and intergenic binding sites. Here, MEIS1 binds exclusively more often (see Figure 10). Binding to intergenic regions can be interpreted as either non-functional binding or preference of the transcription factor to enhancer regions. This is consistent with ChIP-seq experiments in mouse embryos. Penkov et al. (2013) reported that MEIS exclusive peaks are predominantly located in intragenic or intergenic regions in mouse embryos rather than promoters, suggesting that MEIS proteins show a preference for enhancers. For a number of intronic and intergenic MEIS1 peaks in megakaryocytes, we will see in Section 2.3.2 that MEIS1 binds in the vicinity of unannotated transcription start sites.

2.2.3 *Co-localising binding events between MEIS1 and FLI1, GATA1/2, RUNX1 and SCL*

Tijssen et al. (2011) reported a considerable over-representation of co-localisation of the five transcription factors, FLI1, GATA1, GATA2, RUNX1 and SCL, in megakaryocytes. To explore how often and in which combination of these five transcription factors MEIS1 co-localises, I split the overlapping peaks into subgroups based on the combination of transcription factors bound to each region.

High numbers of FLI1 and RUNX1 peaks, which either localise individually or together, overlap with MEIS1 peaks (see Table 5). Similarly, GATA1/2 and SCL co-localise with MEIS1 when no other transcription factor is bound within a 100bp window. However, such co-occupancy is observed at much lower numbers. This is potentially due to the overall lower numbers of peaks identified for these three transcription factors (see Figure 7).

Notably, 135 out of the 144 regions, where all five transcription factors co-localise, are also bound by MEIS1 as well. Although the absolute number is not as high as those highlighted in Table 5, genes bound by all six transcription factors show enrichment for a haematopoietic related KEGG pathway (see Table 6). Conversely, a similar gene ontology term, KEGG and Reactome pathway enrichment analysis for the genes bound by the ten most frequent combinations of the six factors did not reveal such enrichment. One example is the genes bound by both FLI1 and MEIS1, which is the most frequent combination of transcription factor co-localisation (see Table 6).

Peak annotation of the 10 most frequently observed combinations of co-localising transcription factors (highlighted in Table 5) suggests that these followed similar patterns to the

localisation observed for the individual transcription factors (see Figure 11). For example, FLI1 and RUNX1 peaks when co-localising with MEIS1 (either together or separately) retain a strong preference towards promoter regions. The only difference reported is for GATA2-MEIS1 co-localisation which is mainly found in promoter regions, despite the majority of GATA2 only peaks being located in intergenic regions. Gene ontology enrichment analysis of the genes bound by MEIS1 and GATA2 in the promoter regions did not reveal any platelet or megakaryocytic specific function.

Table 5: Combination of transcription factors co-localising with MEIS1. The top 10 most frequent combinations are highlighted in orange.

Transcription factor combination	**# peaks**
GATA1	775
GATA2	248
FLI1	2595
RUNX1	2432
SCL	416
GATA1 - GATA2	113
GATA1 - RUNX1	299
GATA1 - FLI1	175
GATA1 - SCL	260
GATA2 - RUNX1	36
GATA2 - FLI1	210
GATA2 - SCL	11
FLI1 - RUNX1	1349
FLI1 - SCL	252
RUNX1 - SCL	67
GATA1 - GATA2 - FLI1	58
GATA1 - GATA2 - RUNX1	53
GATA1 - GATA2 - SCL	50
GATA1 - FLI1 - RUNX1	191
GATA1 - FLI1 - SCL	223
GATA1 - RUNX1 - SCL	117
GATA2 - FLI1 - RUNX1	31

Continued on next page

Table 5 – *Continued from previous page*

Transcription factor combination	**# peaks**
GATA2 - FLI1 - SCL	14
GATA2-RUNX1-SCL	2
FLI1 - RUNX1 - SCL	122
GATA1 - GATA2 - FLI1 - RUNX1	77
GATA1 - GATA2 - FLI1 - SCL	52
GATA1 - GATA2 - RUNX1 - SCL	41
GATA1 - FLI1 - RUNX1 - SCL	221
GATA2 - FLI1 - RUNX1 - SCL	1
GATA1 - GATA2 - FLI1 - RUNX1 - SCL	135

Table 6: Enriched Gene Ontology terms, KEGG and Reactome pathways among the genes bound by MEIS1 and FLI1, or RUNX1, or all six transcription factors (MEIS1, FLI1, GATA1/2, RUNX1 and SCL). For categories with multiple terms, only the top 2 terms for each category are shown. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg procedure.

	Category	Term	*P*-value	q-value
FLI1-MEIS1	Biological Process	RNA processing	4.53e-21	2.31e-17
	Biological Process	mRNA metabolic process	7.95e-20	2.03e-16
	Cellular Compartment	intracellular	< 1.00e-60	< 1.00e-60
	Cellular Compartment	intracellular part	< 1.00e-60	< 1.00e-60
	KEGG	Spliceosome	2.24e-10	4.58e-08
	Reactome	Gene Expression	4.12e-40	2.48e-37
	Reactome	mRNA Processing	7.09e-14	1.07e-11
RUNX1-MEIS1	Biological Process	cellular metabolic process	4.87e-16	2.46e-12
	Biological Process	primary metabolic process	7.24e-14	1.42e-10
	Cellular Compartment	intracellular	4.10e-48	2.79e-45
	Cellular Compartment	intracellular part	3.79e-44	1.29e-41
	KEGG	Systemic lupus erythematosus	4.83e-14	9.90e-12
	Reactome	Amyloids	5.84e-23	1.78e-20
	Reactome	Meiotic Recombination	5.84e-23	1.78e-20
all	Reactome	Hemostasis	7.95e-05	0.003
	Reactome	Zinc influx into cells by the SLC39 gene family	8.85e-05	0.003
	Reactome	Zinc transporters	2.66e-04	0.005

2.2.4 *Enrichment of other transcription factor motifs within MEIS1 binding sites*

The low number of available ChIP-seq data sets in megakaryocytes limits analysis to identify transcription factors that might co-occupy genomic regions with MEIS1. Identification of

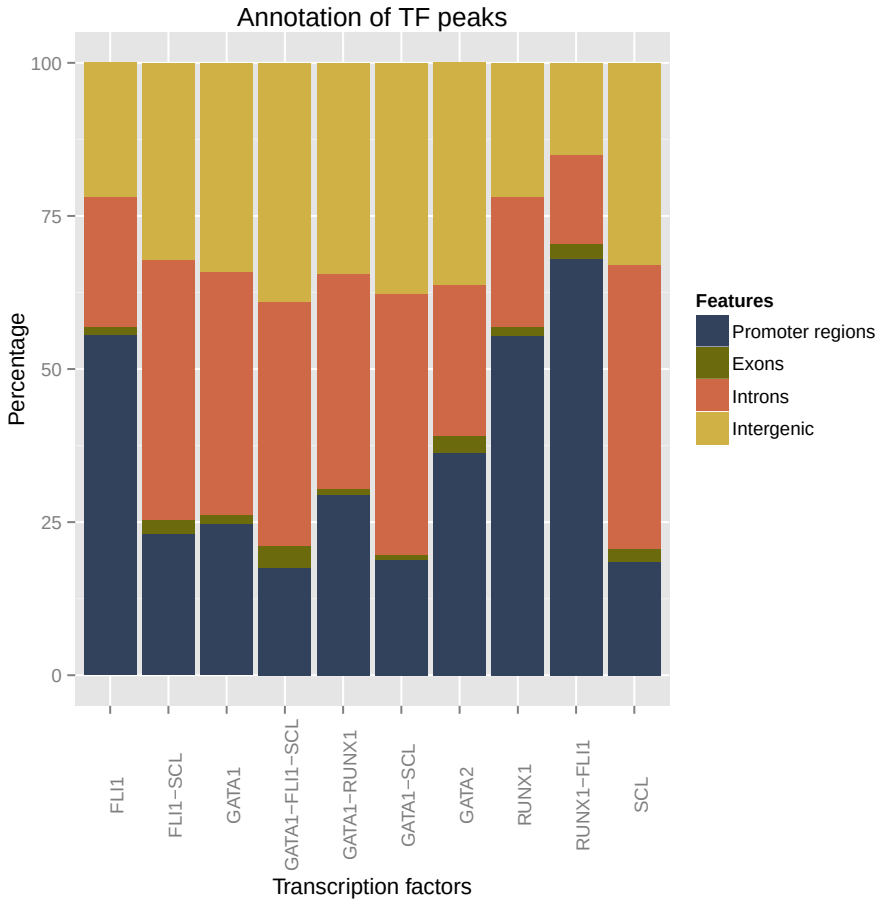


Figure 11: Annotation of ten most frequent co-localisations of MEIS1 peaks with the set of five transcription factors from Tijssen et al. (2011). GATA2 peaks that co-localise with MEIS1 are more frequently located on promoter regions, than GATA2 only peaks.

transcription factor binding sites can also be performed using strictly computational methods. These methods rely on position weight matrices that represent known DNA binding motifs for each transcription factor. However, genomic sequences that match these motifs are not necessarily functional within any cell-type and under any condition. Thus, computational analysis can be used as a discovery tool, but results need to be validated within the cell population of interest in order to identify functional binding sites. Based on a set of curated position weight matrices that model the sequence-specific binding of human transcription factors (Jolma et al., 2013), I examined

the top 1000 MEIS1 peaks (sorted by *P*-value) for enriched binding site motifs.

For this analysis transcription factors are grouped into families based on the type of DNA binding domains they share. Different members of the same transcription factor family bind nearly identical DNA sequences and often exhibit redundant functionality. Members of the TALE homeodomain are an example of this behaviour of closely related members of a transcription factor subfamily (see Section 2.2.1). Hence, any motif matches found for a transcription factor through the computational analysis in the MEIS1 peaks, may be used to infer results for other members of the transcription factor family. Taking advantage of the gene expression data, I filtered out transcription factors that are not expressed in megakaryocytes.

Consistent with the results of the *de novo* motif analysis in Section 2.2.1, there is a large fraction of members of the homeodomain and E-twenty-six transcription factor families with motifs that are enriched within the top 1000 MEIS1 peaks (see Table 7). In addition to the TALE homeodomains, a subfamily of homeodomains identified previously (see Section 2.2.1), two other homeodomain subfamily motifs are also enriched. These include the TGIF and DLX subfamilies, of which TGIF1/2 and DLX1, respectively, were enriched and expressed. Regarding the TGIF proteins, there is evidence in the literature that these co-regulate transcription with MEIS2, through opposing effects (Yang et al., 2000). In the case of DLX1, no evidence exists for co-regulation with MEIS proteins. It does, however, regulate multiple signals from TGF-beta superfamily members during blood production (Chiba et al., 2003).

Binding sites of the ETS family feature predominantly within MEIS1 peaks and most of the members are also expressed in megakaryocytes. Two members of the ETS family, FLI1 and

Table 7: Enriched human transcription factor families in top 1000 MEIS1 peaks. Genes are grouped per transcription factor family. Expressed transcription factors have a FPKM value greater than 1 and their model is enriched if the *P*-value is lower than 0.05.

TF family	Interpro Domain ID	# curated TFs	# TF with model	# expressed TFs	# enriched models	expressed TFs with enriched models
C2H2-type zinc-finger	IPR007087	613	48	518	2	SP4
Homeodomain	IPR001356	244	130	47	32	MEIS3, TGIF2, HOXB3, MEIS2, MEIS1, DLX1, PKNOX1, HOXB2, TGIF1
basic Helix Loop Helix (bHLH)	IPR011598	120	36	53	7	TFAP4, TFEB, BHLHE41, USF1, SREBF2
High Mobility Group (HMG)	IPR009071	59	14	33	0	-
basic leucine Zipper (bZIP)	IPR021802 - IPR011616 - IPR016743 - IPR004827 - IPR004827	58	23	46	5	XBP1, DBP, CREB3
Forkhead	IPR001766	49	16	12	1	-
Nuclear hormone receptor	IPR001628	45	23	19	0	-
E-Twenty-Six (ETS)	IPR000418	29	24	17	22	ELF4, ETV2, ERF, ELK3, ETV3, ELF1, ELK1, ETS1, ETV6, FLI1, GABPA, ERG, ELK4, ETV5
Runt box (RUNX)	IPR013524	3	2	3	2	RUNX2, RUNX3
Myeloblastosis (MYB)	IPR017877	23	2	22	0	-
GATA zinc finger	IPR000679	15	3	11	3	-
T-box	IPR001699	17	12	1	0	-
BED zinc finger	IPR003656	14	1	11	0	-
E2F/DP	IPR003316	11	6	10	0	-
CENPB	IPR004875	11	1	9	0	-
Interferon Regulatory Factor (IRF)	IPR001346	9	6	7	0	-
Rel Homology Domain (RHD)	IPR011539	10	4	10	0	-
SAND	IPR000770	8	1	6	1	GMEB2
Regulatory Factor X (RFX)	IPR003150	9	4	7	4	RFX3, RFX2, RFX5
Heat Shock Factor (HSF)	IPR000232	8	4	3	0	-
CP2	IPR007604	6	2	2	0	-
Activating Protein 2 (AP-2)	IPR013854	5	3	0	0	-
MADS-box	IPR002100	5	4	5	0	-
Nuclear Factor I (NFI)	IPR020604	4	3	4	1	NFIX
P53	IPR011615	3	1	1	0	-
GCM	IPR003902	2	2	0	0	-
NRF	IPR019525	1	1	1	0	-
PROX	IPR007738	2	1	0	0	-
TEA	IPR000818	4	3	1	1	-

GABPA, are already known regulators of late and early megakaryopoiesis, respectively (Pang et al., 2006). In the case of FLI1 the co-localisation with MEIS1 has been already described in megakaryocytes through analysis of ChIP-seq data (see Table 5). Other regulators of megakaryopoiesis from this transcription factor family include ETS1 and ERG. Over-expression of ETS1 promotes megakaryocytic differentiation to the erythroid lineage (Lulli et al., 2006), while ERG is required for both normal haematopoietic stem cell and megakaryocyte homeostasis (Kruse et al., 2009).

The integrative analysis of ChIP-seq data for the MEIS1, GATA1/2, FLI, RUNX1 and SCL transcription factors in Section 2.2.3, showed that MEIS1 frequently co-occupies genomic regions with one or more of the other five transcription factors. In this computational motif analysis FLI1 was enriched, however I also expected to find GATA1/2, RUNX1 and SCL enriched. None of these four factors were reported, due to the lack of a position weight matrix that models their sequence specificity. Nevertheless, other members of the transcription factor families to which they belong were identified.

SCL is a member of the basic helix loop helix (bHLH) family. Five bHLH family members were enriched and expressed in megakaryocytes (see Table 7), including USF1. USF1 expression has been studied within haematopoietic progenitor cells and was found to be induced by thrombopoietin, the cytokine that drives megakaryocytic differentiation (Kirito et al., 2003). Moreover, USF1 trans-activates platelet factor 4, a megakaryocytic specific gene, along with MEIS1 in rat megakaryocytes (Okada et al., 2004). Regarding GATA family members, the curated dataset of position weight matrices contains data for only GATA4/5/6. Although these three GATAs are all found enriched within MEIS1 peaks, none of them was reported because they are not ex-

pressed in megakaryocytes. The paralogues for RUNX1, RUNX2 and RUNX3, are both enriched and expressed (see Table 7).

The basic leucine zipper (bZIP) transcription factor family is also involved in regulation of megakaryopoiesis through Nfe2 and Mafg proteins. Mice deficient in either of these two proteins exhibit reduced platelet production, despite having a normal number of megakaryocytes (Shavit et al., 1998; Shivdasani et al., 1995b). Although these two proteins are included in the curated dataset of position weight matrices, neither was reported as enriched. Nonetheless, three other members of this family, XBP1, DBP, and CREB2, were identified. No evidence was found in the literature however, that links any of these transcription factors to megakaryopoiesis or platelet production.

Finally, among the enriched transcription factor families, there are also two, RFX and NFI, that have not been previously studied in the context of megakaryopoiesis. In the case of RFX members, Fuhrken et al. (2008) through microarray experiments and gene ontology enrichment analysis identified RFX5 as a potential regulator in megakaryocytes. There is no evidence in the literature that links NFI proteins to megakaryopoiesis. However, another member of the NFI transcription family has been shown to promote erythroid differentiation (Starnes et al., 2010).

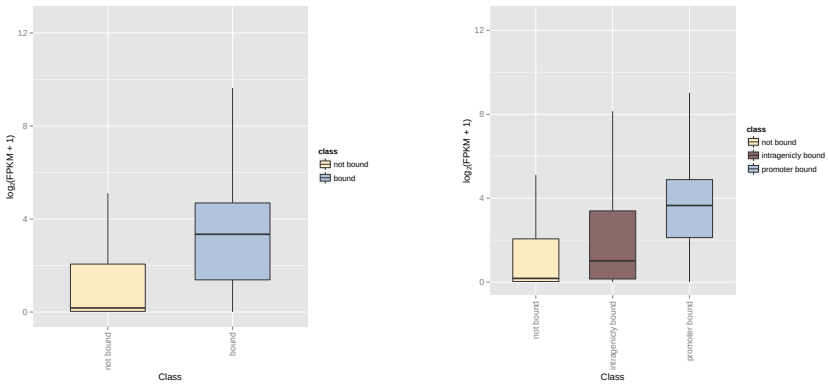
In summary, analysis of MEIS1 ChIP-seq data confirms MEIS1 as an important regulator of megakaryopoiesis, and shows its regulatory function is tightly intertwined with those of other key transcription factors of haematopoiesis. Using computational analysis to identify known transcription factor motifs present in the MEIS1 binding regions, and considering only genes that were determined to be expressed in megakaryocytes by RNA-seq, I identified a list of candidate factors that may interact with MEIS1 and participate in the regulation of megaka-

ryopoiesis. Further experiments are required to validate these candidates and elucidate their role in megakaryopoiesis. In Chapter 4, I investigate the function of NFI proteins in megakaryopoiesis.

2.3 INTEGRATION OF MEIS1 PEAKS WITH MEGAKARYOCYTIC GENE EXPRESSION

Analysis of the high-throughput sequencing techniques described above - ChIP-seq and RNA-seq - provides substantial understanding for functional genomic studies in megakaryocytes. Moreover, integrative analysis of these data can offer a deeper insight into dynamic genome function.

2.3.1 *MEIS1 and gene expression in megakaryocytes*



(a) (b)

Figure 12: Comparison of range of gene expression: (a) between genes bound and not bound by MEIS1 and (b) between genes with a MEIS1 peak in either promoter or intergenic regions.
(a) MEIS1 has a positive effect on gene expression with genes bound by MEIS1 showing a $\log_2$ transformed median gene expression of 3.34, compared to 0.18 for those not bound by MEIS1. (b) MEIS1 peaks located in promoter regions have a stronger effect on gene expression compared to those in intragenic regions. The first set has a a $\log_2$ transformed median expression value of 3.65, whereas the latter is 1.02

To study the effect of MEIS1 on gene expression in megakaryocytes, I initially compared the expression levels between

genes bound by MEIS1 and those not bound by it. MEIS1 has a positive effect on gene expression, as the genes bound by MEIS1 (median = 3.34) show higher expression values to the ones not bound by it (median = 0.18) (see Figure 12(a)). The location of the MEIS1 peak appears to be important for the effect on gene expression. The expression values of genes with a MEIS1 peak in their promoter regions are higher compared to those genes that have an intragenic MEIS1 peak (see Figure 12(b)). The latter gene set has in turn higher expression values than the genes not bound by MEIS1 at all.

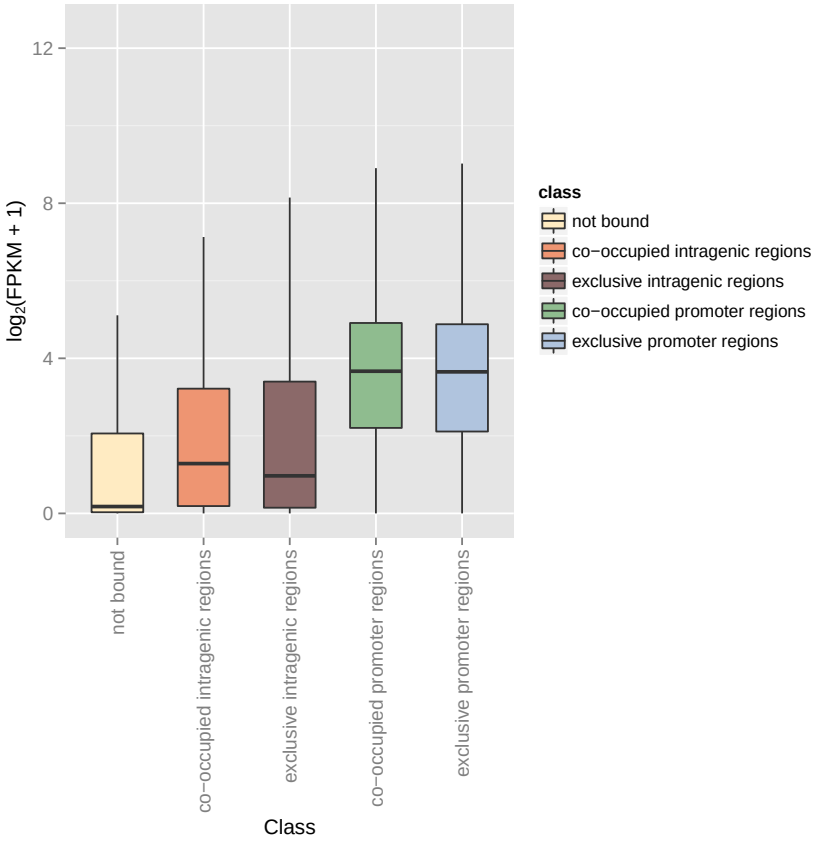


Figure 13: Comparison of the range of gene expression of five different gene sets in megakaryocytes suggest that there is no difference in the gene expression levels between genes bound exclusively MEIS1 or MEIS1 co-localising with any other transcription factor.

As described in Section 2.2.3 MEIS1 peaks often co-localise with other transcription factors. Therefore, I further split the

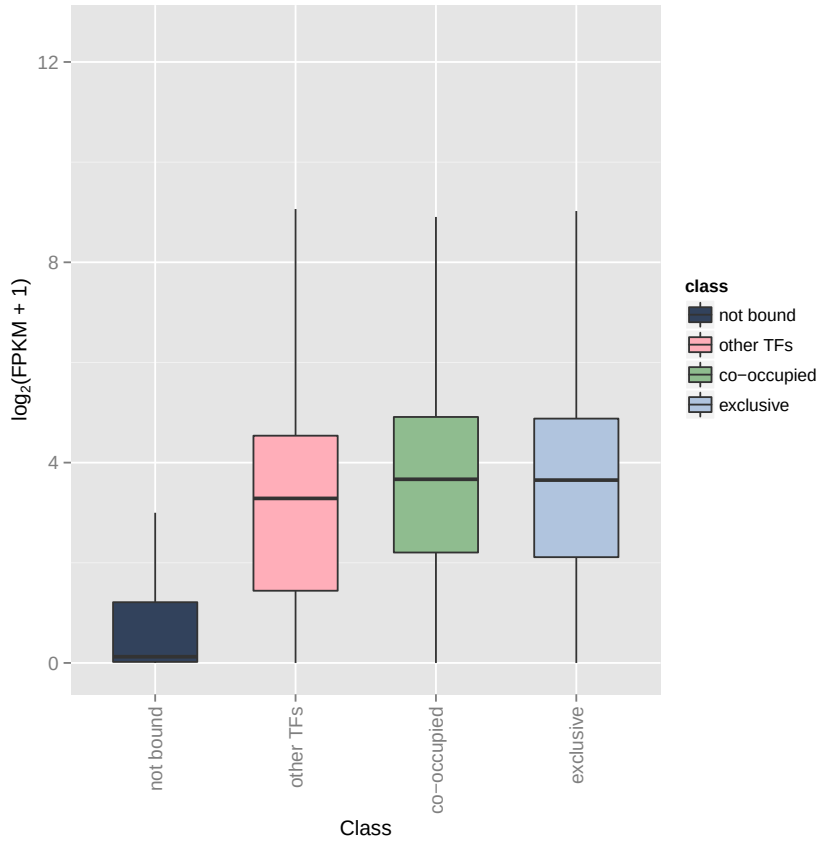


Figure 14: Ranges of expression for four groups of genes; those with an exclusive MEIS1 peak in the promoter region (exclusive), genes with a MEIS1 peak co-localising with any of the other five transcription factors (co-occupied), genes without a MEIS1 peak, but with a binding site for any of the other five transcription factors (other TFs), and genes devoid of binding sites occupied by any of the six transcription factors (not bound). Genes with a binding site for any of the six transcription factors within their promoter region show higher gene expression than those bound by none.

MEIS1 peaks into two subsets; those where MEIS1 binds exclusively and those where MEIS1 co-localises with at least one other transcription factor. Gene expression in megakaryocytes does not appear to be altered if MEIS1 binds exclusively or co-localises with any other transcription factor (see Figure 13). Finally, I focused solely on promoter regions. In the last comparison, I further classified genes not bound by MEIS1 into two subsets; those bound by any of the FLI1, GATA1/2, RUNX1 and SCL transcription factors and those bound by none in the promoter region. Promoter bound genes by any of the five transcription factors, but not MEIS1, show a slight decrease in their

expression values compared to those bound by MEIS1. However, these genes are still expressed at much higher levels, compared to those not bound by any of the six transcription factors (see Figure 14).

2.3.2 *MEIS1 and novel transcription start sites*

Alternative transcription start sites generate different RNA molecules, contributing to protein complexity. But even in the case where the protein product is identical among different RNA molecules, different transcription start sites generate different 5′ untranslated regions, that can also be involved in the regulation of gene expression. For example, the 5′ UTR may harbour a binding site for a miRNA (Bartel, 2009). Nürnberg et al. (2012) reported that MEIS1 marks an alternate promoter in the megakaryocytic-like cell line, CHRF, for the DNM3 gene, leading to the production of a shorter DNM3 transcript. Following this finding, I was interested in identifying if any novel transcription start sites are being used in human megakaryocytes and how often these coincide with a MEIS1 binding site within a window of 500bp around the novel transcription start site.

Using the exons reported by the transcriptome assembly step and the set of novel splice junctions, described in Section 2.1.4, I identified 120 transcripts that contain unannotated first exons. 70 of these were assigned by Cufflinks v.2.0.1 (Trapnell et al., 2010a) to known genes. The remaining 50 transcripts could not be assigned to annotated genes and were reported as novel. For 22 of the 50 novel genes (see Table 9) and for 35 of the 70 novel transcripts (see Table 8), a MEIS1 peak is located in close proximity to the novel transcription start site.

To determine if these novel transcripts can be translated into proteins, I used the NCBI Open Reading Frame Finder

Table 8: Transcripts identified in megakaryocytes with a novel first exon. Transcripts were assembled using Cufflinks v2.0.2 and then the first exons were compared to the annotated exons in Ensembl v.70. Only those having a MEIS1 binding site within 500bp away from their transcription start site are listed below. Novel transcripts were then examined for potential open reading frames using the NCBI ORF Finder. The inferred open reading frames were finally compared to those annotated in Ensembl v.70 and grouped into the following categories: no annotated CCDS, when the transcript does not have an open reading frame, novel, when the inferred translation start and end site are different than any annotated ones, TSS or TES annotated only, when the translation start or end site, respectively, is shared with one of the annotated ones. If both translation start and end sites are annotated, then I provide the Ensembl identifier of the transcript and the CCDS identifier if it exists.

Gene Name	Novel Trascript Coordinates	Novel First Exon Coordinates	Novel Splice Junction Coordinates	TF	Dominant Transcript	Inferred ORF
AC051649.16	chr11:1918230-1929145	chr11:1918230-1918375	chr11:1918375-1923819:+	MEIS1	YES	NO ANNOTATED CCDS
ADAT2	chr6:143745899-143772274	chr6:143771700-143772274	chr6:143768035-143771699:-	MEIS1, RUNX1, FLI1	YES	NOVEL
ADIPOR2	chr12:1800176-1897901	chr12:1800176-1800379	chr12:1800379-1863423:+	MEIS1, RUNX1, FLI1	NO	TSS annotated only (ENST00000357103 : CCDS8511)
ARHGAP35	chr19:47362955-47507709	chr19:47362955-47364466	chr19:47364466-47421744:+	MEIS1, FLI1	MAYBE YES	TSS annotated only (ENST00000404338 : CCDS46127)
BBC3	chr19:47724078-47734504	chr19:47734186-47734504	chr19:47730011-47734185:-	MEIS1	NO	TES annotated only (ENST00000439096 : CCDS12697)
CD226	chr18:67528843-67629068	chr18:67628828-67629068	chr18:67614669-67628827:-	MEIS1, RUNX1, FLI1	YES	ENST00000280200 : CCDS11997
CLDN16	chr3:190120036-190129932	chr3:190120036-190120125	chr3:190120125-190122550:+	MEIS1	YES	novel
CRYBB2P1	chr22:25843828-25908348	chr22:25843828-25844251	chr22:25844251-25849282:+	MEIS1, RUNX1	NO	no annotated CCDS
ELMO1	chr7:36894203-37393629	chr7:37391917-37393629	chr7:37382367-37391916:-	MEIS1	YES	ENST00000310758 : CCDS5449
GS1-124K5.12	chr7:66018825-66057496	chr7:66057244-66057496	chr7:66041916-66057243:-	MEIS1, RUNX1, FLI1	MAYBE YES	no annotated CCDS
KIAA0125	chr14:106345831-106399098	chr14:106345831-106346009	chr14:106346009-106386921:+	all BUT GATA2	MAYBE YES	novel
LINC00674	chr17:66097725-66129305	chr17:66097725-66098100	chr17:66098100-66099083:+	MEIS1	YES	no annotated CCDS
MLH3	chr14:75480046-75491920	chr14:75491890-75491920	chr14:75489721-75491889:-	MEIS1, FLI1, GATA1	NO	novel
NFIB	chr9:14081842-14214522	chr9:14214332-14214522	chr9:14179779-14214331:-	MEIS1, GATA1, RUNX1, SCL	YES	ENST00000543693
NOTCH2NL	chr1:145209070-145281718	chr1:145209070-145209436	chr1:145209436-145248820:+	MEIS1, RUNX1, FLI1	NO	TSS annotated only (ENST00000362074 : CCDS909)
PBX1	chr1:164689978-164821067	chr1:164689978-164690317	chr1:164690317-164761730:+	MEIS1, RUNX1, FLI1	NO	novel
PER2	chr2:239164443-239197474	chr2:239197285-239197474	chr2:239186596-239197284:-	MEIS1, RUNX1	YES	TSS annotated only (ENST00000355768)
PLCH1	chr3:155197762-155273032	chr3:155272957-155273032	chr3:155271985-155272956:-	MEIS1, RUNX1, FLI1, SCL	YES	TES annotated only (ENST00000414191 : CCDS33881)
PVT1	chr8:129057556-129113348	chr8:129057556-129057676	chr8:129057676-129082405:+	MEIS1, RUNX1, GATA1	YES	no annotated CCDS
PVT1 (II)	chr8:129061664-129113348	chr8:129061664-129061994	chr8:129061994-129082405:+	MEIS1	NO	no annotated CCDS
SEC16A	chr9:139334549-139377574	chr9:139377390-139377574	chr9:139372136-139377389:-	MEIS1, FLI1	YES	ENST00000313050 : CCDS55351
SLC23A2	chr20:4833002-4982176	chr20:4982052-4982176	chr20:4951561-4982051:-	MEIS1	MAYBE YES	TES annotated only (ENST00000424750)
SMOX	chr20:4135883-4167959	chr20:4135883-4136113	chr20:4136113-4155676:+	all	NO	TSS annotated only (ENST00000278795 : CCDS13078)
SYNCRIP	chr6:86324618-86353096	chr6:86352515-86353096	chr6:86350282-86352514:-	MEIS1	NO	novel
THBS1	chr15:39872612-39891084	chr15:39872612-39872789	chr15:39872789-39873331:+	MEIS1	NO	novel
TMCC2	chr1:205236716-205240626	chr1:205236716-205237187	chr1:205237187-205238077:+	all BUT GATA2	YES	novel
TNFAIP2	chr14:103587785-103603776	chr14:103587785-103587836	chr14:103587836-103589773:+	MEIS1	MAYBE YES	TSS annotated only (ENST00000560869 : CCDS9979)
TRANK1	chr3:36868311-36986569	chr3:36986278-36986569	chr3:36949945-36986277:-	MEIS1	YES	TES annotated only (ENST00000428977)
TREML2	chr6:41156774-41169778	chr6:41169649-41169778	chr6:41168824-41169648:-	all	YES	ENST00000483722 : CCDS54853
TREML2	chr6:41156774-41169778	chr6:41169649-41169778	chr6:41168847-41169648:-	all	YES	TES annotated only (ENST00000483722 : CCDS54853)
UNC50	chr2:99225307-99234978	chr2:99225307-99225474	chr2:99225474-99226105:+	MEIS1, RUNX1, FLI1	NO	novel
ZMYM2	chr13:20533016-20642489	chr13:20533016-20533275	chr13:20533275-20534097:+	MEIS1	NO	novel
ZNF777	chr7:149128452-149157812	chr7:149157121-149157812	chr7:149153128-149157120:-	MEIS1	YES	TES annotated only (ENST00000247930 : CCDS43675)
ZNF833P	chr19:11750574-11797382	chr19:11750574-11750711	chr19:11750711-11759172:+	MEIS1	YES	no annotated CCDS

Table 9: Novel candidate gene targets with a MEIS1 peak within 500bp from their transcription start site. The gene names are unique identifiers reported by Cufflinks. The third column includes the transcription factors that are located within 500bp from the transcription start site.

Gene ID	Novel Gene Coordinates	TF
NOVEL_GENE.2	chr3:185278080-185278538:-	MEIS1, FLI1
NOVEL_GENE.4	chr4:56794308-56805121:-	MEIS1, FLI1, RUNX1
NOVEL_GENE.6	chr4:119199927-119199947:+	MEIS1
NOVEL_GENE.7	chr4:119199931-119200292:+	MEIS1
NOVEL_GENE.11	chr5:55422258-55429200:+	all but GATA2
NOVEL_GENE.13	chr6:27093591-27094241:-	MEIS1, RUNX1
NOVEL_GENE.14	chr6:32862006-32862704:+	MEIS1
NOVEL_GENE.18	chr6:137105075-137105412:+	MEIS1
NOVEL_GENE.20	chr7:123174647-123175479:+	MEIS1
NOVEL_GENE.21	chr8:8993765-8999186:-	MEIS1, GATA1
NOVEL_GENE.22	chr1:111201512-111202208:-	MEIS1
NOVEL_GENE.27	chr10:112135904-112136252:-	MEIS1, FLI1
NOVEL_GENE.28	chr10:112877718-112894332:+	all but GATA1/2
NOVEL_GENE.31	chr11:94801818-94804388:+	MEIS1
NOVEL_GENE.36	chr12:127650445-127650712:+	all but SCL
NOVEL_GENE.38	chr15:102035169-102038884:-	MEIS1, GATA1, SCL
NOVEL_GENE.39	chr15:102035169-102039950:-	MEIS1, GATA1, SCL
NOVEL_GENE.44	chr21:16333556-16340847:-	MEIS1, RUNX1
NOVEL_GENE.47	chrX:65235277-65237031:+	all but GATA2
NOVEL_GENE.48	chrX:65236907-65241220:+	all but GATA2
NOVEL_GENE.50	chr2:68693204-68693605:-	MEIS1
NOVEL_GENE.51	chr2:70021205-70026314:-	MES1

(http://www.ncbi.nlm.nih.gov/projects/gorf/) on the set of novel transcripts assembled by Cufflinks (see Table 8). I then compared the inferred open reading frames to the open reading frames annotated in Ensembl v70. These were classified as:

1. novel, when neither of their ends is annotated,
2. no annotated CCDS, when there is no annotated coding sequence for the gene in Ensembl,
3. TSS or TES annotated only, when the translation start site (TSS) or the translation end site (TES), respectively, are annotated.

The list of transcripts with a novel first exon and a MEIS1 peak in the vicinity include transcription factors (see Table 8) such as ARHGAP35, NFIB and PBX1. ARHGAP35 is a transcription factor that has not been directly linked to megakaryopoiesis, but has been shown to act as a repressor on the glucocorticoid receptor (LeClerc et al., 1991). The ARHGAP35 transcript expressed in megakaryocytes contains an additional exon relative to the annotated isoform (see Figure 16(a)), creating an RNA molecule with a longer 5′ UTR. The inferred translation start site does not seem to be affected. In the case of PBX1 and TMCC2, however, the megakaryocytic transcript is much shorter than those annotated (see Figure 16(b, c)), and the inferred coding sequences are novel.

There is evidence in the literature that PER2 and TMCC2 play a role in platelet formation. Although Per2 null mice do not display a phenotypic reduction in megakaryocyte production, an increase in megakaryocytic polyploidy is observed with a 50% reduction in the number of platelets in the circulating blood (Zhao et al., 2011). A recent genome-wide association study identified the TMCC2 locus as correlated to mean platelet volume in Europeans (Gieger et al., 2011).

These examples suggest that MEIS1 may play a role in gene isoform expression in megakaryocytes. However, to systematically test the hypothesis that MEIS1 drives megakaryocyte-specific isoform expression, all expressed transcripts must be considered without limiting the analysis to annotated genes. To do this, RNA sequencing data are required for closely related blood cell types where MEIS1 is not expressed. It will then be possible to compare the expressed transcript variants at a genome-wide level and draw conclusions about the role of MEIS1 in megakaryocyte-specific transcription.

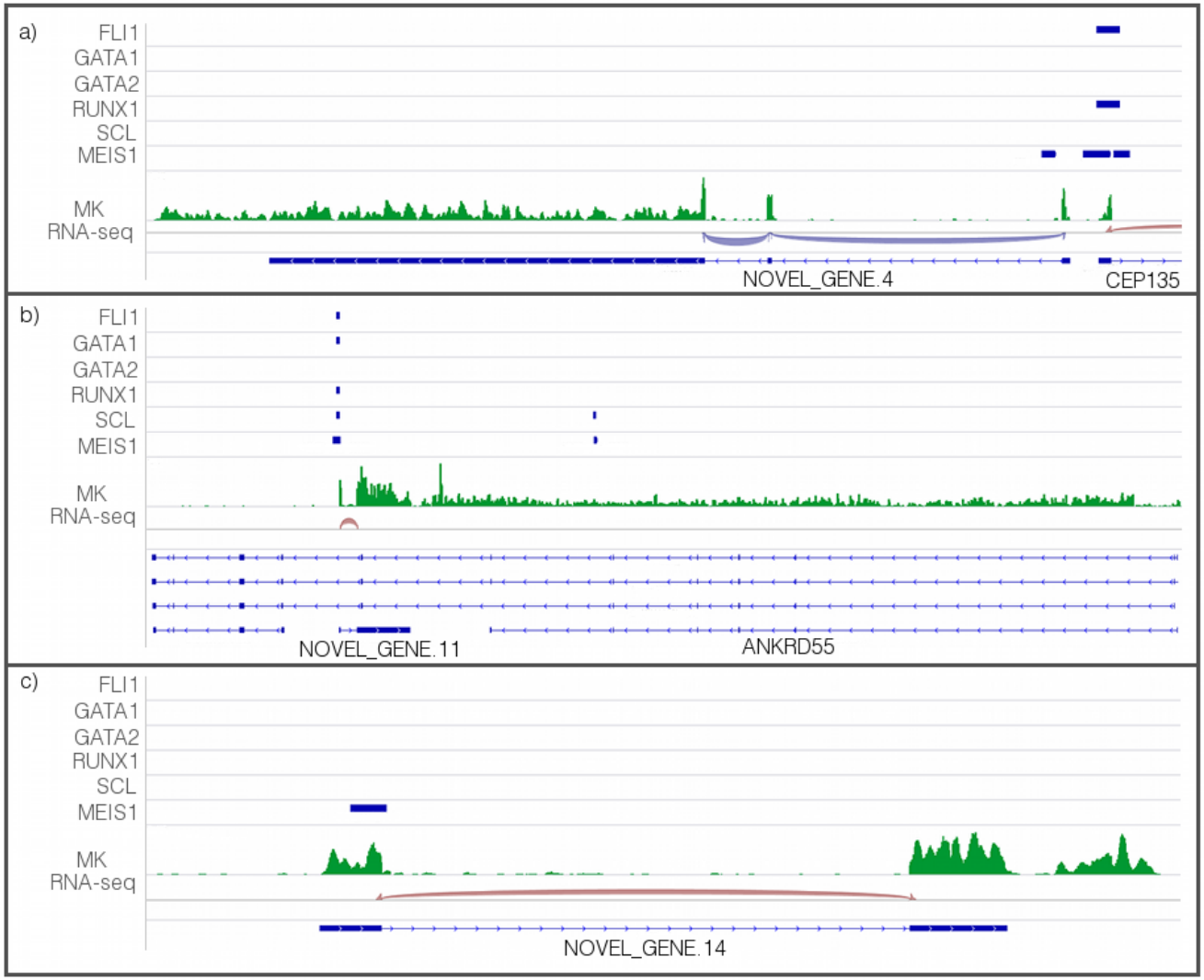


Figure 15: Characteristic examples of novel genes from Table 9. Although it is not possible to infer strand specificity, GSNAP and Cufflinks predict the strand based on the canonical splicing of the reads spanning a splice junction.
(a) NOVEL_GENE.4 is located immediately upstream of the CEP135 gene and is an example of an unannotated long non-coding RNA or an enhancer RNA.
(b) NOVEL_GENE.11 is located within the ANKRD55 gene and shows a higher expression than the annotated gene.
(c) NOVEL_GENE.14 is located in an intergenic region and is not in the vicinity of any annotated protein-coding gene.

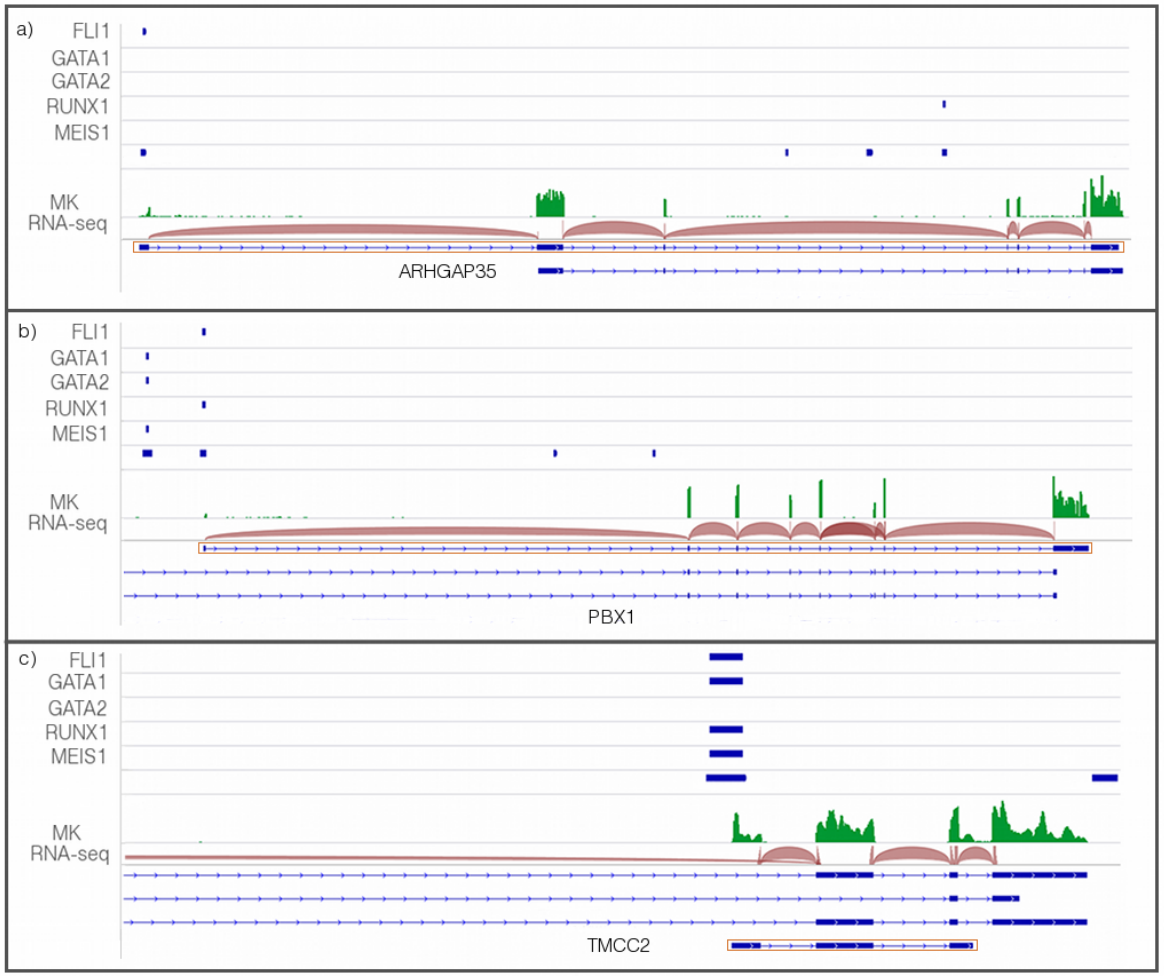


Figure 16: Three examples of novel transcripts in megakaryocytes, for which unannotated transcription start sites are located within close proximity to a MEIS1 peak. The novel transcripts are highlighted in the orange boxes.
(a) Our RNA-seq data reveal a longer isoform for the transcription factor ARHGAP35. The unnanotated isoform contains a novel 5′ exon. (b) The unannotated isoform for the transcription factor PBX1 is much shorter than those annotated. Its transcription start site is close to a intronic MEIS1 binding site. (c) The novel isoform identified for TMCC2 is also much shorter than those annotated.

2.4 METHODS

All wet-lab experiments described below have been performed by members of Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK and are only provided for completeness. I was responsible for the computational analysis of the data sets produced.

2.4.1 *Library preparation and sequencing of megakaryocytic RNA*

The cell culture and the RNA-seq library were prepared by Dr Peter Smethurst and Dr Katrin Voss, in Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge. Megakaryocytes were obtained from cord blood-derived $CD34^+$ haematopoietic stem cells cultured for 7-12 days in a medium supplemented with human recombinant thrombopoietin, a cytokine that promotes the differentiation of HSCs into the megakaryocytic lineage, and interleukin-1 β (IL1B), that promotes the expansion of the stem cell pool. At the completion of the culture about 70-90% of the cells had a megakaryocytic phenotype with the majority being $CD41^+CD42b^+$ and $CD34^-$.

Total RNA was prepared from cultured megakaryocytes that were obtained using the protocol above and RNA was extracted using Trizol, according to the manufacturer's protocol (Invitrogen, Paisley, UK). The RNA pellet was resuspended in nuclease-free water (Applied Biosystems, Warrington, UK) and analysed using the Agilent Bioanalyser 2100 (Agilent, Waldbronn, Germany), which gave a RNA integrity number (RIN) of 8.4. Following DNAase treatment (Turbo DNA-Free, Applied Biosystems), 5 μg of total RNA was used as input for the mRNA Sequencing kit (Illumina) following the manufacturer's instructions, except PCR was performed before gel extraction of a size range of 150-200bp in the first run and 300-450bp in the second

run, to obtain the purified library. The library was quantified by qPCR followed by paired-end sequencing.

Two libraries of poly(A)$^+$-selected RNA originated from the same biological sample, but with a difference in fragment size as described above. Both libraries were sequenced on Illumina Genome Analyzer II yielding 27.5M and 40.5 M paired-end 76 bp reads, respectively.

2.4.2 *Pre-alignment quality control*

Pre-alignment assessment of the reads was done using the FASTQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads were then aligned to the *Homo sapiens* high coverage assembly (hg19) (release February 2009) using GSNAP version 2011-11-25. Read trimming was disabled and I allowed for up to 5 mismatches and novel splicing sites at most 300,000 bp apart. Trimming adapters is not an essential step before using GSNAP, as the tool soft clips adapter sequences.

2.4.3 *Post-alignment quality control*

To assess the quality of the alignments, I calculated the percentage of mapped reads for each lane. GSNAP allows for the mates of a paired-end read to align in the following ways: "concordant", "paired", "unpaired", and "halfmapping".

1. concordant: a paired-end read has a concordant alignment, if both mates align to the same chromosome, in the expected orientation and distance.

2. paired: if either of the orientation or distance is not as expected, then GSNAP reports a paired alignment. However, both mates need to align onto the same chromosome.

3. halfmapping: if only one of the two mates can be aligned onto the genome, then GSNAP reports a halfmapping alignment.

4. unpaired: in case the criterion of alignment onto the same chromosome is violated, then the alignment is classified as unpaired.

The majority of the reads align uniquely to the human genome (concordant uniq). Although the second lane had a smaller fraction of reads aligned to the genome, the absolute numbers of reads with a concordant uniq alignment are comparable between the two lanes.

Custom scripts were used to extract the splice junction coordinates from the spliced alignments. For downstream analysis, the two technical replicates were pooled together.

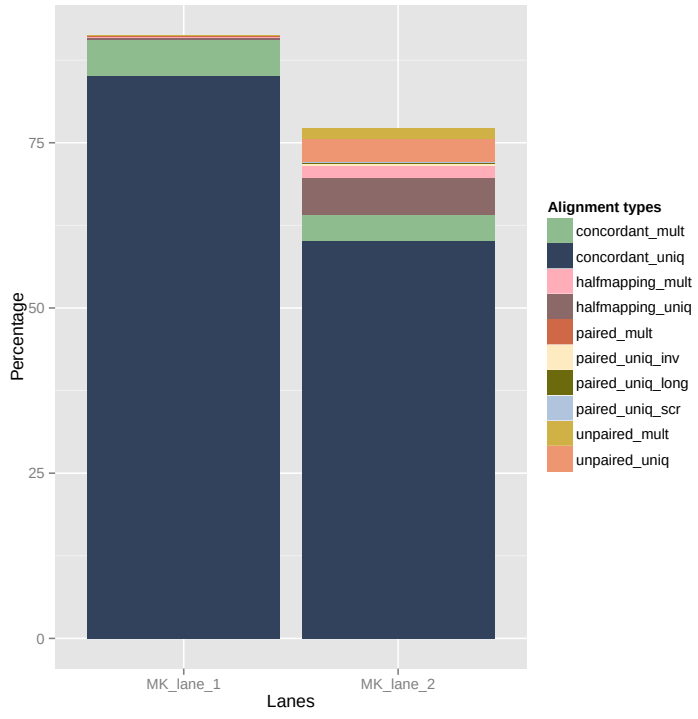


Figure 17: Percentage of fragments aligned to the human genome for each of the two megakaryocytic RNA-seq libraries. Post alignment assessment of RNA-seq data shows that the majority of our fragments are uniquely aligned to the human genome.

2.4.4 *Quantification of gene expression and Transcriptome Assembly*

Gene and isoform quantification was performed using Cufflinks v2.0.2 (Trapnell et al., 2010a) and reported in Fragments Per Kilobase of exon per Million fragments mapped (FPKM). To catalogue the transcripts present in megakaryocytes, I used Cufflinks v2.0.2 (Trapnell et al., 2010a). The assembly step used the existing annotation as a guide, allowing for identification of novel transcripts as well. The annotation used was Ensembl v70.

2.4.5 *Library preparation and sequencing of MEIS1 ChIP*

1 million $CD34^+$ cells (10e 5 per 1mL; 1mL per well ; 10 wells) were cultured with TPO 100ng/mL and IL1B 10ng/mL in CellGro. At the end of the culture, the cells were 75% $CD41^+$ and 58% $CD42b^+$. Culture and immunophenotyping was peformed as described in Tijssen et al. (2011). ChIP for MEIS1 was performed as described previously in Nürnberg et al. (2012) and showed 4 to 5-fold enrichment of a known MEIS1 genomic binding site (element around rs2038479; identified in Nürnberg et al. (2012)) by qPCR. The ChIP-seq library was prepared according to Illumina ChIP-Seq preparation kit (Part No 1003473).

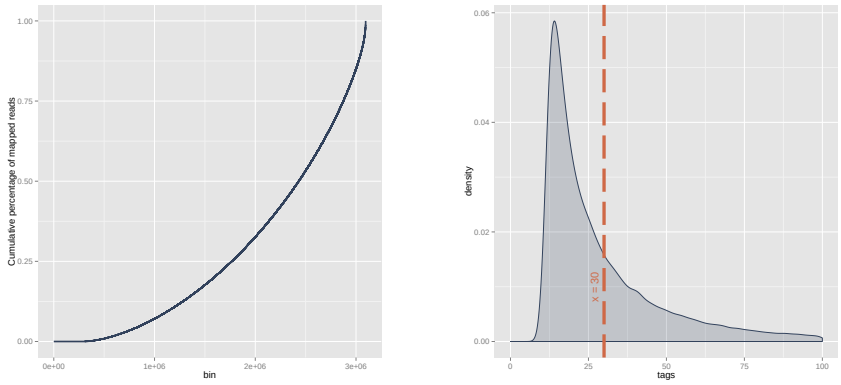
Sequencing was performed on a HiSeq 2000 Illumina sequencer, producing two lanes of 50bp paired-end reads. Again, the read quality assessment was performed using FASTQC. Reads were then aligned to the *Homo sapiens* high coverage assembly (hg19) (release February 2009) using Bowtie v1.0.0. The set of parameters used to run Bowtie was "-y –nomaqround -e 120 -v 3 -m 1 -p 8 –best –strata –sam".

Post-alignment quality control was based on the IP enrichment, a control step introduced by Diaz et al. (2012). In this analysis, the human genome was subdivided into bins of 1kb

in length and I counted the number of reads aligning to each bin. For plotting purposes, bins were then sorted based on the number of reads aligning to each one of them.

2.4.6 *MEIS1 downstream analysis*

Identification of enriched regions was performed using MACS v1.4 without a control sample. Only uniquely mapped reads were considered and duplicated reads were discarded prior to peak calling. In total, 26,726 peaks were identified. The distribution of aligned reads per peak shows that the majority of the peaks were supported by a low number of reads. Therefore, to limit false positive enriched regions, I filtered out any peak that was covered by fewer than 30 reads.



(a) (b)

Figure 18: Post-alignment quality control of the MEIS1 ChIP-seq data.
(a) ChIP enrichment across genome 1kb windows.
(b) Distribution of reads in the MEIS1 enriched regions identified by MACS.

Subsequent peak annotation was performed using custom R scripts. Motif analysis was performed on the top 1,000 MEIS1 peaks using MEME. I allowed for the identification of a maximum of 3 motifs with a length between 6 and 30bp. Finally, motif enrichment analysis of a curated set of transcription factors was performed using custom R scripts. The set of curated position weight matrices used, which represent the binding spe-

cificity of each transcription factor, was based on the study by Jolma et al. (2013). To calculate the *P*-value of each match I used 1,000 sequences that were randomly sampled from the pool of sequences that had the same length as the input and had the same dinucleotide frequency, as this was defined by a 1st order Markov model. Input sequences were masked for repeats using RepeatMasker, before scanning for motifs.

3

TRANSCRIPTOME ANALYSIS OF HAEMATOPOIETIC PROGENITOR CELLS: THE BLUEPRINT PROJECT

3.1 BLUEPRINT AND ITS AIMS

BLUEPRINT (Adams et al., 2012) is a EU-founded consortium that constitutes the European contribution to the International Human Epigenome Consortium (IHEC). IHEC aims to map 1,000 human epigenomes. BLUEPRINT's contribution will be the study of the human epigenome in more than 100 normal and malignant blood cell types. The goals of BLUEPRINT can be summarised as follow:

1. Data production and methods development: these constitute the mapping phase. Data sets from normal cell types will be used to draw the relationships between genomics and epigenomics.

2. Study of epigenome variation: establishing the variation of the epigenome across healthy and diseased individuals.

3. Identification of biomarkers: in acute lymphoblastic leukemia for prognostic or therapeutic purposes.

4. Drug screening: testing the efficiency of compounds to reverse disease-related epigenetic states.

Considerable effort is being made in the mapping phase of BLUEPRINT to isolate healthy blood cell types, from which reference epigenomes are obtained. Different sequencing applications are used to characterise the reference epigenome: RNA-

seq for transcriptome analysis, whole genome bisulfite sequencing for methylome analysis, DNaseI hypersensitivity coupled with sequencing for identification of hypersensitive sites and ChIP-seq for the IHEC core set of six histone modification marks (H3K4me3, H3K4me1, H3K9me3, H3K27me3, H3K27ac, H3K6me3).

This chapter is dedicated to the analysis of the transcriptome of rare haematopoietic progenitors using RNA sequencing. The requirement for a large number of cells for the majority of the above applications led us to focus on transcriptome analysis in cell types where we were only able to isolate a small number of cells. This project is a collaboration among four different research groups: Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, Dr Roderic Guigo's group, Centre for Genomic Regulation, Barcelona, Prof Henk Stunnenberg's group, Nijmegen Centre of Molecular Life Sciences and Dr Paul Bertone's group, EMBL-EBI, Cambridge.

3.2 RNA SEQUENCING OF RARE HUMAN HAEMATOPOIETIC PROGENITORS

3.2.1 *Gene expression profiles of haematopoietic progenitors using whole-genome expression arrays*

Despite the fact that the haematopoietic system is one of the most extensively studied systems in the human body, there are still many aspects to be explored. The intermediate stages of differentiation affect the correct formation of mature blood cells. It is therefore important to study gene expression in haematopoietic progenitors and to identify changes that occur at the transcriptional level. Advances in fluorescence activated cell sorting and monoclonal antibodies have made possible the isolation of these rare blood cell types (see Section 1.1.3). To this end several groups have performed gene expression profiling of

Table 10: Genome-wide expression studies on human and murine haematopoietic progenitors.

	HSC	MPP	CMP	LMPP	MEP	GMP	Pro B	B-NK prog.	mature cells	Reference
Human	a		a		a	a	a		erythrocytes, megakaryocytes, granulocytes, monocytes & various lymphoid	(Novershtern et al., 2011)
	($CD90^+$ & $CD90^-$)[a]		a	a		a	a	a		(Laurenti et al., 2013)
			($CD34^+$ HSPCs)[c,f]						ex-vivo differentiation of HSPCs into erythroblasts	(Xu et al., 2012)
Mouse	(LT- & ST-)[b]			a						(Mansson et al., 2007)
	(LT- & ST-)[b]								NK[e], T-cells[e], B-cells[e], monocytes[e], granulocytes[b], erythrocytes[b],	(Chambers et al., 2007)
	($CD41^+$c-kit^+CD34^+)[d], E11.5[e], E.12.5[f], E13.5[f], E14.5[f], adult[b], z									(McKinney-Freeman et al., 2012)

a: umbilical cord
b: bone marrow
c: peripheral blood
d: yolk sac
e: spleen
f: fetal liver
g: emryonic stem cell differentiated

haematopoietic progenitors using microarray technology (see Table 10). Below I provide a summary of these experiments, in both human and mouse haematopoietic progenitors, from various sources and developmental stages.

Mansson et al. (2007) studied the gene expression profiles of three early cell populations in mouse haematopoiesis, long-term and short-term haematopoietic stem cells and multilymphoid progenitors. To identify the different gene regulatory networks that govern immature and terminally differentiated haematopoietic cells, Chambers et al. (2007) profiled haematopoietic stem cells and a set of mature blood cell types (T-cell, B-cells, NK cells, nucleated erythrocytes, monocytes and granulocytes) from peripheral blood, bone marrow and spleen of adult mice. Complementary to the previous studies of haematopoietic stem cells from adult mice, McKinney-Freeman et al. (2012) focused on the characterisation of gene expression of embryonic haematopoietic stem cells in a large study that included 2,500 murine embryos. Haematopoietic stem cells were purified from embryos at various developmental stages from mid-gestation to adulthood. In addition they compared these to haematopoietic stem cells derived from cultured embryonic stem cells.

Gene expression analysis has also been performed on human haematopoietic cells, with the study published by Novershtern et al. (2011) being the largest to date (described already in 2.1.1). Xu et al. (2012) compared fetal to adult erythropoiesis using differentiated erythrocytes isolated from human fetal liver and peripheral blood. Combining gene expression data with histone modifications and transcription factor occupancy data, they identified enhancers specific to each developmental stage and new regulators of erythropoiesis. Finally, to study the equivalent of multilymphoid progenitors in humans, Laurenti et al. (2013) isolated haematopoietic stem and progenitor cells from human cord blood. Their analysis centered around B cell differentiation of the lymphoid system and multilymphoid progenitors in humans.

A common characteristic of the above-mentioned studies is that they have all been performed on microarray platforms. BLUEPRINT aims to incorporate sequencing applications to investigate in further detail the expression profiles of haematopoietic progenitor cells. This will provide greater opportunities to identify novel features, perform analyses at transcript level and characterise cell-/lineage-specific splicing events.

3.2.2 *RNA-seq library preparation*

Rare haematopoietic progenitors were isolated from cord blood using the cell surface markers summarised in Table 11. The cell isolation was performed in Prof Willem Ouwehand group, Department of Haematology, University of Cambridge, UK. RNA isolation and subsequent RNA-seq libraries were prepared in Prof Stunnenberg's lab, Nijmegen, the Netherlands. Poly(A)$^+$ RNA was sequenced on a HiSeq 2000 machine yielding on average 160 million paired-end reads of 100bp length (see Section 3.5.1 for detailed RNA-seq library preparation).

The haematopoietic cell types profiled include: haematopoietic stem cells (HSC), multipotent progenitors (MPP), common lymphoid progenitors (CLP), common myeloid progenitors (CMP), granulocyte/monocyte progenitors (GMP), and megakaryocyte/erythrocyte progenitors (MEP) (summarised in Figure 19). BLUEPRINT aims to generate three biological replicates per cell type. However, at the time of this writing not all samples have been produced (available replicates are indicated in Figure 19).

Table 11: Cell surface markers for the isolation of haematopoietic progenitors (reviewed in (Seita and Weissman, 2010)). The third column represents the proportion of each subclass found within the $CD34^+$ fraction.
HSC: haematopoietic stem cell, MPP: multipotent progenitor, CLP: common lymphoid progenitor, CMP: common myeloid progenitor, MEP: megakaryocyte/erythrocyte progenitor, GMP: granulocyte/monocyte progenitor

Cell type	Cell surface phenotype	HSPCs from $CD34^+$ cells (%)
HSC	$Lin^-CD34^+CD38^-CD90^+CD45RA^-$	1
MPP	$Lin^-CD34^+CD38^-CD90^-CD45RA^-$	7
CLP	$Lin^-CD34^+CD38^+CD10^+CD45RA^+$	0.5
CMP	$Lin^-CD34^+CD38^+IL\text{-}3Ra^{low}CD45RA^-$	8
MEP	$Lin^-CD34^+CD38^+IL\text{-}3Ra^-CD45RA^-$	4
GMP	$Lin^-CD34^+CD38^+IL\text{-}3Ra^+CD45RA^+$	4

3.2.3 *Quality assessment of progenitor datasets*

To determine the quality of our RNA-seq data, we computed basic post-alignment statistics, such as the percentage of mapped reads, computed in collaboration with Pawan Poudel, Prof Ouwehand's lab, Department of Haematology, University of Cambridge (see Section 3.5.2.1). Following quantification of gene expression in each sample, I examined various cell surface markers and known regulators of the haematopoietic system.

Principal component analysis indicated a clear separation between haematopoieitc stem cells and the rest of the samples (see Figure 20), as seen also by Laurenti et al. (2013). However,

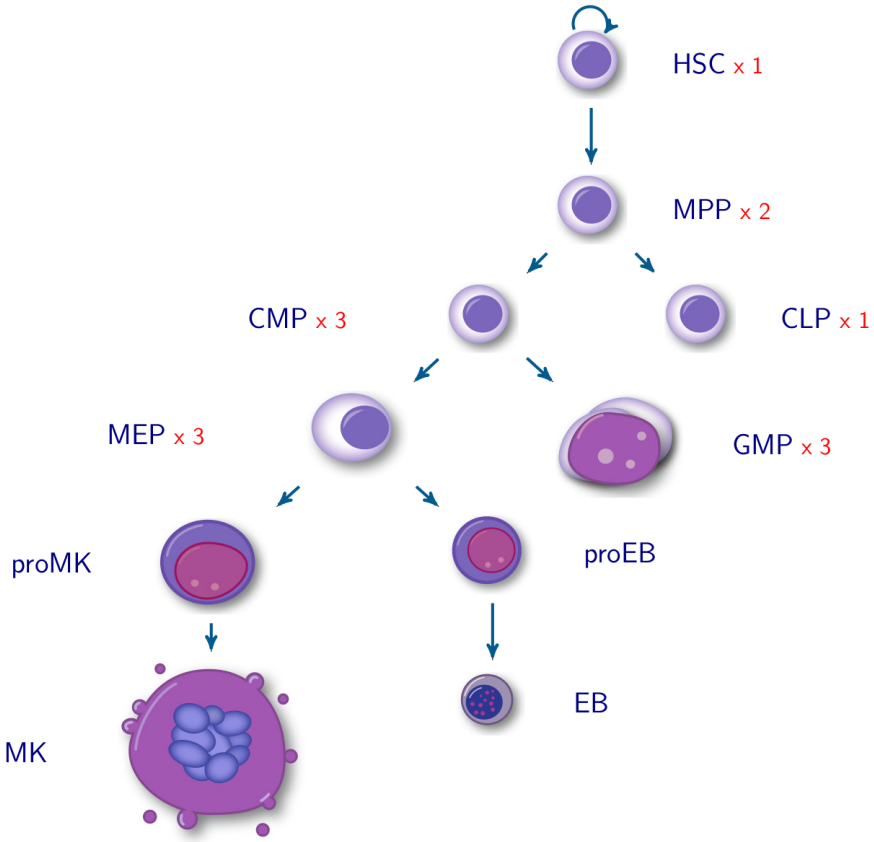


Figure 19: Hierarchical representation of the haematopoetic progenitor cells profiled for the mapping part of the BLUEPRINT project. The numbers in red indicate biological replicates produced so far.
HSC: haematopoietic stem cell, MPP: multipotent progenitors, CLP: common lymphoid progenitors, CMP: common myeloid progenitors, GMP: granulocyte/monocyte progenitors, MEP: megakaryocyte/erythrocyte progenitors.

all other samples clustered closely, with the MPPs being closer to HSCs than any other cell type. The close clustering of all samples but HSCs can be explained by the fact that each of these cell types are similar immature haematopoietic progenitor cells. Moreover, the reduced complexity of the sequencing samples caused by low initial cell numbers (see Table 11) reduces our capacity to identify subtle changes and differences in genes with low level of expression.

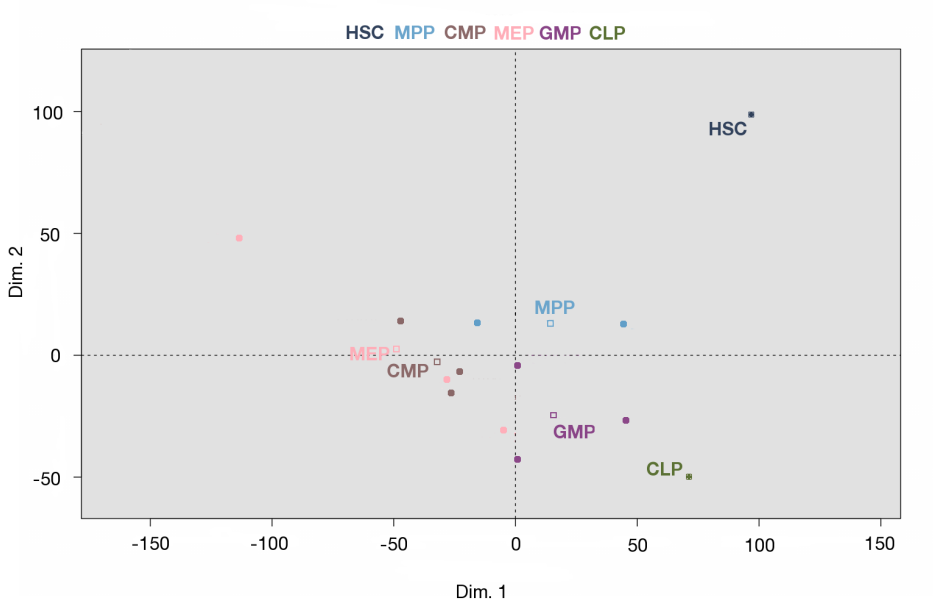


Figure 20: Sample clustering based on the first two dimensions of the principal component analysis (PCA) on the haematopoietic progenitors RNA-seq datasets. PCA was performed on protein coding genes that are expressed (FPKM > 1) in at least one cell type.

3.2.3.1 *Expression levels of cell surface markers*

Overall expression of the genes encoding various cell surface markers, listed in Table 11, agrees with their cell surface presentation (see Figure 21). In some cases, discussed in more detail below, expression does not follow the expected pattern. In interpreting these discrepancies, it is important to keep in mind that mRNA expression does not always correlate with protein expression and/or cell surface expression of the antigens. Therefore, despite the fact that these cells express genes that encode for surface receptors, there is no guarantee that these are translated and that they would be on the cell surface. The discrepancies I observe may be due to intermediate regulatory mechanisms between gene expression and cell surface expression of proteins.

Each of the multipotent progenitors (one haematopoietic stem cell sample and two multipotent progenitor samples) transcribe CD38 (see Figure 21), despite the fact that multipotency

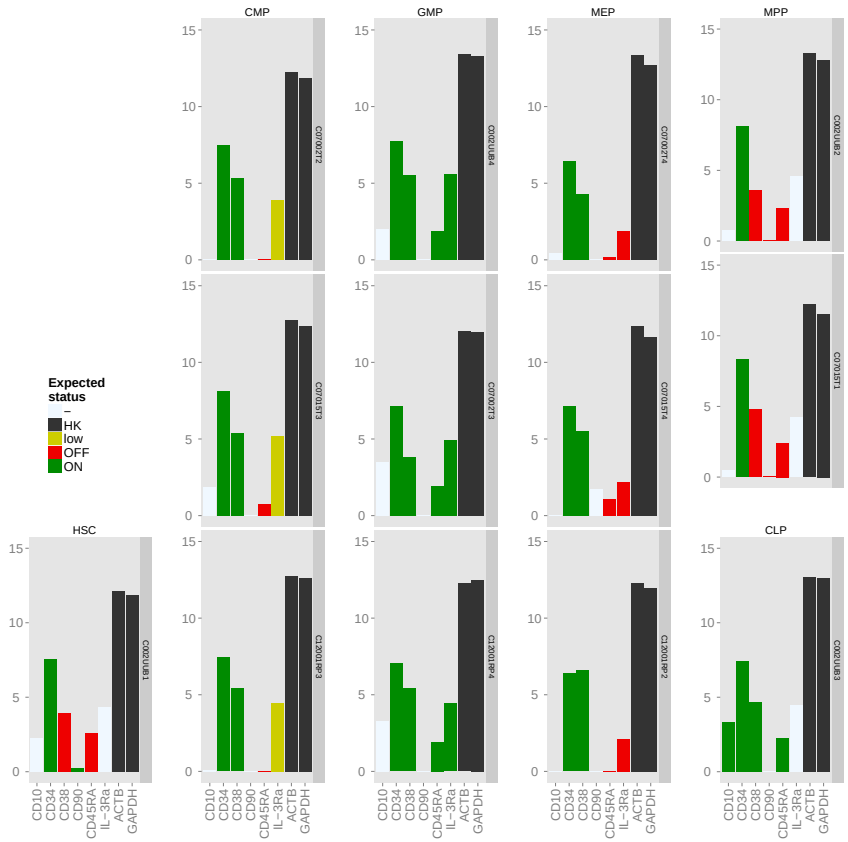


Figure 21: Expression levels of various cell surface markers in the BLUEPRINT data.
The selected marker genes are: CD10: Neprilysin, CD34: Hematopoietic progenitor cell antigen CD34, CD38: ADP-ribosyl cyclase 1, CD90: Thy-1 membrane glycoprotein, CD45RA: Receptor-type tyrosine-protein phosphatase C, IL-3Ra: Interleukin-3 receptor subunit alpha, ACTB: Actin cytoplasmic 1, GAPDH: Glyceraldehyde-3-phosphate dehydrogenase, HK: housekeeping genes.

in the haematopoietic cells is characterised by the absence of CD38 on the surface of these cells (Terstappen et al., 1991). In the case of interleukin-3 receptor (IL-3Ra), cell surface expression distinguishes among the common myeloid progenitors (low expression), granulocyte/monocyte progenitors (positive) and megakaryocyte/erythrocyte progenitors (negative). However, my analysis showed that at the transcriptional level common myeloid progenitors and granulocyte/monocyte progenitors express this gene at similar levels, while the expression decreases in the megakaryocyte/erythrocyte progenitors. However, IL-3Ra transcription is not abolished there, as one might

expect due to its absence from the cell surface of these cells (see also Figure 24). Finally, the expression of CD45RA is high in granulocyte/monocyte progenitors and common lymphoid progenitors in accordance with its presence on the cell surface of these cell types. However, these are not the only haematopoietic progenitors where CD45RA is expressed, as it is also found in multipotent haematopoietic progenitors where there is no expression of this marker on the cell surface.

The discrepancies described above highlight the differences between gene expression levels and cell surface presentation and are in no case indicative of any problems with the samples.

3.2.3.2 *Expression levels of key haematopoietic regulators*

The second assessment I made using the RNA-seq data was based on key regulators with known lineage specific expression and function in haematopoiesis. The gene set includes those expected to be expressed in:

1. haematopoietic stem cells due to their regulatory role in stem cell proliferation and renewal (HOXB4, NOTCH1, GATA2),
2. common lymphoid progenitors because they are involved in the lymphoid development (EBF1, SPI1),
3. granulocyte/monocyte progenitors due to their role in the granulocyte/monocyte development (IRF8), and
4. megakaryocytic/erythroid progenitors because they promote the erythroid (GATA1) and
5. megakaryocytic differentiation (GATA2).

The expression levels of these key transcription factors in our dataset are all consistent with their described regulatory roles in blood development (see Figure 22).

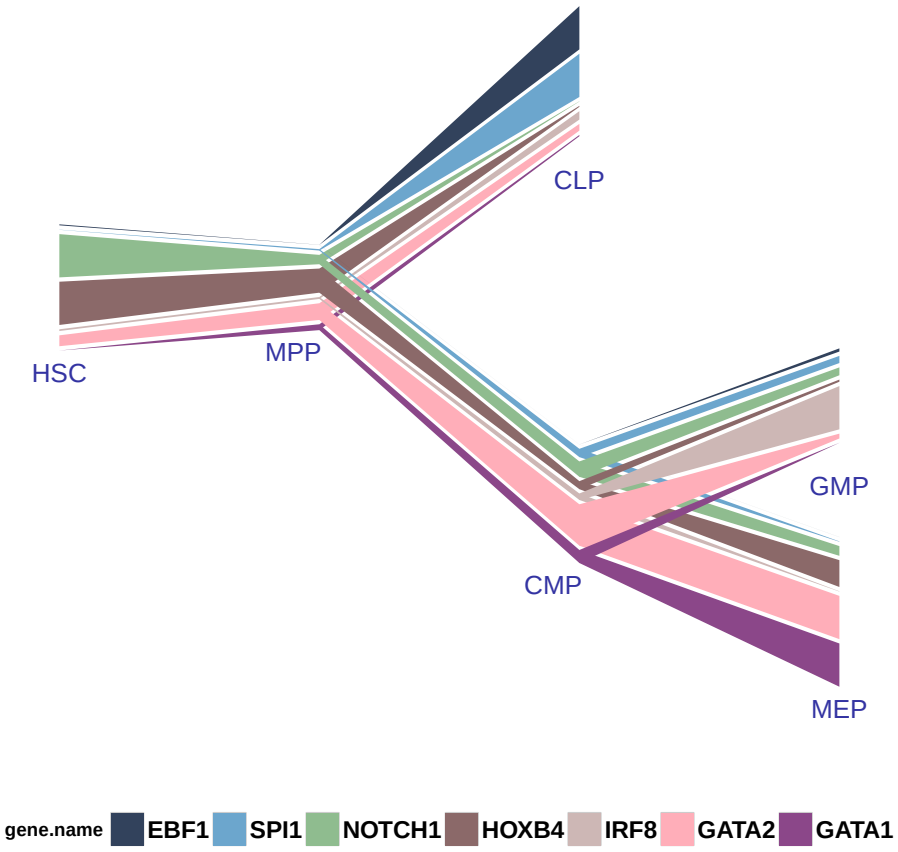


Figure 22: Gene expression of known key haematopoietic regulators. Width corresponds to expression level relative to the maximum expression of the gene across all cell types.
The selected marker genes are: EBF1: Early B-cell factor, SPI1: Transcription factor PU.1, NOTCH1: Neurogenic locus notch homolog protein 1, HOXB4: Homeobox protein Hox-B4, IRF8: Interferon regulatory factor 8, GATA2: Endothelial transcription factor GATA-2, GATA1: Erythroid transcription factor.

3.2.4 *Identification of cell-type and lineage-specific genes*

Gene expression constitutes the basis of a cell's phenotype. By identifying sets of genes that are specific to a cell type or lineage, we may be able to delineate the mechanisms that drive developmental progression. The absence of replicates for some cell types limits the power of the analysis presented below. However, the same methods will be applied to the full data set when the remainder of the samples have been produced.

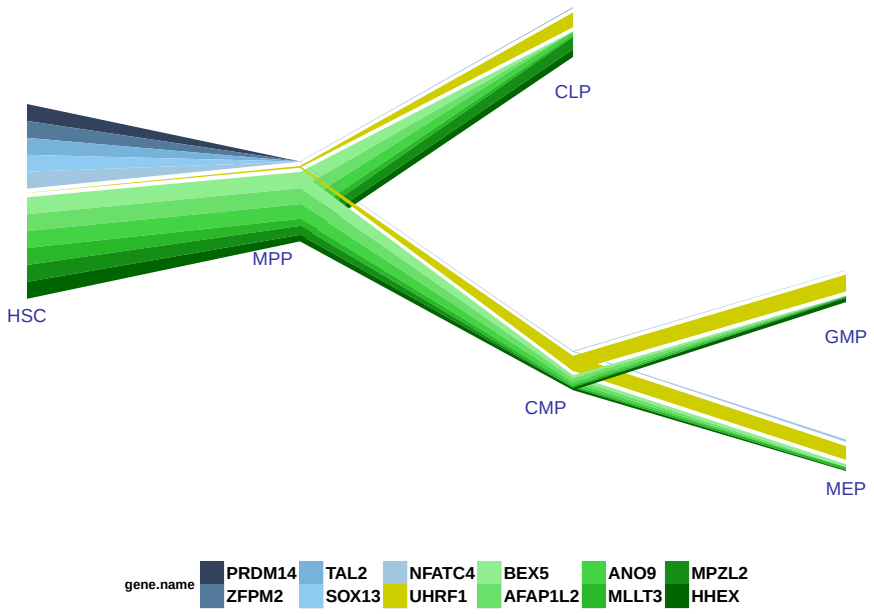


Figure 23: Expression patterns of multipotent progenitor-specific genes. Example of gene sets that are either expressed in the stem cell compartment (blue shaded), all progenitors but haematopoetic stem cells (yellow), or in the multipotent cells (green shaded). Width corresponds to expression level relative to the maximum expression of the gene across all cell types.
PRDM14: PR domain containing 14, ZFPM2: zinc finger protein - FOG family member 2, TAL2: T-cell acute lymphocytic leukemia 2, SOX13: sex determining region Y-box 13, NFATC4: nuclear factor of activated T-cells cytoplasmic calcineurin-dependent 4, UHRF1: ubiquitin-like with PHD and ring finger domains 1, BEX5: brain expressed X-linked 5, AFAP1L2: actin filament associated protein 1-like 2, ANO9: anoctamin 9, MLLT3: myeloid/lymphoid or mixed-lineage leukemia translocated to 3, MPZL2: myelin protein zero-like 2, HHEX: hematopoietically expressed homeobox.

To identify cell type-specific genes, I performed differential expression analysis among all available cell types aiming to identify genes that change in any cell type (see Section 3.5.2.1). Cell type-specific genes can then be determined by filtering those that change in a single cell type. For the lineage-specific examples demonstrated below, I considered genes that are over-expressed during the lineage differentiation and are absent in all the other cell types.

3.2.4.1 *Cell type-specific expression of epigenetic factors*

Epigenetic modifications constitute an additional level of regulation of gene expression, as described in Chapter 1. The

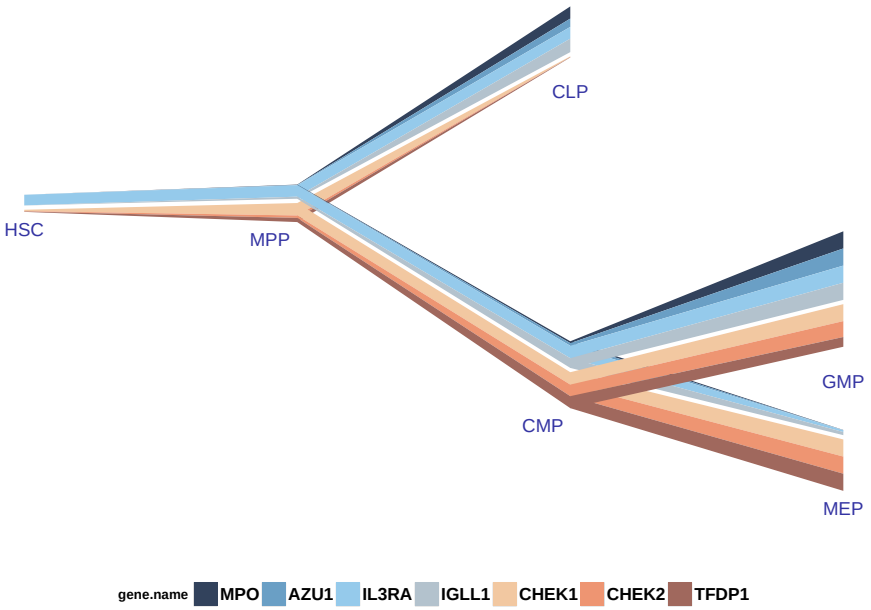


Figure 24: Shared gene expression patterns between lineages. Width corresponds to expression relative to the maximum expression of the gene across all cell types.
MPO: myeloperoxidase, AZU1: azurocidin 1, IL-3RA: interleukin 3 receptor, alpha, IGLL1: immunoglobulin lambda-like polypeptide 1, CHEK1: checkpoint kinase 1, CHEK2: checkpoint kinase 2, TFDP1: transcription factor Dp-1.

low number of haematopoietic progenitor cells makes direct epigenetic profiling impossible with current technologies. Therefore, it is of great interest to identify epigenetic regulators with distinct gene expression profiles among haematopoietic progenitors.

In Chapter 1, I described that hypomethylated regulatory regions are usually permissive for transcription factor binding and subsequent transcriptional activation, while hypermethylated regions are associated with epigenetic silencing. The resulting trascriptional repression is mediated by transcription factors that recognise methylated DNA and inhibit transcription via recruitment of co-repressors that modify chromatin (Klose and Bird, 2006). Three families of chromatin modifiers have been identified that mediate this process; the methyl-binding domain proteins, including MBD1, MBD2, MBD4 and MECP2; the zinc-

finger proteins KAISO (or ZBTB33), ZBTB38 and ZBTB4 (Filion et al., 2006; Prokhortchouk et al., 2001); and the SET-and-RING finger-associated proteins, including UHRF1 and UHRF2 (reviewed in (Clouaire and Stancheva, 2008; Prokhortchouk and Defossez, 2008; Sasai and Defossez, 2009)).

In my analysis, two of these chromatin modifiers were identified with distinct profiles in the haematopoietic progenitors; UHRF1 is expressed in all progenitor cells except the haematopoietic stem cell, while ZBTB38 is specific to common lymphoid progenitor.

In contrast to UHRF1, transcription of PRDM14 is specific to haematopoietic stem cells. PRDM14 is a member of the PRDM transcription factor family that possesses histone methyltransferase activity (Volkel and Angrand, 2007). Another member of this family, PRDM16, is required for maintenance of haematopoietic stem cells (Chuikov et al., 2010), while PRDM14 is highly expressed in haematopoietic malignancies (Dettman and Justice, 2008; Dettman et al., 2011) and has been extensively studied in embryonic stem cells, where it promotes self-renewal (Tsuneyoshi et al., 2008). In embryonic stem cells, PRDM14 binds DNA in a sequence-specific manner *in vitro*, inhibiting the expression of differentiating markers, while promoting the expression of genes involved in self-renewal (Ma et al., 2011). One of the targets of PRDM14 is DNMT3B, a DNA methyltransferase that functions in *de novo* methylation, and through PRDM14 embryonic stem cells inhibit the expression of DNMT3B, retaining a DNA hypomethylated state (Leitch et al., 2013). Taken together, these findings suggest that a similar DNA methylation program might be in place in haematopoietic stem cells to retain a permissive chromatin state.

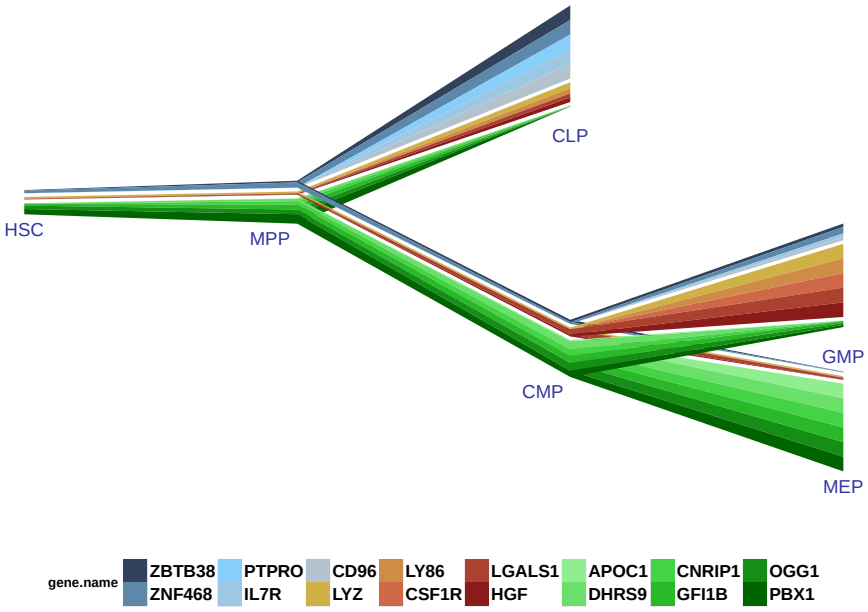


Figure 25: Gene expression patterns of cell type specific genes. Width corresponds to expression relative to the maximum expression of the gene across all cell types.
ZBTB38: zinc finger and BTB domain containing 38, ZNF468: zinc finger protein 468, PTPRO: protein tyrosine phosphatase receptor type O, IL7R: interleukin 7 receptor, CD96: CD96 molecule, LYZ: lysozome, CSF1R: colony stimulating factor 1 receptor, LGALS1: lectin galactoside-binding soluble 1, HGF: hepatocyte growth factor, APOC1: apolipoprotein C-I, DHRS9: dehydrogenase/reductase member 9, CNRIP1: cannabinoid receptor interacting protein 1, GFI1B: growth factor independent 1B transcription repressor, OGG1: 8-oxoguanine DNA glycosylase, PBX1: pre-B-cell leukemia homeobox 1.

3.2.4.2 *Lineage-specific gene expression*

The sets of lineage-specific genes identified contain both genes with known lineage-specific functions and genes with no known function in haematopoiesis (see Figure 25). The majority of the genes that are expressed exclusively in common lymphoid progenitors include adaptive immune response related genes, such as PTPRO (Aijo et al., 2012), CD96 (Fuchs et al., 2004) and IL-7R (Corcoran et al., 1996), and a zinc-finger transcription factor, ZNF468 that has not yet been studied within the lymphoid system. Similarly, for the granulocytes/monocyte progenitors, the genes showing cell type-specific expression are linked to innate immune responses, including LYZ (Kitaguchi

et al., 2009), LY86 (Miyake et al., 2000) and HGF (Nakamura et al., 1994), or CSF1R (Sherr, 1990), which is a receptor for colony stimulating factor (CSF), a cytokine that controls the production, differentiation, and function of macrophages.

In contrast, for most of the megakaryocyte/erythrocyte progenitor specific genes identified there is no evidence in the literature for any role within these two lineages. Only GFI1B has a known function in erythroid/megakaryocyte biology (Randrianarison-Huetz et al., 2010). Finally, within the stem cell compartment I also observed a set of lineage specific transcription factors expressed. These include important transcription regulators in megakaryocytes like ZFPM2 (Gieger et al., 2011) and NFATC4 (Muller et al., 2009), or SOX13 in granulocytes (Chambers et al., 2007). This observation is consistent with other studies, where lineage-specific factors were found to have dual roles in haematopoiesis; maintaining stem cells and promoting lineage-specific differentiation (Cai et al., 2012; Iwasaki et al., 2005).

3.2.4.3 *Shared expression of genes*

Another expression profile identified was that of genes that show similar expression profiles in multiple lineages. The first set consists of immune response related genes that are expressed in both lymphoid progenitors and granulocyte/monocyte progenitors, in accordance with the role of the mature cells of these two lineages in the immune system. The second set consists of three members of the cell cycle progression KEGG pathway, CHEK1, CHEK2 and TFDP1 (see Figure 24), which are specific to the myeloid branch. These findings are explained by the constant high demand for myeloid cells in the circulating blood, such as the two most abundant cells, platelets and erythrocytes, require a continuous proliferation of their precursors.

3.3 COMMON MYELOID PROGENITOR BREAKPOINT

A more in-depth analysis was performed on the CMP breakpoint where the presence of the full replicate set made this possible. Moreover, this breakpoint is of special interest to me because it contains the precursors of megakaryocytes and leads to the production of platelets. One of the main aims of this project was to identify changes in the expression profiles of the different cell types. Using RNA-sequencing data such changes can be identified at the gene and the transcript isoform level. The differential expression analysis at the transcript isoform level has become possible with RNA-sequencing that, for the first time, allows the identification, for example, of isoform switching.

3.3.1 *Gene-level differential expression analysis*

The tree structure of the haematopoietic system requires a more complex differential expression analysis that extends beyond the simple pair-wise comparison models for the identification of differentially expressed genes. I performed an analysis that aimed to capture genes that show a different expression pattern, irrespective of the type of change, in at least one of the cell types involved in the comparison (see Section 3.5.2.1). At the common myeloid progenitor breakpoint, after filtering for non-expressed genes (FPKM<1 in all three cell types), there were 750 genes identified as differentially expressed among the common myeloid, the megakaryocyte/erythrocyte and the granulocyte/monocyte progenitors. As mentioned previously, the analysis strategy described in this section can be applied to the whole dataset once this becomes available.

Of the 750 differentially expressed genes, four are epigenetic factors (FAM83D, ASH2L , CHAF1B, TFPT), five are splicing factors (DNAJC6, PRPF3, NSRP1, CTNNBL1, FAM50B) and thirty nine are transcription factors (see Figure 26). Among the

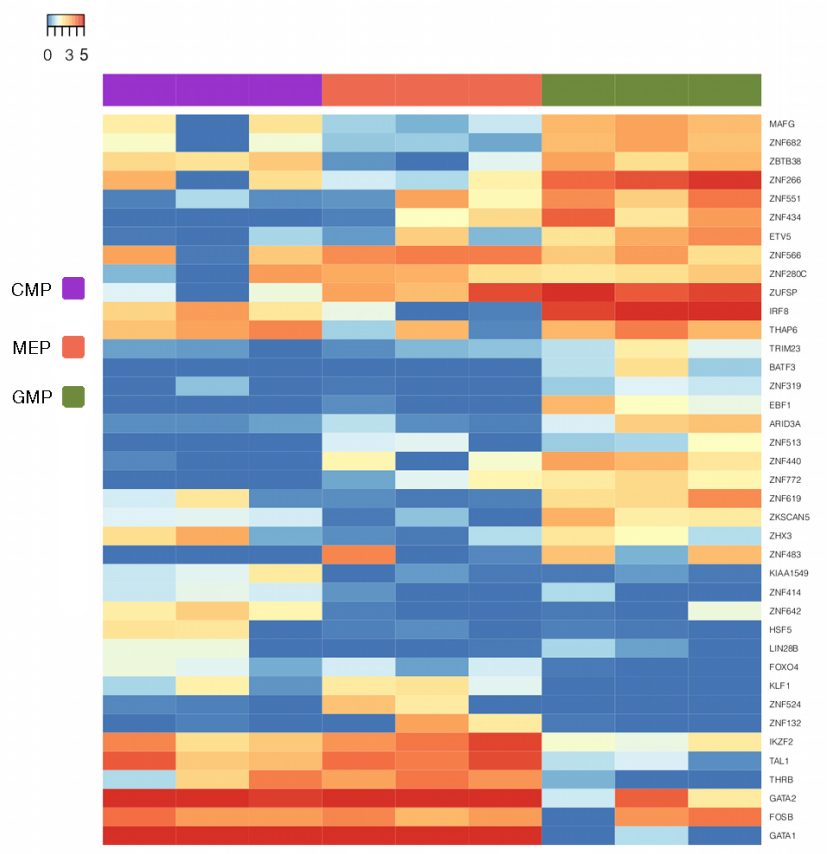


Figure 26: Heatmap of differentially expressed transcription factors in the CMP breakpoint. Genes were called differentially expressed if the posterior probability of the second model is greater than 0.5.

differentially expressed transcription factors identified, there are several known regulators of haematopoiesis, such as GATA1, GATA2, KLF1, TAL1 (see Section 1.2.4), but also numerous Kruppel C2H2-type zinc-finger proteins, with no function associated. To test for enriched biological functions, I examined for any over-represented gene ontology terms and pathways. The gene set is highly enriched for immune system terms and the haematopoiesis KEGG pathway (see Table 12).

The enriched GO terms and pathways identified include terms associated with immune system processes and functions, as well as haematopoietic cell lineage differentiation, which are consistent with the three blood cell types compared. Two of the differentially expressed genes of the enriched cytokine-cy-

tokine receptor signaling pathway are CSF1 and its receptor CSF1R (see Figure 27). CSF1 is a cytokine that controls the macrophage differentiation and function (reviewed in (Chitu and Stanley, 2006)). The significant over-expression in granulocyte/monocyte progenitors of its receptor (previously identified in Figure 25) is, therefore, consistent with its role. However, it is of interest that the cytokine itself is significantly up-regulated in common myeloid and megakaryocyte/erythrocyte precursors. In 1958, Bessis (1958) described that the maturation of erythrocytes occurs in islands in the presence of a macrophage that extends and surrounds the maturing red blood cells (Palis, 2004). Taken together, the significantly higher expression of CSF1 in megakaryocyte/erythrocyte precursors could be a signal for differentiating macrophages to home around the maturing blood erythroblasts enabling the normal release of mature red cells in the blood.

Table 12: Enriched Gene Ontology terms, KEGG and Reactome pathways among the differentially expressed genes at the CMP breakpoint. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg procedure. Only the top 5 terms for the Biological Function category are shown.

Category	Term	*P*-value	q-value
Biological Process	immune system process	6.03e-08	2.41e-04
Biological Process	immune response	2.72e-07	5.43e-04
Biological Process	regulation of immune system process	3.74e-06	4.22e-03
Biological Process	defense response to Gram-positive bacterium	4.22e-06	4.22e-03
Biological Process	myeloid leukocyte mediated immunity	8.86e-06	7.08e-03
Cellular Compartment	cell periphery	5.68e-08	7.93e-06
Cellular Compartment	plasma membrane	1.59e-07	2.42e-05
KEGG	Hematopoietic cell lineage	1.65e-08	2.65e-06
KEGG	Malaria	9.20e-05	7.41e-03
KEGG	Cytokine-cytokine receptor interaction	8.08e-04	0.043

3.3.2 *Transcript-level differential expression analysis*

As mentioned above, with the use of RNA-seq technology we can also examine differences at the transcript level. Using a similar approach to the differential expression analysis at

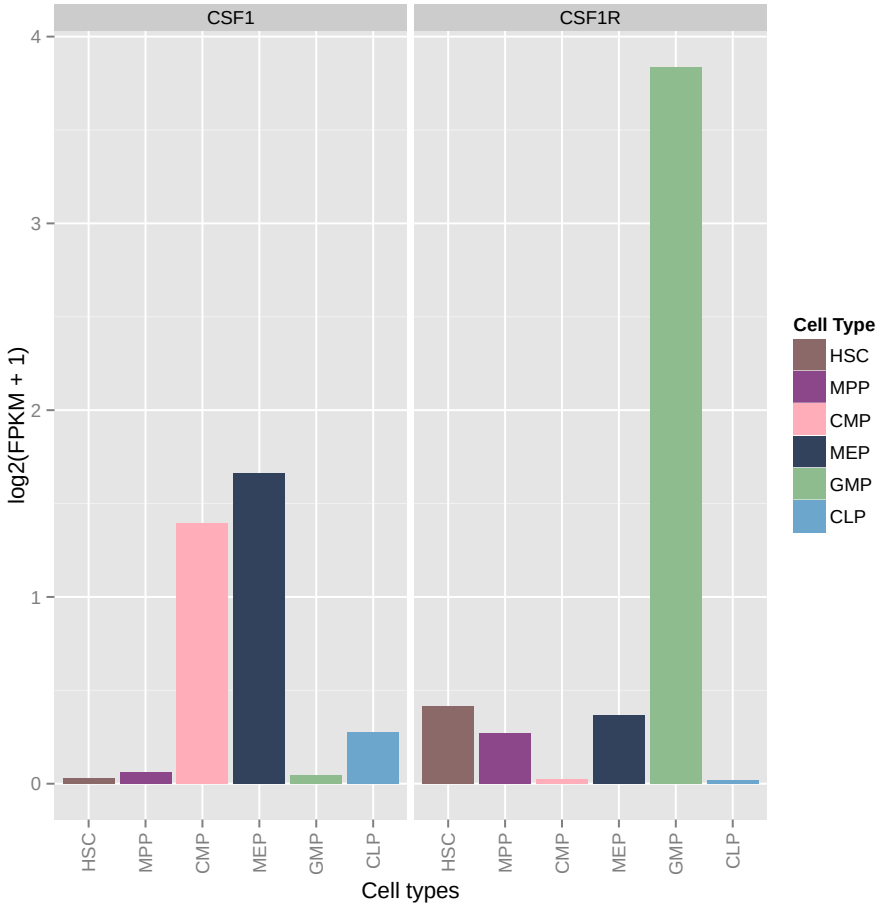


Figure 27: Gene expression of CSF1 and its receptor, CSF1R, in the Blueprint data.

the gene level described above (see Section 3.5.2.1), I identified 1,803 differentially expressed transcripts at the CMP breakpoint. Of these, 28 are epigenetic factors, 48 are splicing factors and 150 are transcription factors. Among the top 10 enriched Reactome pathways, three are related to the function of the immune system and haemostasis, which are in concordance with the cell types that were analysed (see Table 13). However, no cell type specific function is enriched in the top 10 GO terms. Other enriched Reactome and GO terms are related to metabolism. The enrichment analysis results differ from those obtained at the gene level, suggesting that cell type functions may be regulated at the gene level, whereas differential expression of

transcripts seems to be involved in the fine-scale tuning of cellular mechanisms, such as metabolism and RNA processing.

Table 13: Enriched Gene Ontology terms, KEGG and Reactome pathways among the differentially expressed transcripts at the CMP breakpoint. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg procedure. Only the top ten results are shown for the categories with more enriched terms.

Category	Term	*P*-value	q-value
Biological Process	cellular metabolic process	9.79e-13	5.42e-09
Biological Process	cellular macromolecule metabolic process	4.82e-12	1.33e-08
Biological Process	primary metabolic process	4.59e-11	8.46e-08
Biological Process	metabolic process	6.66e-11	9.21e-08
Biological Process	gene expression	8.70e-11	9.45e-08
Biological Process	RNA processing	4.82e-12	1.16e-06
Biological Process	nitrogen compound metabolic process	1.19e-08	8.24e-06
Biological Process	RNA metabolic process	3.59e-08	2.20e-05
Biological Process	cellular nitrogen compound metabolic process	4.66e-08	2.58e-05
Molecular Function	RNA binding	2.54e-09	3.67e-06
Molecular Function	binding	6.88e-07	4.973-04
Molecular Function	catalytic activity	1.91e-06	7.14e-04
Molecular Function	protein binding	1.98e-06	7.14e-04
Molecular Function	nucleoside phosphate binding	2.93e-06	7.31e-04
Molecular Function	organic cyclic compound binding	3.03e-06	7.31e-04
Molecular Function	nucleotide binding	3.97e-06	8.20e-04
Molecular Function	nucleic acid binding	3.62e-05	6.53e-03
Molecular Function	ATP binding	4.61e-05	7.40e-03
Molecular Function	small molecule binding	5.42e-05	7.62e-03
Cellular Compartment	intracellular	1.60e-30	1.18e-27
Cellular Compartment	intracellular part	2.41e-29	8.90e-27
Cellular Compartment	organelle	3.71e-23	9.12e-21
Cellular Compartment	membrane-bounded organelle	7.24e-23	1.34e-20
Cellular Compartment	intracellular organelle	1.09e-22	1.61e-20
Cellular Compartment	intracellular membrane-bounded organelle	2.39e-22	2.94e-20
Cellular Compartment	cytoplasm	2.08e-21	2.19e-19
Reactome	Gene Expression	3.34e-16	2.00e-13
Reactome	Metabolism	8.94e-14	2.68e-11
Reactome	Immune System	8.87e-12	1.77e-09
Reactome	Metabolism of proteins	2.15e-07	3.22e-05
Reactome	Disease	1.11e-06	1.33e-04
Reactome	Membrane Trafficking	1.63e-06	1.63e-04
Reactome	Signalling by NGF	9.11e-06	7.81e-04
Reactome	Cytokine Signaling in Immune system	1.23e-05	9.22e-04
Reactome	Adaptive Immune System	1.39e-05	9.28e-04
Reactome	Hemostasis	2.10e-05	1.26e-03
Reactome	Metabolism of lipids and lipoproteins	2.39e-05	1.31e-03
Reactome	Metabolism of carbohydrates	4.07e-05	1.99e-03
Reactome	tRNA Aminoacylation	4.31e-05	1.99e-03

3.3.3 *Isoform usage differential expression analysis*

In addition to identifying transcripts with a differential expression profile among the three cell types at the CMP breakpoint, differential expression analysis can also be performed for the proportion of transcript isoforms expressed for each gene. This provides a direct way to interrogate whether isoforms are being used in a different proportion independently of the expression level of the gene. I focused on differentially expressed protein coding transcripts, as a change in the produced protein isoform may imply a change in protein function. The function of proteins relies on their functional domains, therefore isoform switching is particularly interesting if there is a difference in protein domain between the differentially expressed transcript isoforms. Manual inspection of the differentially expressed protein coding transcript isoforms revealed two interesting examples in which differential splicing in each cell type encodes for proteins with alternative protein domains.

The first example is a zinc finger transcription factor, ZNF836, that has not been functionally characterised. The full-length protein isoform of ZNF836 contains two protein domains: one Kruppel domain at the N-terminus and a twenty four C2H2 zinc-finger domain at the C-terminus (see Figure 28(b)). At the transcript level the zinc-finger domain is encoded by the last exon, which is only expressed in the GMPs (see Figure 28(a)), while CMPs and MEPs express a shorter transcript that contains only the Kruppel domain.

The second interesting finding is GFI1B, a transcription factor belonging to the C2H2 zinc-finger family. It has a known regulatory role in megakaryotic and erythroid development and differentiation (Randrianarison-Huetz et al., 2010). In accordance with its function in haematopoiesis, GFI1B has already been identified as a MEP specific gene through my initial analysis

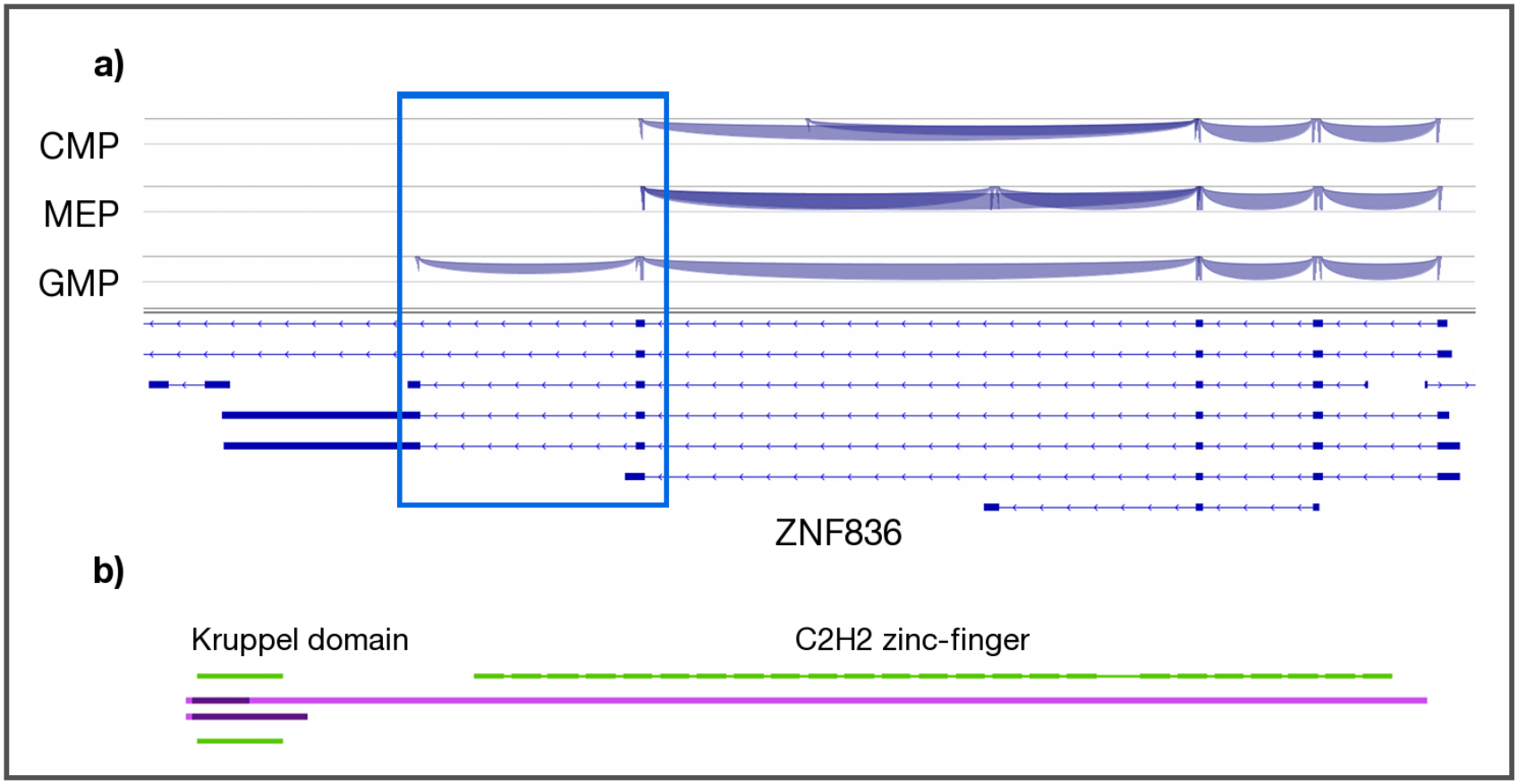


Figure 28: ZNF836 transcript and protein isoform expression in the CMP breakpoint. (a) Differential transcript isoform in the ZNF836 locus. All three cell types express the long isoform that contains all the annotated exons. However, CMPs express a shorter isofom too, that contains an exon skipping event (highlighted in the blue rectangle). CMPs may also express a novel transcript isoform too based on the novel splice junction observed at their 3′ end. (b) The shorter isoform contains two C2H2 zinc-finger proteins less than the longer one.

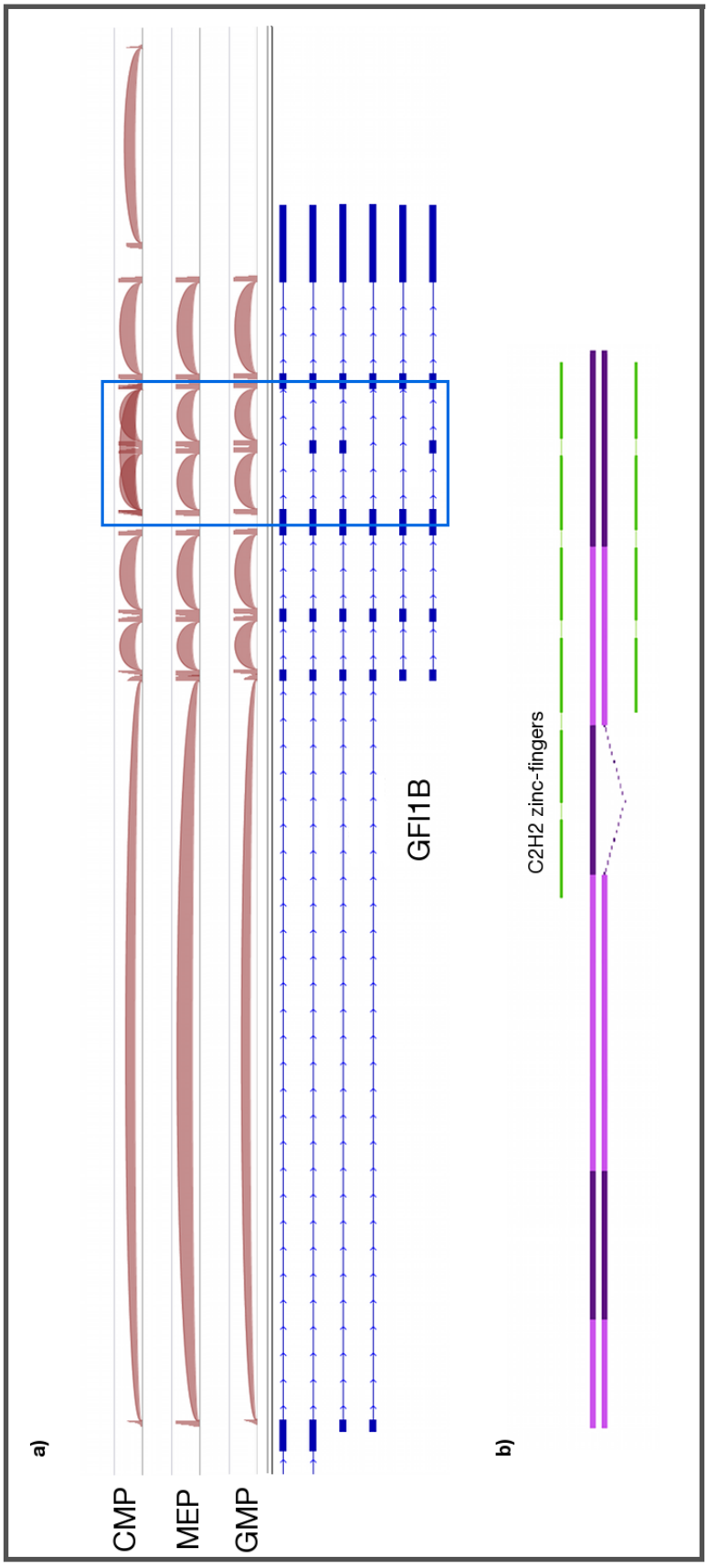


Figure 29: GFI1B transcript and protein isoform expression in the CMP breakpoint. (a) Differential transcript isoform in the GFI1B locus. All three cell types express the long isoform that contains all the annotated exons. However, CMPs express a shorter isofom too, that contains an exon skipping event (highlighted in the blue rectangle). CMPs may also express a novel transcript isoform too based on the novel splice junction observed at their 3′ end. (b) The shorter isoform contains two C2H2 zinc-finger proteins less than the longer one.

(see Section 3.2.4.2) and as a gene down-regulated in GMPs according to the polytomous analysis (see Section 3.3.4). At the transcript level, all three cell types express the full-length transcript isoform that encodes for a protein that contains six C2H2-type zinc-finger domains (see Figure 29(b)). However, CMPs also express a shorter isoform that skips one of the exons contained in the full-length transcript (see Figure 29(a)). This exon skipping event results in the production of a protein isoform that contains only four C2H2-type zinc-finger domains.

The changes in transcript isoform expression among the different cell types may reflect changes in protein structure and function, especially if these changes result in protein domain disruptions as in the examples described above. Functional assays are required, however, to establish if these differentially expressed transcript isoforms are translated into proteins and what is the role of the different resulting protein isoforms in haematopoiesis.

3.3.4 *Cell type-specific gene and transcript isoform expression*

The differential expression analysis described above defines which genes and transcript isoforms change among the different cell types of the CMP breakpoint. However, it is essential to define the directionality of these changes and to identify clusters of genes that show similar gene expression profiles. To determine sets of distinct gene expression patterns, I performed a Bayesian polytomous analysis. The principle behind polytomous analysis is the simultaneous estimation of the posterior probabilities of all possible models among the three cell types. Therefore, apart from the general differential expression analysis described above, I performed three additional comparisons, where each of the three cell types was compared to the other two. The input to the polytomous analysis was the posterior probabilities of five models, of which model-0 con-

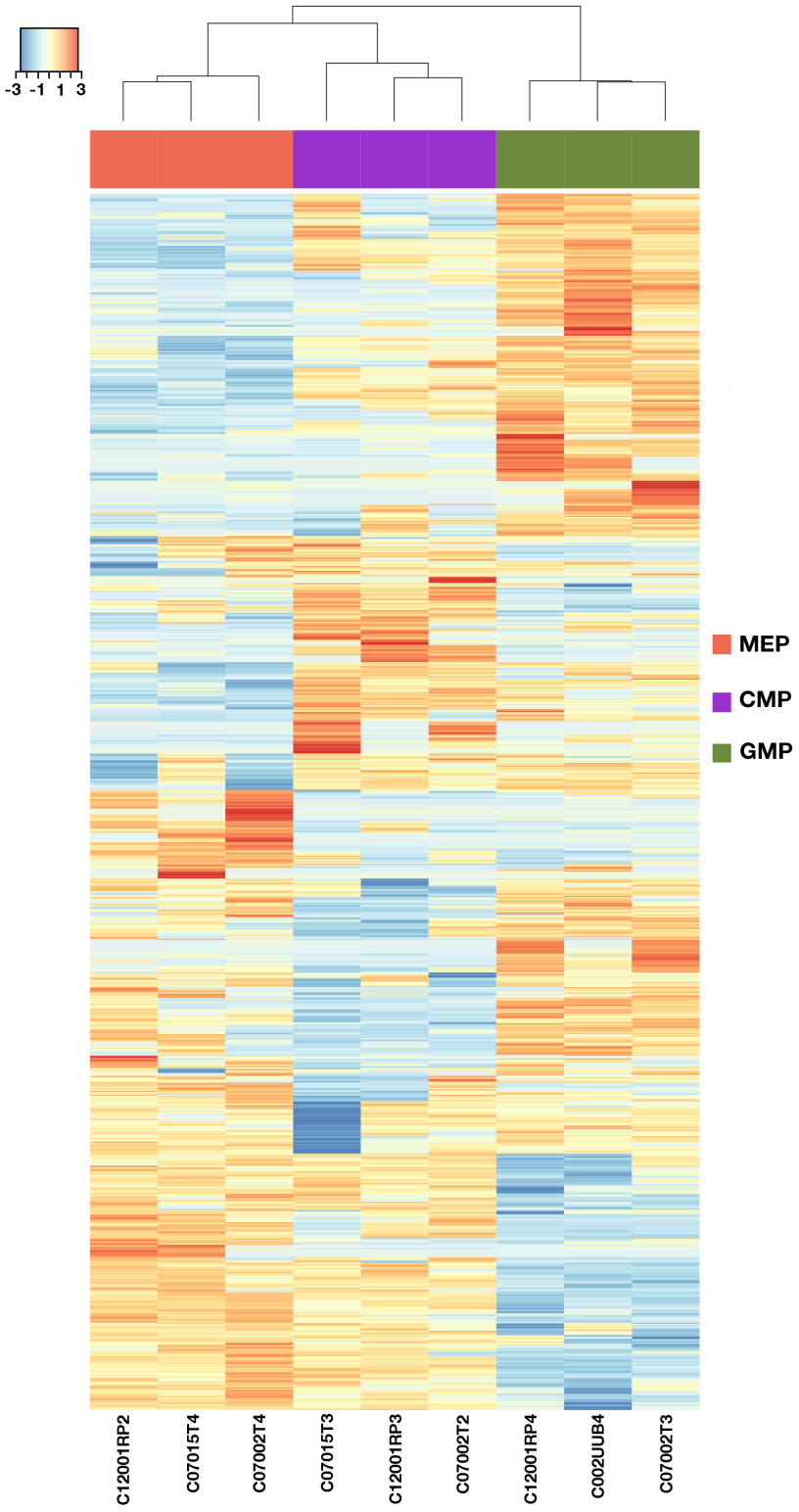


Figure 30: Heatmap of normalised expression of the cell type specific genes identified through the polytomous analysis.

stitutes the baseline model, where no expression changes are observed, while models 2, 3 and 4 describe cell type-specific changes. For instance, model-2 includes genes that are differentially expressed in CMPs compared to both MEPs and GMPs (see Section 3.5.2.1 and Figure 34).

The polytomous analysis computes a posterior probability for each of the models per gene (see Figure 35). The majority of the genes do not change among the three cell types, having a higher posterior probability for model-0. By filtering out these genes and performing a K-means clustering on the model posterior probabilities of the remaining ones, I initially identified three clusters of genes, one for each cell type of the CMP breakpoint (see Figure 30). Then based on the normalised expression values of the genes, each cluster is split into two sub-clusters, one for the up-regulated and one for the down-regulated genes within each cell type. The same analysis was performed at the transcript level (see Figure 31). The numbers of genes and transcripts in each of the six sub-clusters are summarised in Table 14.

Granulocyte/monocyte progenitors show a higher number of cell type specific genes and transcripts within the common myeloid progenitor breakpoint compared to the other two cell types, with the majority of these genes being up-regulated. This observation is in accordance with the isolation strategy, in which common myeloid progenitors and megakaryocyte/erythrocyte progenitors differ only at the level of surface expression of IL-3Ra, and with the notion that common myeloid progenitors are required to differentiate more frequently in the megakaryocyte/erythrocyte lineage. Also there is evidence that granulocyte/monocyte progenitors are a mixed population, originating from both common lymphoid primed progenitors and common myeloid progenitors (Adolfsson et al., 2005; Gorgens et al., 2013).

Table 14: Identification of clusters of cell type specific genes and transcripts in the CMP breakpoint. Genes and transcripts were divided into three clusters using a K-means algorithm (K=3), firstly on the posterior probabilities of the five models to identify the cell type specific features and then on the normalised expression values of the genes and transcripts.

	CMP		MEP		GMP	
	UP	DOWN	UP	DOWN	UP	DOWN
genes	80	190	78	198	189	171
	Total #: 906					
transcripts	133	443	127	534	549	374
	Total #: 2,162					

A gene ontology and pathway enrichment analysis in each of the six sub-clusters (see Table 15) revealed the following:

1. Enrichment of ion binding related terms for the genes that are down-regulated in the CMPs, which is consistent for example with the iron uptake function of red blood cells, one of the products of the megakaryocyte/erythrocyte lineage.

2. Cell cycle related genes are enriched among the up-regulated MEP specific genes, in accordance with the much higher demand for platelet and erythrocyte production and consequent higher proliferation of their progenitors.

3. Genes related to the immune response were enriched both among the down-regulated genes in MEPs and up-regulated genes in GMPs.

4. Genes related to the megakaryocytic and erythroid lineages were enriched among the down-regulated genes specific to GMPs.

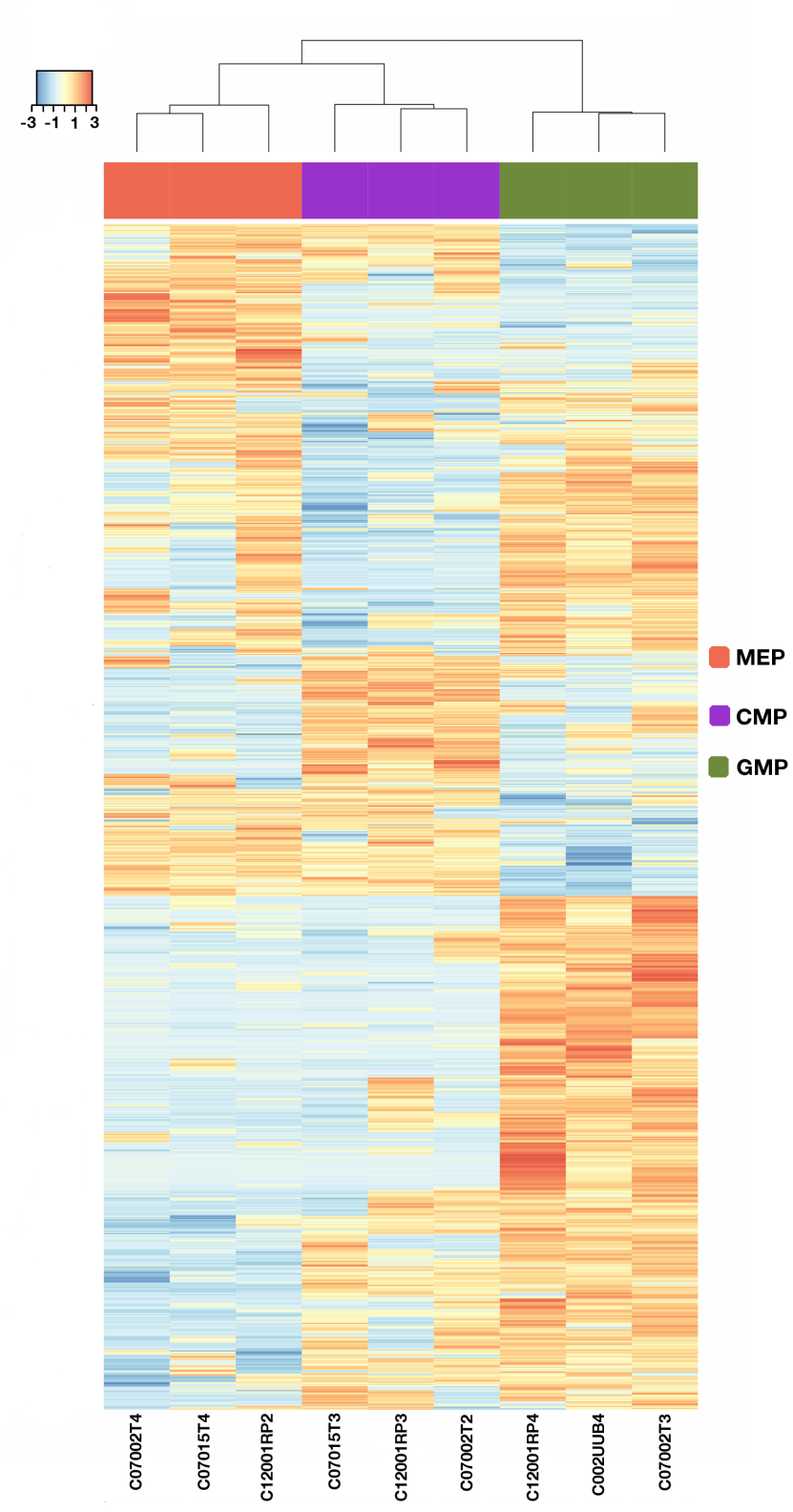


Figure 31: Heatmap of standardised expression of the cell type specific transcript isoforms identified through polytomous analysis.

3.3.5 *Novel splice junctions*

In addition to the analysis described above on the annotated features of the human genome, we can also identify novel splice junctions based on the aligned reads onto the human genome using a spliced aligner (see Section 3.5.2.2). Identification of novel splice junctions was performed in collaboration with Dr Lu Chen, Department of Haematology, University of Cambridge. Alignments were only considered if at least 10 bp of the read had been aligned on either side of the splice junction and to limit any false positive splice junctions, the set of splice junctions were further filtered to retain those that were supported by a minimum of ten reads.

Table 15: Enriched Gene Ontology terms, KEGG and Reactome pathways among the six clusters of cell type specific genes identified at the CMP breakpoint. For categories with multiple terms, only the five most significant terms for each category are shown. *P*-values have been corrected for multiple testing using the Benjamini-Hochberg procedure.

		Category	Term	p-value	q-value
CMP specific	UP	Cellular Compartment	tight junction	4.95e-04	4.65e-02
		KEGG	Tight junction	2.49e-03	4.99e-02
		KEGG	Cell adhesion molecules (CAMs)	2.57e-03	4.99e-02
	DOWN	Molecular Function	ion binding	3.20e-05	1.29e-02
		Molecular Function	cation binding	7.19e-05	1.45e-02
		Molecular Function	zinc ion binding	2.01e-04	2.70e-02
		Molecular Function	metal ion binding	2.72e-04	2.74e-02
		Molecular Function	transition metal ion binding	4.31e-04	3.47e-02
MEP specific	UP	KEGG	Systemic lupus erythematosus	1.57e-17	5.34e-16
		Reactome	Nucleosome assembly	4.83e-14	1.76e-12
		Reactome	Cell Cycle	7.37e-04	8.07e-03
	DOWN	Biological Process	immune system process	6.18e-10	1.55e-06
		Biological Process	immune response	1.50e-09	1.88e-06
		Biological Process	leukocyte migration	1.90e-07	1.58e-04
		Biological Process	regulation of immune system process	1.85e-06	1.16e-03
		Biological Process	defense response	5.72e-06	2.82e-03
		Biological Process	leukocyte chemotaxis	9.31e-06	3.33e-03
		Biological Process	cell activation	1.24e-05	3.87e-03
		Biological Process	interferon-gamma-mediated signaling pathway	1.95e-05	4.91e-03
		Biological Process	immune effector process	2.22e-09	4.91e-03
		Biological Process	cellular component movement	2.29e-05	4.91e-03
		Biological Process	cell migration	2.35e-05	4.91e-03
		Molecular Function	phosphatidylinositol 3-kinase binding	2.57e-05	1.25e-02
		Molecular Function	cytokine receptor binding	1.11e-04	1.80e-02
		Molecular Function	CD8 receptor binding	1.11e-04	1.80e-02
		Molecular Function	2'-5'-oligoadenylate synthetase activity	3.31e-04	3.22e-02
		Cellular Compartment	plasma membrane part	2.51e-04	3.03e-02
		Cellular Compartment	cell periphery	2.55e-04	3.03e-02
		Cellular Compartment	actin cytoskeleton	4.48e-04	3.03e-02
		Biological Process	immune response	1.51e-07	3.44e-04
		Biological Process	innate immune response	2.88e-07	3.44e-04
		Biological Process	defense response	1.36e-06	1.08e-03

Continued on next page

UP

Table 15 – *Continued from previous page*

		Category	Term	*P*-value	q-value
	DOWN	Biological Process	response to cytokine stimulus	8.70e-06	4.15e-03
		Biological Process	cellular response to organic substance	1.37e-05	5.46e-03
		Biological Process	cellular response to interferon-gamma	1.90e-05	5.68e-03
		Biological Process	cytokine-mediated signaling pathway	4.42e-05	1.17e-02
		Biological Process	interferon-gamma-mediated signaling pathway	5.11e-05	1.22e-02
		Biological Process	signal transduction	8.80e-05	1.50e-02
		Cellular Compartment	cell periphery	3.59e-05	5.03e-03
		Cellular Compartment	membrane part	7.33e-04	3.30e-02
		Cellular Compartment	ruffle	7.50e-04	3.30e-02
		Cellular Compartment	focal adhesion	8.14e-04	3.30e-02
		Cellular Compartment	actin filament	8.72e-04	3.30e-02
		Cellular Compartment	cell-substrate adherens junction	9.95e-04	3.30e-02
		Cellular Compartment	cell surface	1.00e-03	3.30e-02
		Biological Process	immune system development	6.54e-07	1.67e-03
		Biological Process	hematopoietic or lymphoid organ development	4.61e-06	5.87e-03
		Biological Process	immune system process	9.69e-06	8.07e-03
		Biological Process	myeloid cell differentiation	1.27e-05	8.07e-03
		Biological Process	hemopoiesis	2.83e-05	1.44e-02
		Biological Process	erythrocyte development	6.11e-05	2.33e-02
		Biological Process	megakaryocyte differentiation	8.90e-05	2.33e-02
		Biological Process	platelet formation	9.13e-05	2.33e-02
		Biological Process	regulation of biological quality	1.10e-04	2.33e-02
		Biological Process	basophil differentiation	1.10e-04	2.33e-02
		Biological Process	eosinophil fate commitment	1.10e-04	2.33e-02
		Biological Process	embryonic hemopoiesis	1.25e-04	2.45e-02
		Biological Process	single-organism developmental process	2.31e-04	4.20e-02
		Biological Process	unsaturated fatty acid biosynthetic process	2.74e-04	4.65e-02
		Molecular Function	enhancer sequence-specific DNA binding	5.50e-06	2.95e-03
		Molecular Function	enhancer binding	9.14e-05	3.10e-03
		Molecular Function	glutathione transferase activity	9.14e-04	1.63e-02
		Molecular Function	identical protein binding	1.58e-04	2.12e-02
		Molecular Function	phosphotransferase activity, nitrogenous group as acceptor	2.69e-04	2.71e-02
		Molecular Function	1-phosphatidylinositol-5-phosphate 4-kinase activity	3.03e-04	2.71e-02
		Molecular Function	protein dimerization activity	3.76e-04	2.88e-02
		Molecular Function	creatine kinase activity	6.01e-04	4.03e-02
		Cellular Compartment	cytoplasm	1.90e-04	1.70e-02
		Cellular Compartment	mitochondrial matrix	5.74e-04	3.36e-02
		Cellular Compartment	mitochondrion	1.00e-03	4.90e-02
		KEGG	Systemic lupus erythematosus	3.05e-07	3.75e-05
		Reactome	Factors involved in megakaryocyte development and platelet production	9.98e-05	2.64e-02
		Reactome	Meiotic Recombination	2.81e-04	2.65e-02

In total, there were 101,496 splice junctions that fulfilled the above criteria. To determine whether any of these splice junctions are novel, he compared them to annotated splice sites from Ensembl v70. The majority of the splice junctions are already annotated (see Table 16). The novel splicing events were then classified based on whether either of their ends is already annotated (partially novel), not annotated or the splicing pattern was novel between known splicing sites (novel splicing pattern). Only few of these belonged to the latter category, while the ma-

jority of the novel splice junctions are either novel or partially novel.

Table 16: The Table shows the number of splice junctions present at the three cell types of the CMP breakpoint, as these were defined by the spliced alignments of the RNA-seq dataset. In total there are 109,729 splice sites that are spanned by at least 10 reads in at least two of the three replicates per cell type. The majority of these splice junctions is annotated in Ensembl v70. The novel splice junctions were classified into three categories: *novel*: none of the ends of the splice junction is annotated, *partially novel*: one of the ends of the splice junction is annotated, *novel splicing pattern*: the splice junction is novel, but its ends are annotated.

Class	Junctions	Percentage
Annotated	101,496	92.50
Novel	1,626	1.48
Partially novel	3,822	3.48
Novel splicing pattern	416	0.25
Total: 109,729		

3.4 CONCLUSIONS

The production of blood cells is a tightly regulated process that includes consecutive intermediate stages of increasingly restricted differentiating potential. A part of the BLUEPRINT project mapping phase, focused on a set of six rare haematopoietic progenitors that were isolated from cord blood and subjected to transcriptome profiling using RNA sequencing. In contrast to previous gene expression studies performed on these cell types (summarised in Section 3.2.1), the datasets presented here constitute the first genome-wide analysis of the poly(A)$^+$ transcriptome of these cell types.

The goal of this project was to catalogue and analyse the transcripts expressed in these rare cells by building an analysis pipeline that will allow for the identification of cell type and lineage specific expression at two different levels: genes

and transcripts. The structure of the haematopoietic system requires a more sophisticated differential expression analysis model that extends beyond simple pair-wise comparisons. Using the common myeloid, megakaryocyte/erythrocyte and granulocyte/monocyte progenitors I first identified changes in any of the three cell types and then classified the directionality of the changes. The latter was achieved using a Bayesian polytomous analysis that does not rely on pre-defined thresholds for the capture of different scenarios.

In addition to known regulators of haematopoiesis, the analysis presented in this chapter identified a host of genes not currently associated with blood development. Also highlighted is a further level of regulation for those closely related cell types. Genes are not only turned activated or repressed, but isoform switching preferentially occurs with protein domains being included or excluded in transcripts. Follow-up experiments are required to examine whether these candidate genes or transcript isoforms play an important role in blood development.

3.5 MATERIAL AND METHODS

The cord blood collection and all wet-lab experiments described below have been performed by members of Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK and Prof Henk Stunnenberg's group, Nijmegen Centre of Molecular Life Sciences, Nijmegen, The Netherlands and are only provided for completeness. I was responsible for the computational analysis of the data sets produced and have acknowledged any other person involved in the data analysis process.

3.5.1 *Cord blood collection and progenitors cell isolation*

Cord blood was collected with informed consent (REC 12/EE/0040) at the Rosie Hospital, Cambridge University Hospitals and processed within 18 hours. $CD34^+$ progenitors were isolated from cord blood using the EasySep progenitor cells enrichment kit with platelet depletion (STEMCELL cat. 19356) according to the manufacturer instructions. Isolated cells from up to 3 cord blood units were pooled and then sorted on an ARIA III flow cytometer in six different populations (defined as shown in Table 11), using the following antibodies (for an example of cell sorting and gate setting see Figure 32):

1. Lin^- PECy5 (CD2 #BD 555328 20µl / test; CD3 #BD 561006 20µl / test; CD11b #BD 561686 20µl / test; CD14 #Invitrogen MHCD1418 5µl / test; CD19 #BD 555414 20µl / test; CD56 #BD 555517 20µl / test; GPA #NHSBT 1/100)
2. CD38 FITC (#BD 560982 20µl / test)
3. CD34 APC (#BD 560940 20µl / test)
4. CD90 PECy7 (#BD 561558 5µl / test)
5. CD45RA Pacific Blue (#Invitrogen MHCD45RA28 5µl / test)
6. CD123/IL-3Ra PerCP (#BD 560904 20µl / test)
7. CD10 PE (#BD 561002 20µl / test)

Sorted cells (20000 to 200000; with purity > 99%) were centrifuged, resuspended in Trizol (Invitrogen) and stored at -80 degrees Celsius. Cell isolation and purification was performed in Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge.

RNA-seq libraries were prepared using the SMARTer Ultra Low RNA and Advantage 2 PCR kit both from Clontech,

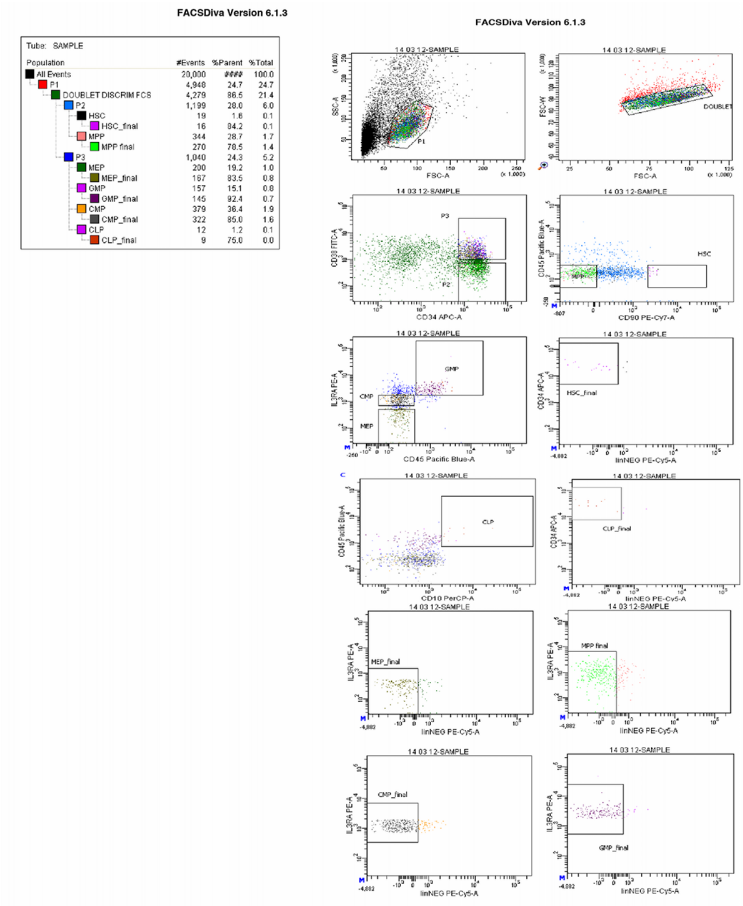


Figure 32: Example of progenitor cell sorting and gate setting. Gates were established using a fluorescence minus one strategy.

following manufacturer instructions using 100pg of total RNA as input for each sample. Samples were indexed using NEXTflex adapters (Bioo Scientific) and paired-end sequencing was performed on a HiSeq 2000 machine using TruSeq reagents (Illumina) according to manufacturer's instructions.

3.5.2 *Bioinformatic analysis workflow*

Large projects like this one require a well-thought analysis pipeline. Any analysis, however, depends on the questions and hypothesis that have generated such data. The project's aim are to create a catalogue of transcripts expressed in the early human haematopoietic progenitors, and to identify changes in gene expression occurring during the first stages of blood form-

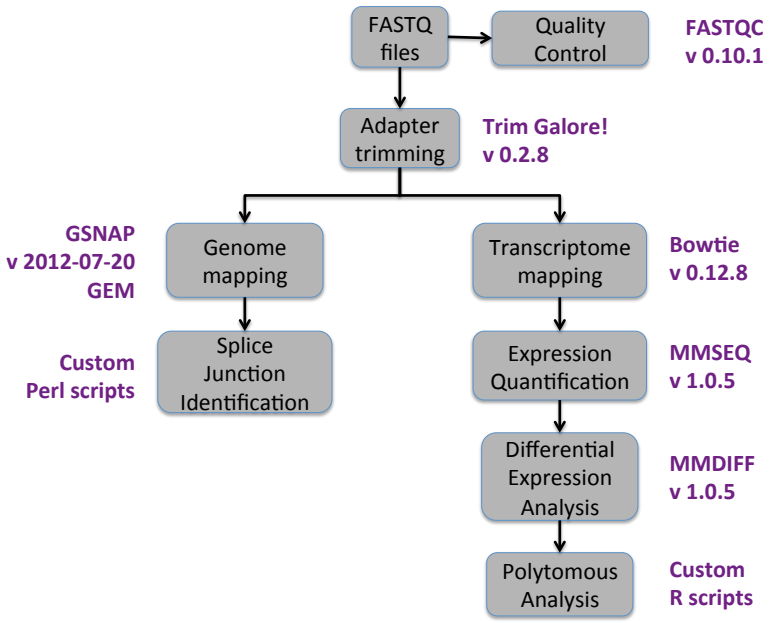


Figure 33: Workflow of the bioinformatics analysis performed on the RNA-seq data. The tools used in each step and their versions are listed on the side.

ation. Therefore, the first step was to decide on which tools to use to achieve such goals.

To address these questions, we built two separate analysis workflows that would focus on the already annotated transcripts and on the identification of novel transcript isoforms. A graphic overview of this pipeline can be found in Figure 33. The pipeline design was a result of extensive discussions between members of Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, including Augusto Rendon, Lu Chen, Ernest Turro, and members of the EMBL - European Bioinformatics Institute, including Myrto Kostadima, Remco Loos from Dr Paul Bertone's lab and David Richardson from Dr Paul Flicek's lab.

For both pipelines, the FASTQ files were trimmed first for the sequencing adapters and then for the SMARTer kit adapters

using Trim Galore! v 0.2.8. Adapter trimming was performed by Pawan Poudel, Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge.

3.5.2.1 *Transcriptome mapping*

The trimmed reads were aligned to the Ensembl v70 human transcriptome using Bowtie 0.12.8 by Pawan Poudel. The number of reads aligned to the human transcriptome are summarised on Table 17. I then used the MMSEQ v1.0.5 (Turro et al., 2011) analysis pipeline to quantify gene and transcript isoform expression.

Table 17: Percentage of reads aligning to the human transcriptome Ensembl v70 using Bowtie v0.12.8.

	Sample ID	Total number of reads	Number of reads mapped	Percentage	Number of unmapped reads	Percentage
HSC	C002UUB1	172,481,943	73,755,811	(42.76%)	98,723,265	(57.24%)
MPP	C07015T1	181,799,291	139,333,441	(76.64%)	42,464,044	(23.36%)
	C002UUB2	169,991,894	98,124,196	(57.72%)	71,866,287	(42.28%)
CLP	C002UUB3	172,009,365	94,999,977	(55.23%)	77,003,407	(44.77%)
CMP	C07015T3	184,806,802	148,709,976	(80.47%)	36,096,081	(19.53%)
	C07002T2	195,410,930	143,030,270	(73.19%)	52,379,152	(26.80%)
	C12001RP3	167,136,700	127,126,407	(76.06%)	40,008,538	(23.94%)
GMP	C07002T3	171,032,184	100,611,206	(58.83%)	70,417,647	(41.17%)
	C002UUB4	176,049,028	106,984,181	(60.77%)	69,060,982	(39.23%)
	C12001RP4	137,506,194	87,205,573	(63.42%)	50,297,372	(36.58%)
MEP	C07002T4	195,846,106	150,294,464	(76.74%)	45549504	(23.26%)
	C12001RP2	178,155,896	112,904,383	(63.37%)	65,245,035	(36.62%)
	C07015T4	156,581,333	126,241,633	(80.62%)	30,338,778	(19.38%)

Differential expression analysis was performed using MMDIFF v 1.0.5, that allows for a flexible model comparison, which is ideal for the tree structure of the haematopoietic system. MMDIFF was used for the analysis described in Sections 3.2.4, 3.3.1, 3.3.2 and 3.3.3.

In Section 3.2.4, differential expression analysis was performed using all available progenitors RNA-seq samples. The two models compared were:

1. gene expression in all cell types is similar
2. gene expression is different in at least one cell type

The prior probability that the second model is true was set to 0.1. Differentially expressed genes were then selected if they were expressed in at least one cell type (FPKM > 1) and the posterior probability of the second model was greater than 0.5.

Similarly, MMDIFF was used at the CMP breakpoint to identify differentially expressed genes (Section 3.3.1), transcripts (Section 3.3.2) and isoform usage (Section 3.3.3). The second model used in these comparisons is animated in Figure 34, designated model-1.

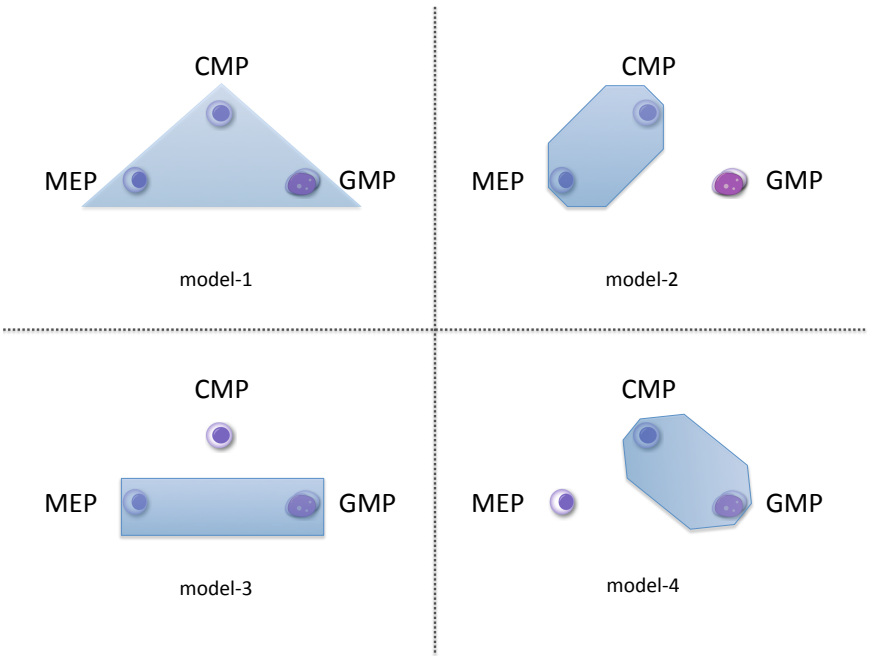


Figure 34: Polytomous analysis alternative models at the CMP breakpoint. The baseline model (model-0) represents features in which expression does not change.

The analysis described above was used to identify changes in expression, it does not however identify the type of changes. To determine the directionality of changes, I performed polytomous model selection that compares multiple models. The baseline model (model-0) was always the same and assumed that expression is similar in all cell types. The alternative models considered at the CMP breakpoint are shown in Figure 34.

Model-1 is described above, and the other models compared any two cell types to the third.

Then assuming a prior probability of 0.5 for model-0 and 0.125 for each of the other models, I computed the posterior probability of each model for each feature. I filtered any features that were not expressed in at least one cell type and features that showed a higher posterior probability for model-0 (meaning that they do not change expression among the different cell types), I used a K-means (K=3; one for each cell type) clustering on the posterior probabilities of the five models for the remaining features, aiming to identify cell type specific changes. To identify whether genes were up- or down-regulated, I performed a second K-means clustering (K=2) on the normalised expression values of the features of each of the three sets identified by the previous clustering step.

3.5.2.2 *Genome mapping*

Identification of novel splicing events cannot be achieved using the alignments onto the transcriptome. For this purpose, the trimmed reads were aligned to the *Homo sapiens* high coverage assembly (hg19) (release February 2009) using GSNAP version 2012-07-20 (Wu and Nacu, 2010) (step performed in collaboration with Dr Lu Chen, Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge) and GEM (Marco-Sola et al., 2012). The latter alignment step was done in Dr Roderic Guigo's group. For the GSNAP alignment, read trimming was disabled, a maximum of 7 mismatches were allowed and novel splicing sites were limited to a maximum of 300,000 bp apart. Following this step, in collaboration with Dr Lu Chen, custom Perl scripts were used to identify the splice junctions that were present in each sample. Splice junctions that were not supported by both aligners and by less than ten reads were removed.

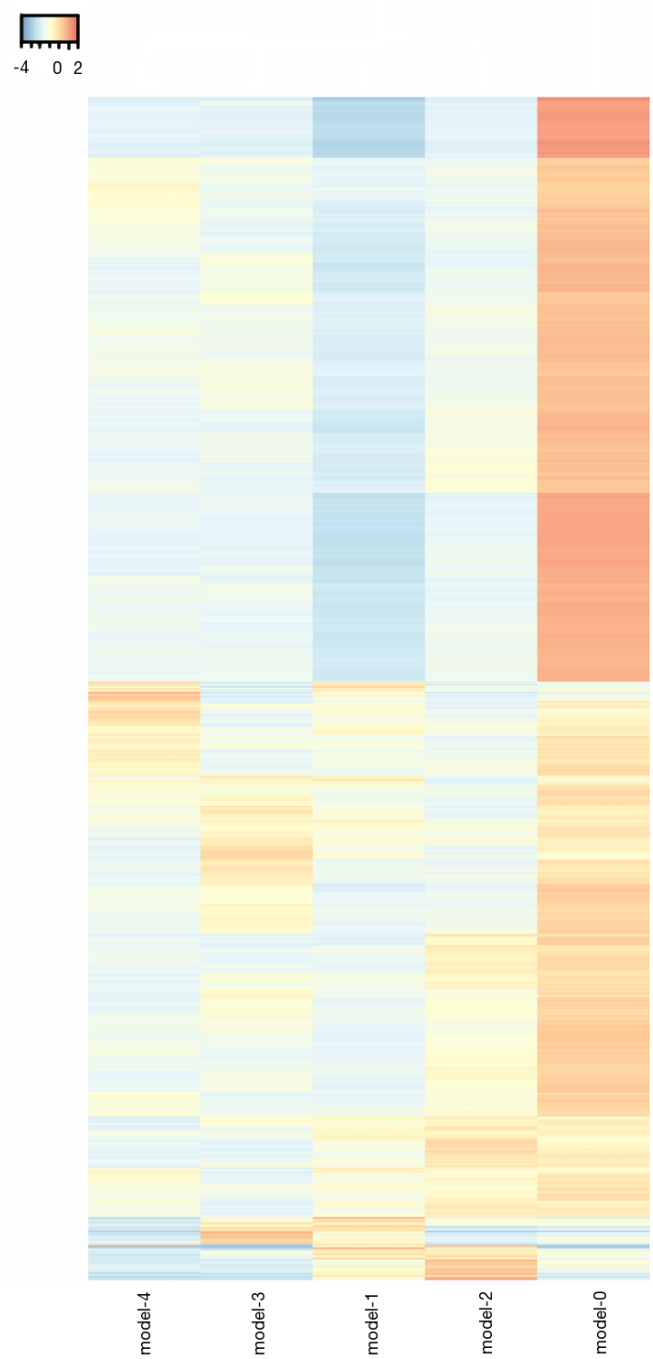


Figure 35: Posterior probability of the five different models studied in the polytomous analysis at the CMP breakpoint.

4

THE ROLE OF NFIB IN MEGAKARYOPOIESIS

4.1 INTRODUCTION

In Chapter 2, I focused on the MEIS1 transcription factor and its binding profile in human megakaryocytes. Integrative analysis of MEIS1 binding data with data of other key transcriptional regulators of haematopoiesis showed that these transcription factors were frequently co-localising, suggesting that MEIS1 is highly integrated in the gene regulatory circuit in megakaryocytes and providing evidence that other factors might be involved.

The aim of the following analysis was to identify other transcription factors that may be involved in the regulation of megakaryopoiesis. To do so, I analysed the gene expression profiles of all human transcription factors across early haematopoietic progenitors and megakaryocytes, integrating the RNA-seq datasets presented in the previous chapters. My analysis focused only on those human transcription factors for which experimental evidence of regulatory function existed (Vaquerizas et al., 2009). Candidate transcription factors were selected if they were found to be expressed at higher levels in megakaryocytes compared to their precursor cells, the megakaryocyte/erythrocyte progenitors (MEPs). In addition, I required that these transcription factors were lowly or not expressed in erythrocytes.

The set of transcription factors that fulfill the above-mentioned criteria includes: BCL6B, DLX1, IRX3, MAFB, MEIS3P2, MSC, MYCN, NFIB, RCOR2, SALL2, SIX5, TEAD4, TFEB, VDR, ZFPM2,

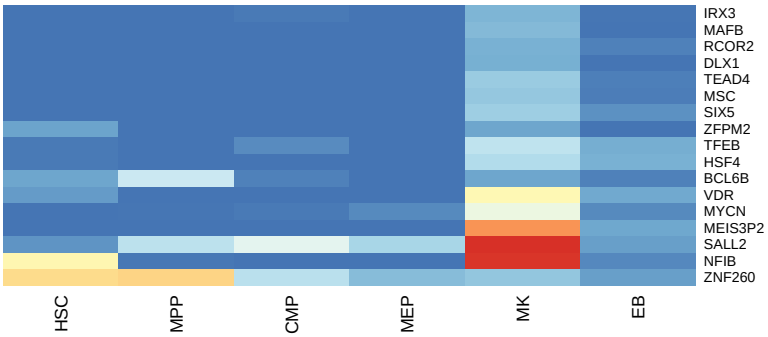
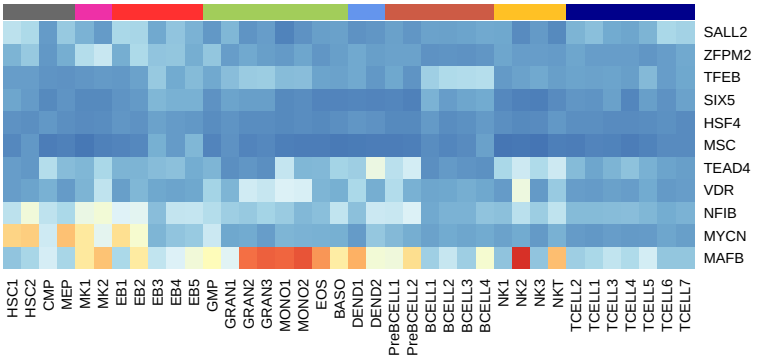


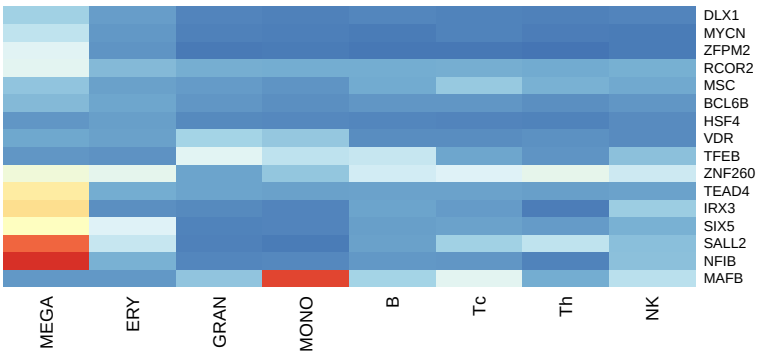
Figure 36: Gene expression of candidate transcription factors across various haematopoietic cell types. Expression data for early haematopoietic progenitors come from the BLUEPRINT dataset (see Chapter 3 for details), while the megakaryocytic and erythrocytic RNA-seq libraries were produced independently in Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK (see Chapter 2 and section 4.4.3, respectively).
HSC; haematopoetic stem cell, MPP: multipotent progenitor, CMP: common myeloid progenitor, MEP: megakaryocyte/erythrocyte progenitor, MK: megakaryocyte, and EB: erythrocyte

and ZNF260. With the exception of five genes, the rest show a rather low expression in megakaryocytes, but still over the threshold of 1 FPKM (see Figure 36). This gene set was further filtered for genes that have low or no expression in other mature blood cell types. Given that the BLUEPRINT dataset, described in Chapter 3, did not encompass mature blood cell types, I referred to publicly available microarray studies, such as the DMAP (Novershtern et al., 2011) and HaemAtlas (Watkins et al., 2009) datasets, for the expression of these transcription factors (see Figure 37).

Upon examination of the expression values of the candidate transcription factors, I discarded three from further analysis because they showed no noticeable difference in gene expression across the different cell types; these are BCL6B, HSF4, and MSC. Four more transcription factors (MAFB, TFEB, VDR and ZNF260) were dropped from the list because their expression profiles showed higher expression in other cell types. For example, consistent with its role in monocytic differentiation (Kelly



(a)



(b)

Figure 37: Gene expression of 16 candidate transcription factors in: (a) the DMAP dataset (Novershtern et al., 2011), and (b) the HaemAtlas dataset (Watkins et al., 2009).

et al., 2000), MAFB is highly expressed in monocytes and granulocytes.

The remaining nine transcription factors showed a higher expression in megakaryocytes, however for only two of these (NFIB and SALL2) there was an agreement among our RNA-seq datasets and the HaemAtlas indicating that their expression values were significantly higher (see Figures 36 and 37(b)). However, this finding was not supported for either of the genes in the DMAP dataset 37(b), where expression of NFIB and SALL2 was low in megakaryocytes. Such discrepancies could be attributed to the difference in maturation stages of the mega-

karyocytes profiled by each study, an observation discussed previously in Chapter 2.

To summarise, my analysis identified NFIB and SALL2 as the most promising candidates for megakaryocyte specific transcription factor. To my knowledge, neither gene has been studied within the context of hematopoiesis. SALL2 is a C2H2-type zinc finger protein (Sweetman and Munsterberg, 2006) that plays a significant role during the development of the neural tube (Bohm et al., 2008). NFIB is a member of the nuclear factor I transcription family. Members of this family have already been identified as potential MEIS1 co-binding transcription factors in megakaryocytes. These findings prompted me to look further into the nuclear factor I proteins and their potential role in megakaryopoiesis.

4.2 THE NUCLEAR FACTOR I TRANSCRIPTION FACTOR FAMILY

The nuclear factor I (NFI) family of genes was initially discovered in a study investigating adenovirus DNA replication (Nagata et al., 1982). Host NFI was shown to recruit the viral DNA polymerase onto the origin of replication, mediating the formation of the DNA replication pre-initiation complex (Armentero et al., 1994; Chen et al., 1990; de Jong and van der Vliet, 1999). Later studies revealed that the NFI family of proteins is identical to the sequence-specific CCAAT-box binding transcription factor (Jones et al., 1987; Santoro et al., 1988) that had already been identified as an important transcription initiation protein (Benoist et al., 1980; Efstratiadis et al., 1980; Morgan et al., 1987). Therefore, several distinct functions have been ascribed to NFI proteins. Even within the context of regulation of gene expression, different NFI protein products exhibit different roles, either activating or repressing transcription (reviewed in (Gronostajski, 2000)).

The NFI family consists of four family members: NFIA, NFIB, NFIC and NFIX (Kruse et al., 1991; Rupp et al., 1990), which are encoded by four paralogous genes in mammals (Qian et al., 1995). Analysis of individual gene knockout mice have revealed a range of phenotypes (see Table 18). Both NFIA and NFIB deficient mice die around birth. NFIA mutants suffer from a range of neurological deficits (das Neves et al., 1999; Lu et al., 2007; Shu et al., 2003), while loss of NFIB results in abnormal lung maturation. In addition to these acute phenotypes, mice lacking NFIB also exhibit nervous system defects (Steele-Perkins et al., 2005). On the contrary, NFIC or NFIX knockout do not cause lethality. NFIX deficient mice show abnormal brain development (Chaudhry et al., 1997), while NFIC mutants exhibit tooth defects (Steele-Perkins et al., 2003). Absence of a reported platelet phenotype is not surprising, as NFIB is not expressed in mouse megakaryocytes or platelets based on publicly available RNA-seq data (Rowley et al., 2011).

Table 18: Phenotypes from knockout studies of nucelar factor I proteins in mice.

Protein	Knock-out phenotype	References
NFIA	perinatal lethality	(das Neves et al., 1999; Shu et al., 2003)
	neurological deficits	(Lu et al., 2007)
NFIB	perinatal lethality	(Koehler et al., 2010)
	neurological deficits	
NFIC	tooth defects	(Steele-Perkins et al., 2003)
NFIX	brain development abnormalities	(Campbell et al., 2008)

The NFI proteins usually contain two domains located at the N- and C-termini of the proteins (reviewed in (Gronostajski, 2000), see Figure 38(a)). The N-terminal DNA-binding and dimerisation domain is highly conserved in all NFI proteins and species. The functions assigned to this domain include site-specific DNA binding, dimerisation with NFI proteins and interaction with adenovirus DNA polymerase (Gounari et al., 1990; Mermod et al., 1989). The C-terminal part of the DNA-binding and dimerisation domain contains four cysteine residues con-

served among all NFI proteins that are sufficient for low-affinity DNA binding, while the N-terminal part is a basic alpha helix domain that increases the affinity of the DNA binding (Dekker et al., 1996). NFI proteins bind to DNA either as hetero- or homodimers to the dyad symmetric consensus sequence TTGGC(N5)GCCAA with high affinity (Hennighausen et al., 1985; Leegwater et al., 1985; Nowock et al., 1985), or to consensus half sites with reduced affinity (Meisterernst et al., 1988).

The C-terminal transactivation and repression domain of NFI proteins is highly variable in amino acid sequence and rich in proline residues (proline rich domain) (Paonessa et al., 1988; Santoro et al., 1988). The functions ascribed to this domain are transactivation or repression of expression and inhibition of DNA binding (Mermod et al., 1989; Roulet et al., 1995). The proposed transactivation mechanisms include interaction with basal transcription factors under the effect of growth factors, direct displacement of repressive histones or interaction with co-activators (Alevizopoulos et al., 1995). Similarly, repressive regulation of gene expression can occur through direct competition with other transcription factors or attachment to heterologous DNA-binding domains (Gronostajski, 2000).

4.2.1 *Gene expression of Nuclear Factor I proteins in haematopoiesis*

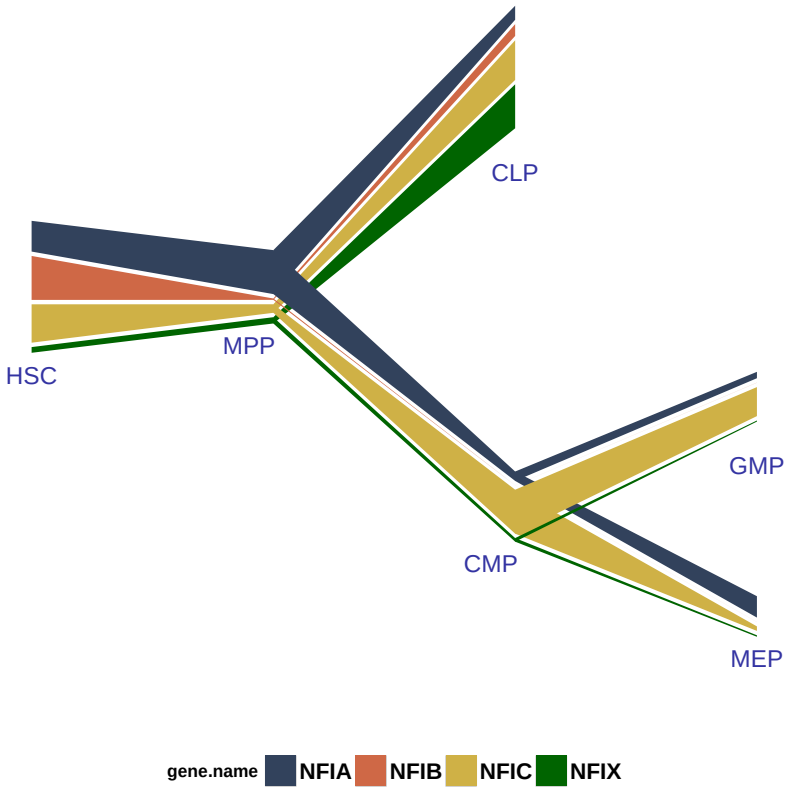
Of the four members of the NFI gene family, NFIA and recently NFIX have been studied within the context of haematopoiesis. NFIA over-expression in haematopoietic progenitors favours the cells differentiation into the erythroid rather than the granulocytic lineage (Starnes et al., 2010). NFIX deficiency has been shown to increase apoptosis of murine stem and progenitor cells (Holmfeldt et al., 2013).

To determine which members of the NFI family may have a role in haematopoiesis, I examined the gene expression pro-

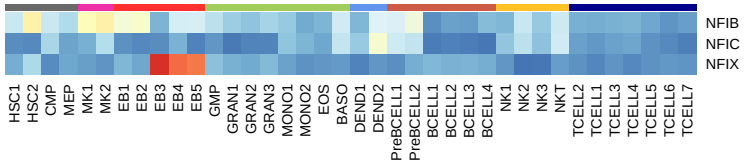
Figure 38: NFI protein domain structure and alternative splicing.
(a) Graphical representation of the general splicing of NFIs and their protein domains (adapted from (Gronostajski, 2000)). NFI proteins are encoded in 11 exons in all species. Exon 2 encodes the largest part of the N-terminal DNA-binding and dimerisation domain. This can be further split into a basic alpha helix loop and a domain, containing four cysteine residues conserved in all NFI proteins and species. Exons 3-11 encode the transactivation or repression domain.
(b) Graphical representation of NFIB and NFIC mRNAs, as these were identified by the megakaryocytic RNA-seq data set. Both NFIB transcript isoforms lack the DNA-binding and dimerisation domain, while NFIC contains both. The addition of exon 12 in the NFIB transcript does not affect the C-terminal domain. The alternative splicing, however, between exons 8-12, encodes a shorter domain.

files of all NFIs in expression datasets of human haematopoetic cells (HaemAtlas, DMAP and BLUEPRINT, see Figure 39). The expression pattern of NFIA is consistent with its role in erythroid differentiation, as NFIA is highly expressed only in erythrocytes among the mature blood cells and its expression decreases in the granulocyte monocyte progenitors. Its expression is also high in the multipotent progenitors.

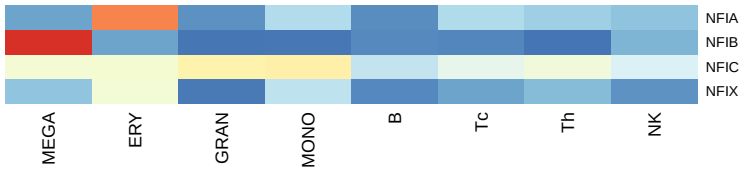
NFIX is also specific to erythrocytes, but rather the late stages of erythroid differentiation (see Figure 39(b)), while its expression is moderate in the progenitors with the exception of common lymphoid progenitors (CLP), where it has the highest expression among the progenitor cell types. Although not a subject of this analysis, it would be interesting to examine if NFIX, along with NFIA, plays a role in erythroid differentiation. NFIC does not have any cell type specific expression. It is expressed in almost all haematopoietic progenitors and mature cells. The



(a)



(b)



(c)

Figure 39: Gene expression of all four members of the NFI gene family in the: (a) BLUEPRINT, (b) DMAP (Novershtern et al., 2011), and (c) HaemAtlas (Watkins et al., 2009) data sets.

only significant change is observed in granulocytes and monocytes, where its expression is slightly higher (see Figure 39(c)).

NFIB is also expressed in the stem cell compartment other than in the megakaryocytes. It has already been shown that NFIB plays a role in stem cell proliferation and/or differentiation within the epithelial-melanocyte compartment (Chang et al., 2013), and in neuronal development (Namihira et al., 2009). Expression of NFIB in haematopoietic stem cells suggests that it might have a similar role in this stem cell compartment as well. It has generally been observed that certain lineage-specific transcription factors are also expressed in the haematopoetic stem cells, either due to lineage priming (Hu et al., 1997) or their dual roles in haematopoiesis (Cai et al., 2012).

To conclude, three of four members of the NFI gene family have a cell type specific expression among mature blood cell types: NFIB in megakaryocytes and NFIA and NFIX in erythrocytes. However, NFIB is not the only member of the family expressed in megakaryocytes, as NFIC is also expressed at similar levels.

4.2.2 *NFIB and NFIC isoform expression in human megakaryocytes*

Several splice variants have been reported for the NFI genes, introducing further diversity in the family (Paonessa et al., 1988; Santoro et al., 1988). To determine the transcript isoforms expressed in megakaryocytes, I examined the RNA-seq data introduced in Chapter 2. As described in Section 2.3.2, I had already identified a novel transcription start site for NFIB that is marked by a MEIS1 binding site. As shown in Figure 40, the novel isoform is the major one, as there is no read coverage over the first two annotated exons. To verify the presence of the novel transcription start site in the NFIB locus, I looked for a compatible chromatin signature. Consistent with the RNA-seq data, publicly available FAIRE-seq data (Paul et al., 2013) show a peak over the transcription start site marking a region of open chromatin that is easily accessible (see Figure 40). The incom-

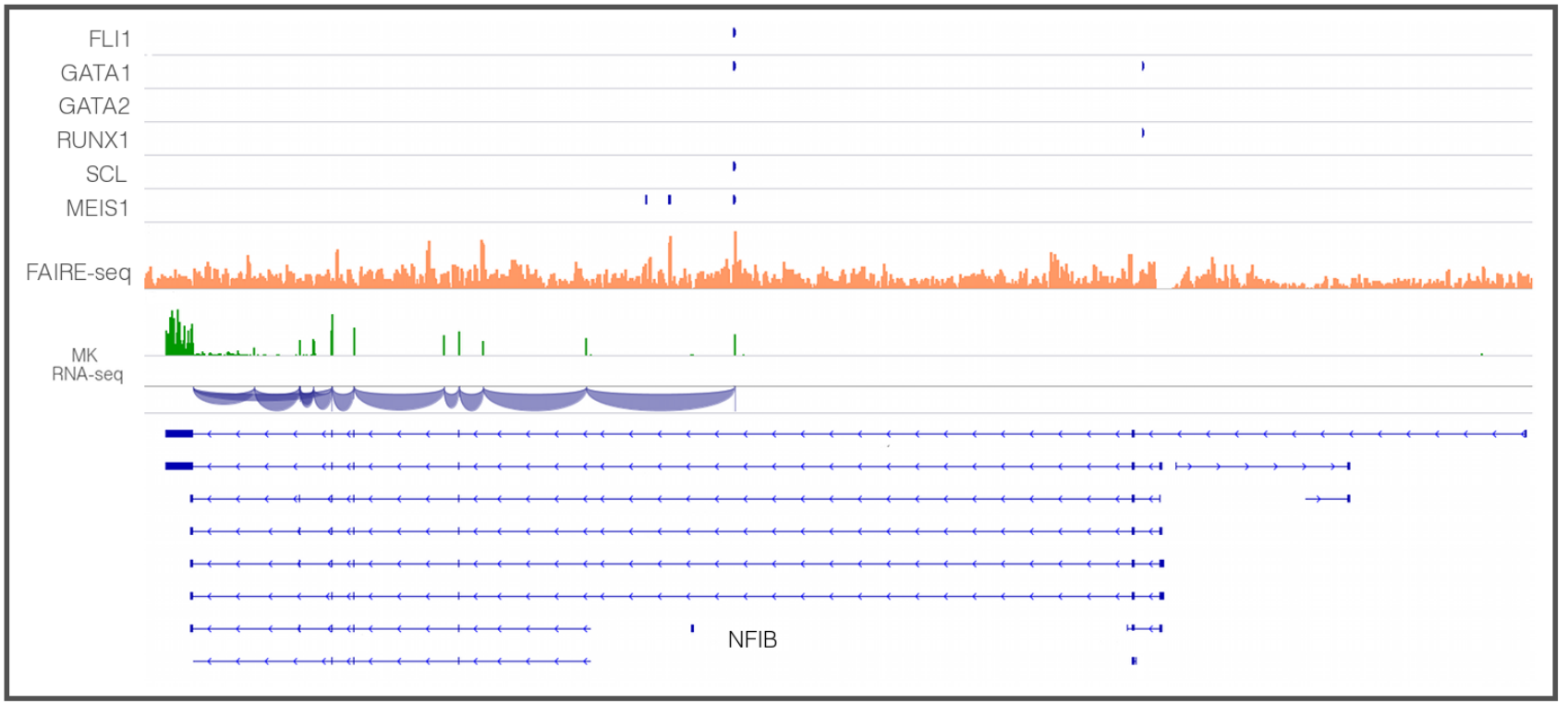


Figure 40: NFIB expression in megakaryocytes. Analysis of the RNA-seq data suggests that the megakaryocytic NFIB gene contains a novel transcription start site that is also marked by a binding site where MEIS1, FLI1, GATA1 and SCL co-localise. No read coverage is observed over the first two annotated exons. In addition, there are two different splicing events at the 3′ end of the gene suggesting more than one transcript isoform is expressed.

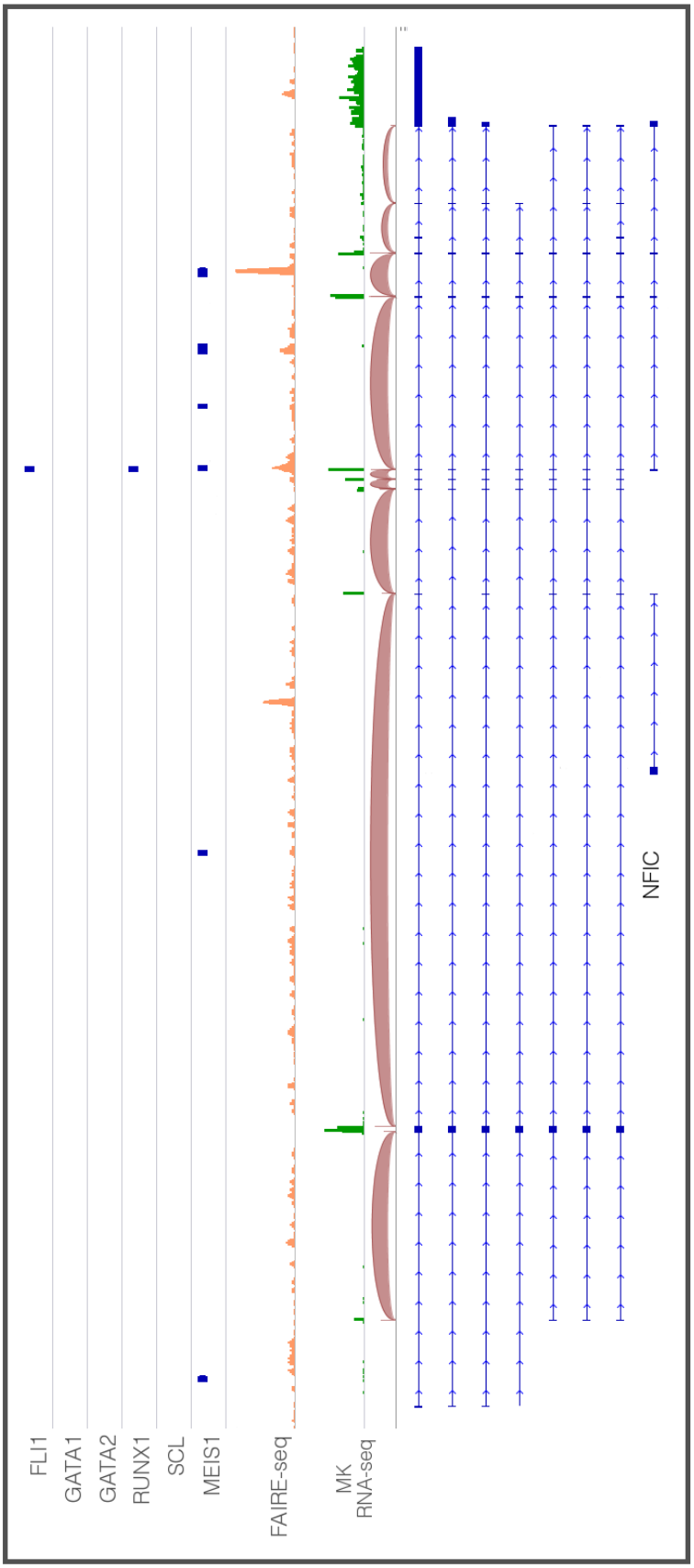


Figure 41: NFIC expression in the megakaryocytes. The RNA-seq data suggests that from the NFIC locus one full-length isoform is transcribed. However, the existence of a high FAIRE-seq peak within the NFIC gene body that is also bound by MEIS1, FLI1 and RUNX1 could suggest that a second shorter transcript isoform is expressed. Subsequent western blot experiments identified only one protein isoform expressed in megakaryocytes (the longest - see Figure 43.

patible splice junctions observed close to the 3′ end of the NFIB locus also suggest that there is more than one splice variant expressed. In the case of NFIC, there is one isoform expressed that exhibits previously annotated splicing events.

Protein domains are distinct functional units and they are usually responsible for the function of a protein. It is therefore essential to identify the domains encoded by the NFIB and NFIC transcriptional isoforms. As described above, NFI proteins usually consist of two domains with distinct functions. Compared to the general structure of the NFI mRNAs (reviewed in (Gronostajski, 2000) and shown in Figure 38(a)), NFIB transcripts usually contain an additional exon at the 3′ end (denoted exon 12 in Figure 38(b)). This exon, however, does not seem to interfere with the transactivation or repression domain. The megakaryocytic NFIB mRNAs lack the DNA-binding and dimerisation domain, which is encoded by exon 2 that is not transcribed in human megakaryocytes. This finding suggests that NFIB lacks the ability to bind directly to DNA and dimerise with other NFI proteins. Differential 3′ end splicing gives rise to variable transactivation and repression domains that may confer altered functional specificity. The NFIC expressed transcript contains both annotated domains. Further functional assays are needed to verify whether these proteins dimerise or bind to the DNA.

To validate the novel NFIB transcript isoform and its transcription start site, Frances Burden (Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK) performed a 5′ race PCR (see section 4.4.4). The PCR products were then sequenced and aligned to the human genome (see Figure 42). The novel transcription start site was validated. Moreover several other products were identified, suggesting that even shorter transcript isoforms of NFIB may be expressed in megakaryocytes. These shorter transcriptional iso-

forms could not be identified in RNA-seq data, as they overlap annotated exons.

4.2.3 *NFIB protein expression in human megakaryocytes*

Immunoblotting analysis was performed by Dr Mattia Frontini (Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK) using megakaryocytes protein extracts to detect protein level of the two NFIs in human megakaryocytes (see section 4.4.6). The western blot confirmed that there are several NFIB protein isoforms in human megakaryocytes, whereas only one protein isoform was detected for NFIC (see Figure 43). All of these are consistent with our RNA-seq and 5' race data.

4.3 DISCUSSION

Integrative analysis of the transcriptome data from haematopoietic progenitors with the megakaryocytic transcriptome data revealed novel potential transcription factors regulating megakaryopoiesis. One of these candidates is Nuclear Factor I/B (NFIB). Investigation of other datasets that contain gene expression profiles of mature blood cells showed that NFIB is a transcription factor specific to megakaryocytes. These analyses prompted further biochemical characterisation of this transcription factor.

NFIB is a member of the nuclear factor I family. Of the four family members, NFIB and NFIC are expressed in megakaryocytes. Our megakaryocyte RNA-seq data showed that one of the isoforms transcribed in the NFIB locus in megakaryocytes is novel. In addition, MEIS1 binds in the proximity of the unannotated transcription start site along with other transcription factors. By western blot analysis, we were able to show that the detected isoforms of NFIB and NFIC were present as protein

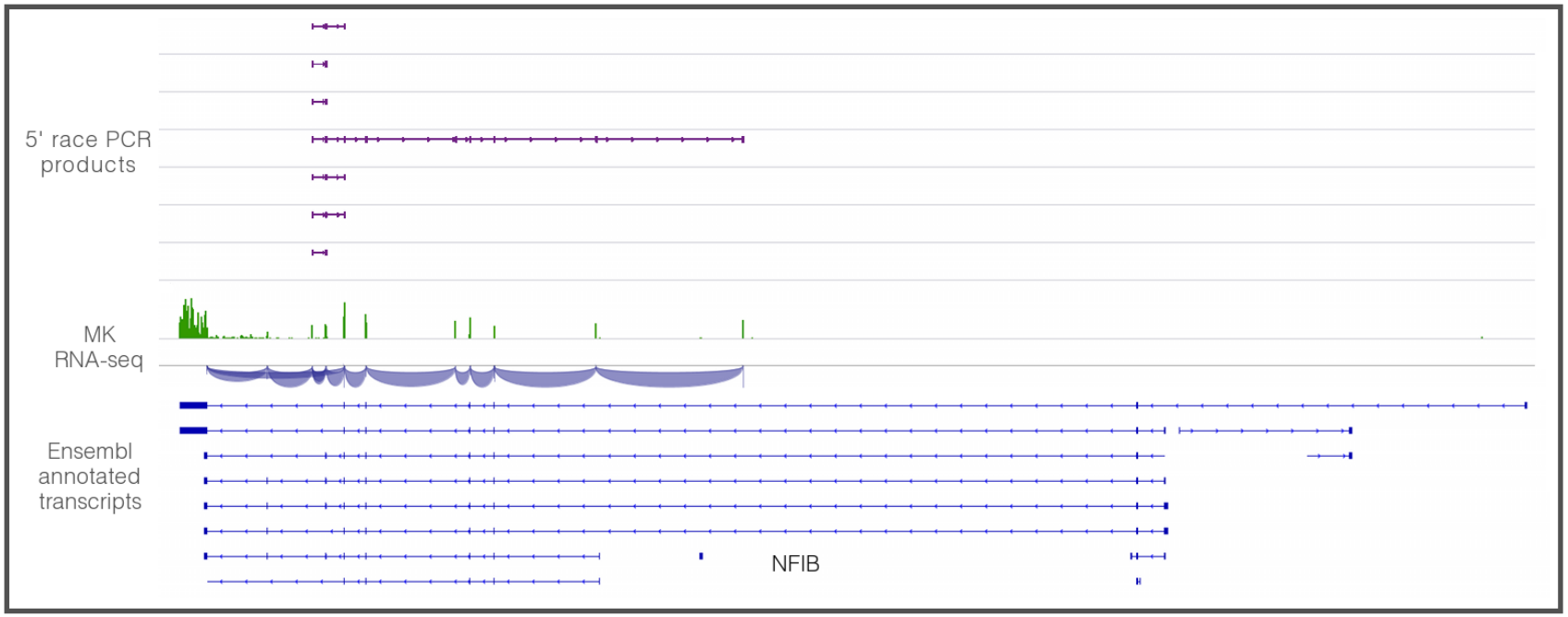


Figure 42: Detection of the NFIB isoforms transcription start site through 5′ race PCR. The analysis confirmd the novel transcription start site identified using the RNA-seq data set. Multiple products also suggest the expression of a much shorter transcript isoform that could have not been identified using only the RNA-seq data because all of its exons overlap with the longer one.

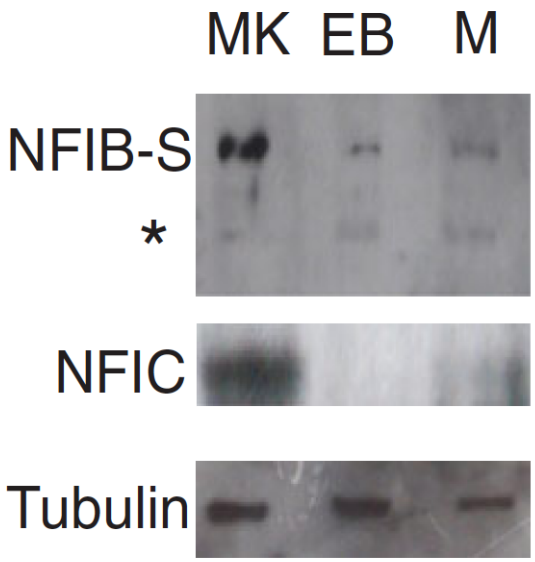


Figure 43: Detection of NFIB and NFIC proteins in human megakaryocytes. Western blot experiment using antibodies against NFIB and NFIC confirm the expression of the two gene products in megakaryocytes. NFIB expresses multiple protein isoforms in megakaryocytes, (around 40, 35 and 32 kDa), however, none of these represents the full-length isoform, which would be expected at around 70 kDa. NFIC expresses only one (full-length) variant. .

in megakaryocytes. Consistent with our RNA-seq and 5′ race data, the western blot analysis showed more than one protein isoform for NFIB.

The diversity of NFI products and functions have already been highlighted in various studies (Alevizopoulos et al., 1995; Mermod et al., 1989; Roulet et al., 1995) and reviewed in (Gronostajski, 2000), suggesting that also in megakaryocytes the various NFI products might perform independent functions. In the case of NFIB, the proteins translated in megakaryocytes do not contain the DNA-binding and dimerisation domain. This finding suggests that NFIB does not dimerise either with itself or NFIC proteins, or binds directly to DNA. This leads to an open question about its functional role.

The C-terminal domain of NFI proteins that is present in megakaryocytes can function in opposite ways in the regula-

tion of transcription, either as an activator or repressor. To determine if and how NFIB regulates gene expression in megakaryocytes we have set up the following experiments:

1. ChIP-seq profiling to identify DNA binding sites, where NFIB might be bound through co-factors. Integration of the binding sites with the megakaryocytic gene expression data will help us define its role in gene regulation.

2. functional assays to identify any megakaryocytic specific phenotypes by knock-down and overexpression. We will knockdown NFIB using shRNA or over-express it (either full-length isoform or the C-terminal domain only) in $CD34^+$ cells and test their colony forming potential during megakaryocytic differentiation. These experiments could also be coupled with gene expression assays to assess any gene expression changes.

Similar experiments are also under way for NFIC.

4.4 MATERIAL AND METHODS

All wet-lab experiments described below have been performed by members of Prof Willem Ouwehand's group, Department of Haematology, University of Cambridge, UK and are only provided for completeness. I was only responsible for the computational analysis of the data sets presented in this chapter.

4.4.1 *Erythroblasts cell culture and RNA-seq library preparation*

The cell culture and the erythroblasts RNA-seq library were prepared by Dr Pete Smethurst and Dr Katrin Voss, in Prof Willem Ouwehand's lab, Department of Haematology, University of Cambridge. Erythroblasts were obtained from cord blood-derived $CD34^+$ haematopoietic stem cells by cultures for 7-12 days in a medium supplemented with erythropoietin, a cytokine

that promotes the differentiation of HSCs into the erythrocytic lineage, and interleukin-3 (IL3). At the completion of the culture about 70-90% of cells are of erythrocytic nature with the majority being $CD36^{+}CD71^{+}CD235a^{+}$.

Total RNA was prepared from erythroblasts that were obtained using the protocol above and extraction of RNA was according to the Trizol method (Invitrogen, Paisley, UK). The RNA pellet was resuspended in nuclease-free water (Applied Biosystems, Warrington, UK) and analysis by Agilent Bioanalyser 2100 (Agilent, Waldbronn, Germany) gave a RNA integrity number (RIN) of 8.4. Following DNase treatment (Turbo DNA Free, Applied Biosystems), 5 µg of total RNA was applied to the mRNA Sequencing kit (Illumina) following the manufacturer's instructions, however, PCR amplification of the library was performed before gel extraction of a band range of 150-200bp to obtain the purified library. This was quantified by qPCR followed by paired-end sequencing.

One library of poly(A)$^{+}$-selected RNA was sequenced on the Illumina Genome Analyzer II, yielding 30.5M paired-end 76 bp reads.

4.4.2 *Pre-alignment quality control*

Pre-alignment assessment of the reads was done using the FASTQC tool (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). Then reads were aligned to the *Homo sapiens* high coverage assembly (hg19) (release February 2009) using GSNAP version 2011-11-25. Read trimming was disabled and I allowed for up to 5 mismatches and novel splicing sites at most 300,000 bp apart. Trimming off adapters is not an essential step before using GSNAP, as the tool soft clips adapter sequences.

4.4.3 *Quantification of gene expression*

Gene and isoform quantification was performed using Cufflinks v2.0.2 (Trapnell et al., 2010a) and reported in Fragments Per Kilobase of exon per Million fragments mapped (FPKM). The annotation used was Ensembl v70.

4.4.4 *5′ race PCR*

5′ race PCR was carried out using 5′/3′ Race kit 2nd generation (Roche 03353621001) according to manufacturer instructions with the following oligonucleotides for NFIB:

1. SP1 CTAGCCTTCGTTGGTGAGATAACCAG
2. SP2 GGAATTAGTGATTGTAAGTGCTGCA
3. SP3 CACATATCGATTGGCTTGAGATGTGC

The RNA was isolated from megakaryocytes cultured from CD34 cells isolated from apheresis cones. See below for culture details.

4.4.5 *Megakaryocytes culture*

$CD34^+$ cells were isolated from apheresis cones using an AUTOmacs pro and cultured in Cellgro (CellGenix) supplemented with TPO (CellGenix) 100ng/ml and IL1β (R&D). On day 10 megakaryocytes were sorted (>95% purity) using an anti-CD42b PE conjugated antibody and a PE positive selection kit (STEMCELL).

4.4.6 *Immunoblotting and antibodies*

Primary megakaryocytes were lysed in Laemmli buffer by boiling the cell pellets for 10 minutes; 10^5 cells equivalents were

loaded in each lane and transferred onto PVDF membranes. Membranes were probed with the following antibodies:

1. NFIB Abcam AB11989 and AB 80835.
2. NFIC Bethyl Laboratories A303123A and A303124A.

Part III

CONCLUSION

5

DISCUSSION

5.1 CONCLUSIONS

Next-generation sequencing has revolutionised the field of genomics, providing a genome-wide dimension to known assays and allowing for the development of a host of new techniques to interrogate gene expression and its regulation. Its predominant applications include the identification of protein-DNA interaction at genomic loci (ChIP-seq), transcriptome analysis (RNA-seq) and DNA methylation analysis. Irrespective of the application, next-generation sequencing provides unprecedented amounts of biological data, creating a need for comprehensive bioinformatics analysis, to handle and process the resulting big data sets, and more importantly to extract biological insights from them. In this thesis, two next-generation sequencing applications have been used on several cell-types of the haematopoietic system in an attempt to study gene expression and its regulation in this well-studied system.

The main focus of my thesis has been to gain further understanding in the process that leads to the formation of platelets. To this end, I examined the gene and transcript isoform expression of megakaryocytes and their precursors. For the first time, RNA sequencing was applied to human megakaryocytes for the identification of splicing events and the quantification of the transcript isoform expression. This data set was integrated with genome-wide binding profiles of known regulators of haematopoiesis, including FLI1, GATA1, GATA2, RUNX1,SCL/TAL1 and MEIS1.

The integrative analysis showed frequent co-localisation of these six transcription factors in megakaryocytes and confirmed the already known key regulatory role of MEIS1 in megakaryopoiesis and platelet production. Moreover I found several examples of novel transcription start sites bound by MEIS1, suggesting that it regulates megakaryopoiesis through the control of gene isoform expression.

Platelet production, similarly to the production of any other mature blood type, proceeds in a hierarchical fashion. During a tightly regulated differentiation process, haematopoietic stem cells commit to differentiation through a sequence of increasingly lineage-restricted progenitor cells. As part of the BLUEPRINT consortium, I analysed RNA sequencing data sets of six rare haematopoietic progenitor cells with the aim to characterise the transcriptome of the intermediate cell types of the haematopoietic system. These datasets constitute the first in-depth analysis of cell-type and lineage-specific expression of transcript isoforms in rare precursors. My analysis provides insight not only into gene-level expression, as partly covered by previous microarray-based studies, but extends to the transcript level. In addition to the genome-wide study of transcription in these cells, I identified two examples where the alternative splicing of transcript isoforms affects functional domains generating protein isoforms with potentially diverse functions among the different cell types.

While genome-wide data sets like those analysed in the first two chapters of my thesis allow to comprehensively describe binding and expression in the given cell types, it remains a major challenge to translate this knowledge into functional insights into gene regulation. The final chapter of my thesis provides an example of how an integrative analysis of various datasets, including those in Chapters 2 and 3, can enable the identification of novel regulators in megakaryopoiesis. Combin-

ing the gene expression profiles of transcription factors in megakaryocyte/erythrocyte precursors with those in megakaryocytes and erythroblasts. I was able to identify candidate regulators of megakaryopoiesis. Further integration with publicly available datasets of mature blood cells' gene expression identified NFIB as a transcription factor with megakaryocyte specific isoforms.

5.2 FUTURE PERSPECTIVES

As sequencing costs decrease, it has become easier to conduct genome-wide assays for a comprehensive characterisation of the functional genome and transcriptome. Large consortia, such as the ENCODE (ENCODE Project Consortium, 2012) and NIH Roadmap Epigenome (Bernstein et al., 2010), have generated an in-depth reference of widely used cell-lines and primary cells. Similarly, as a member of the Blueprint consortium I aspire to continue working towards the generation of an extensive annotation of the epigenome of different primary blood cells, which will assist our continuous efforts to study normal and malignant blood cell functions.

This thesis provided an example of how the integrative analysis of different next-generation sequencing data sets can reveal potential regulators of gene expression in megakaryocytes. However, such findings require further functional validation. In the case of NFIB follow-up studies are needed to establish whether NFIB has a central role in the gene regulatory network of megakaryocytes. Functional assays examining the phenotype of the knockout and the over-expression of different isoforms of this transcription factor will shed light on its role. Screening of NFIB interactions will also help us identify novel interactors. In addition, ChIP-sequencing of NFIB and NFIC will be essential to identify at which genomic loci these transcription factors bind, if at all. If proven essential for megakaryopoiesis, the iden-

tification of NFIB can open new directions in the on-going efforts of many groups to generate functional platelets *in vitro*. Patients with thrombocytopenia are in constant need of platelets transfusions, which are obtained through blood donations. Repeated transfusions, though, might result in rejection of the transfused platelets by the recipients (Schiffer, 2001). Therefore, it has become critical to identify those transcription factors that can enhance the production of platelets from induced pluripotent stem cells, which can be obtained from the same patients (Takayama et al., 2010), overcoming any problems posed by the expression of antibodies in the patients against donor platelets.

Moreover, the analysis of novel transcription start sites in megakaryocytes coinciding with MEIS1 binding revealed several novel regulatory regions in megakaryocytes. However, additional information is needed to perform a comprehensive analysis that catalogues all active gene regulatory elements in megakaryocytes. To achieve this, ChIP sequencing of key histone modification marks, DNaseI hypersensitivity and DNA methylation experiments are required. Those data will shortly be provided by the BLUEPRINT consortium. Such datasets may then be integrated with the transcription factor binding sites presented in Chapter 2 in order to gain further insight to the regulatory mechanisms active during megakaryopoiesis.

The data sets presented in this thesis, along with all others generated within Blueprint, will become publicly available in several formats and through different websites, in order to be accessible to all types of interesting parties, from computational biologists to wet-lab scientists and clinicians. Thus creating a rich information background of the normal function of blood cell types, which can be then used for the delineation of the mechanisms perturbed in disease. High-throughput genotyping has brought to light a wealth of common and rare variants,

a number of which is located outside protein-coding regions, making the identification of their functional consequences difficult. Integrative approaches of functional genomics data within a disease-relevant context are therefore required in order to draw the relationship between genotypes and phenotypes (Knight, 2012). For example, genome-wide characterisation of the regulatory elements in megakaryocytes can be used to study the mechanisms through which functional genomics variants affect platelet function, volume or count. Such systematic analyses will expand our understanding of common variants already identified through GWAS studies (Gieger et al., 2011) and rare bleeding disorders investigated by the BRIDGE-BPD consortium (https://bridgestudy.medschl.cam.ac.uk/bpd.shtml).

RNA sequencing of megakaryocytes has led to the identification of novel intergenic transcripts that originate from previously unanottated genomic regions. In recent years, growing evidence of long non-coding RNAs (lincRNAs) role in regulation of gene expression of neighboring genes has emerged (Paralkar and Weiss, 2013). Combining histone modification marks with transcriptome data in megakaryocytes will enable the identification of such RNAs. In addition, using the Blueprint RNA-sequencing data we can classify the lincRNAs identified into cell-type or lineage-specific within progenitors and the megakaryocytic-erythroid branch. Any interesting findings would then be validated by performing knockout experiments coupled with either microarray gene expression profiling or qPCR to examine changes in gene expression levels.

In this thesis, the megakaryocytes profiled were cord-blood stem cell *in vitro* derived megakaryocytes, which are the closest available model for in vivo megakaryopoiesis. Megakaryocytes are extremely rare in the bone marrow and access to human bone marrow is limited, making any genome-wide studies of these cells impossible due to the high numbers of cells required.

The constant progress in sequencing technologies and the introduction of single-cell sequencing techniques will allow us to study the transcriptome of single bone-marrow derived megakaryocytes (reviewed in (Shapiro et al., 2013)). With this approach, we shall be able to overcome any homogeneity issues present in megakaryocyte cell cultures, such as the different levels of ploidy. Once single cell sequencing becomes more prevalent, transcriptome analysis of megakaryocytes will allow us to investigate the different stages of maturation of these cells to identify revelant changes in gene expression. In the longer term, single-cell epigenetic assays may provide access to epigenome of rare cell types.

C. Abramovich, N. Pineault, H. Ohta, and R. K. Humphries. Hox genes: from leukemia to hematopoietic stem cell expansion. *Ann. N. Y. Acad. Sci.*, 1044:109–116, Jun 2005.

D. Adams, L. Altucci, S. E. Antonarakis, J. Ballesteros, S. Beck, A. Bird, C. Bock, B. Boehm, E. Campo, A. Caricasole, F. Dahl, E. T. Dermitzakis, T. Enver, M. Esteller, X. Estivill, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, C. Giehl, T. Graf, F. Grosveld, R. Guigo, I. Gut, K. Helin, J. Jarvius, R. Küppers, H. Lehrach, T. Lengauer, Å. Lernmark, D. Leslie, M. Loeffler, E. Macintyre, A. Mai, J. H. A. Martens, S. Minucci, W. H. Ouwehand, P. G. Pelicci, H. Pendeville, B. Porse, V. Rakyan, W. Reik, M. Schrappe, D. Schübeler, M. Seifert, R. Siebert, D. Simmons, N. Soranzo, S. Spicuglia, M. Stratton, H. G. Stunnenberg, A. Tanay, D. Torrents, A. Valencia, E. Vellenga, M. Vingron, J. Walter, and S. Willcocks. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224–226, Mar. 2012.

M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013): 1651–1656, Jun 1991.

M. Adli, J. Zhu, and B. E. Bernstein. Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods*, 7(8):615–618, Aug 2010.

J. Adolfsson, O. J. Borge, D. Bryder, K. Theilgaard-Monch, I. Astrand-Grundstrom, E. Sitnicka, Y. Sasaki, and S. E. Jacobsen. Upregulation of Flt3 expression within the bone marrow

Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity*, 15(4):659–669, Oct 2001.

J. Adolfsson, R. Mansson, N. Buza-Vidas, A. Hultquist, K. Liuba, C. T. Jensen, D. Bryder, L. Yang, O. J. Borge, L. A. Thoren, K. Anderson, E. Sitnicka, Y. Sasaki, M. Sigvardsson, and S. E. Jacobsen. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell*, 121(2):295–306, Apr 2005. [DOI:10.1016/j.cell.2005.02.013] [PubMed:15851035].

T. Aijo, S. M. Edelman, T. Lonnberg, A. Larjo, H. Kallionpaa, S. Tuomela, E. Engstrom, R. Lahesmaa, and H. Lahdesmaki. An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human T helper cell differentiation. *BMC Genomics*, 13:572, 2012.

K. Akashi, D. Traver, and L. I. Zon. The complex cartography of stem cell commitment. *Cell*, 121(2):160–162, Apr 2005.

C. A. Albers, A. Cvejic, R. Favier, E. E. Bouwmans, M. C. Alessi, P. Bertone, G. Jordan, R. N. Kettleborough, G. Kiddle, M. Kostadima, R. J. Read, B. Sipos, S. Sivapalaratnam, P. A. Smethurst, J. Stephens, K. Voss, A. Nurden, A. Rendon, P. Nurden, and W. H. Ouwehand. Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. *Nat. Genet.*, 43(8):735–737, Aug 2011.

C. A. Albers, D. S. Paul, H. Schulze, K. Freson, J. C. Stephens, P. A. Smethurst, J. D. Jolley, A. Cvejic, M. Kostadima, P. Bertone, M. H. Breuning, N. Debili, P. Deloukas, R. Favier, J. Fiedler, C. M. Hobbs, N. Huang, M. E. Hurles, G. Kiddle, I. Krapels, P. Nurden, C. A. Ruivenkamp, J. G. Sambrook, K. Smith, D. L. Stemple, G. Strauss, C. Thys, C. van Geet, R. Newbury-Ecob, W. H. Ouwehand, and C. Ghevaert. Compound inheritance of a low-frequency regulatory SNP and a

rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat. Genet.*, 44(4):435–439, Apr 2012.

B. Alberts. *Molecular Biology of the Cell: 5th edition*. Number v. 1 in Molecular Biology of the Cell: 5th Edition. Garland Science, 2008. ISBN 9780815341062.

A. Alevizopoulos, Y. Dusserre, M. Tsai-Pflugfelder, T. von der Weid, W. Wahli, and N. Mermod. A proline-rich TGF-beta-responsive transcriptional activator interacts with histone H3. *Genes Dev.*, 9(24):3051–3066, Dec 1995.

V. G. Allfrey and A. E. Mirsky. Structural Modifications of Histones and their Possible Role in the Regulation of RNA Synthesis. *Science*, 144(3618):559, May 1964.

S. Anders, D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*, 8(9):1765–1786, Sep 2013.

R. G. Andrews, J. W. Singer, and I. D. Bernstein. Monoclonal antibody 12-8 recognizes a 115-kd molecule present on both unipotent and multipotent hematopoietic colony-forming cells and their precursors. *Blood*, 67(3):842–845, Mar 1986.

R. G. Andrews, J. W. Singer, and I. D. Bernstein. Precursors of colony-forming cells in humans can be distinguished from colony-forming cells by expression of the CD33 and CD34 antigens and light scatter properties. *J. Exp. Med.*, 169(5):1721–1731, May 1989.

D. J. Anstee. Blood group antigens defined by the amino acid sequences of red cell surface proteins. *Transfusion medicine (Oxford, England)*, 5(1):1–13, Mar. 1995.

M. T. Armentero, M. Horwitz, and N. Mermod. Targeting of DNA polymerase to the adenovirus origin of DNA replication by interaction with nuclear factor I. *Proceedings of the*

National Academy of Sciences of the United States of America, 91 (24):11537–11541, Nov. 1994.

G. Attardi and F. Amaldi. Structure and synthesis of ribosomal RNA. *Annu. Rev. Biochem.*, 39:183–226, 1970.

L. H. Augenlicht and D. Kobrin. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Res.*, 42(3): 1088–1093, Mar 1982.

V. Azcoitia, M. Aracil, and C. Martínez-A. The homeodomain protein Meis1 is essential for definitive hematopoiesis and vascular patterning in the mouse embryo. *Dev Biol*, 2005.

E. R. Bacon, N. Dalyot, D. Filon, L. Schreiber, E. A. Rachmilewitz, and A. Oppenheim. Hemoglobin switching in humans is accompanied by changes in the ratio of the transcription factors, GATA-1 and SP1. *Mol. Med.*, 1(3):297–305, Mar 1995.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

A. Balduini, M. D'Apolito, D. Arcelli, V. Conti, A. Pecci, D. Pietra, M. Danova, F. Benvenuto, C. Perotti, L. Zelante, S. Volinia, C. L. Balduini, and A. Savoia. Cord blood in vitro expanded CD41+ cells: identification of novel components of megakaryocytopoiesis. *Journal of Thrombosis and Haemostasis*, 4(4):848–860, Apr. 2006.

M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, 12(11):745–755, Nov 2011.

N. L. Barbosa-Morais, M. J. Dunning, S. A. Samarajiwa, J. F. J. Darot, M. E. Ritchie, A. G. Lynch, and S. Tavaré. A re-annotation pipeline for Illumina BeadArrays: improving the

interpretation of gene expression data. *Nucleic acids research*, 38(3):e17, Jan. 2010.

A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007.

D. P. Bartel. Micrornas: Target recognition and regulatory functions. *Cell*, 136(2):215 – 233, 2009. ISSN 0092-8674. doi: http://dx.doi.org/10.1016/j.cell.2009.01.002. URL http://www.sciencedirect.com/science/article/pii/S0092867409000087.

K. Barton, N. Muthusamy, C. Fischer, C. N. Ting, T. L. Walunas, L. L. Lanier, and J. M. Leiden. The Ets-1 transcription factor is required for the development of natural killer cells in mice. *Immunity*, 9(4):555–563, Oct 1998.

C. M. Baum, I. L. Weissman, A. S. Tsukamoto, A. M. Buckle, and B. Peault. Isolation of a candidate human hematopoietic stem-cell population. *Proc. Natl. Acad. Sci. U.S.A.*, 89(7):2804–2808, Apr 1992.

I. Beerman, D. Bhattacharya, S. Zandi, M. Sigvardsson, I. L. Weissman, D. Bryder, and D. J. Rossi. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc. Natl. Acad. Sci. U.S.A.*, 107(12):5465–5470, Mar 2010.

O. Bell, V. K. Tiwari, N. H. Thoma, and D. Schubeler. Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, 12(8):554–564, Aug 2011.

C. Benoist, K. O'Hare, R. Breathnach, and P. Chambon. The ovalbumin gene-sequence of putative control regions. *Nucleic acids research*, 8(1):127–142, Jan. 1980.

R. J. Berenson, R. G. Andrews, W. I. Bensinger, D. Kalamasz, G. Knitter, C. D. Buckner, and I. D. Bernstein. Antigen CD34+ marrow cells engraft lethally irradiated baboons. *J. Clin. Invest.*, 81(3):951–955, Mar 1988.

R. J. Berenson, W. I. Bensinger, R. S. Hill, R. G. Andrews, J. Garcia-Lopez, D. F. Kalamasz, B. J. Still, G. Spitzer, C. D. Buckner, and I. D. Bernstein. Engraftment after infusion of CD34+ marrow cells in patients with breast cancer or neuroblastoma. *Blood*, 77(8):1717–1722, Apr 1991.

S. M. Berget, C. Moore, and P. A. Sharp. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74(8):3171–3175, Aug 1977.

S. A. Berkman. Infectious complications of blood transfusion. *Blood Rev.*, 2(3):206–210, Sep 1988.

B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, 28(10):1045–1048, Oct 2010.

J. Berthelsen, V. Zappavigna, F. Mavilio, and F. Blasi. Prep1, a novel functional partner of Pbx proteins. *EMBO J.*, 17(5): 1423–1433, Mar 1998.

P. Bertone, V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306 (5705):2242–2246, Dec 2004.

M. Bessis. [Erythroblastic island, functional unity of bone marrow]. *Rev Hematol*, 13(1):8–11, 1958.

M. Bhatia, J. C. Wang, U. Kapp, D. Bonnet, and J. E. Dick. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. *Proc. Natl. Acad. Sci. U.S.A.*, 94(10):5320–5325, May 1997.

A. Bird. DNA methylation patterns and epigenetic memory. *Genes Dev.*, 16(1):6–21, Jan 2002.

J. Bohm, A. Buck, W. Borozdin, A. U. Mannan, U. Matysiak-Scholze, I. Adham, W. Schulz-Schaeffer, T. Floss, W. Wurst, J. Kohlhase, and F. Barrionuevo. Sall1, sall2, and sall4 are required for neural tube closure in mice. *Am. J. Pathol.*, 173(5): 1455–1463, Nov 2008.

F. A. Bonilla and H. C. Oettgen. Adaptive immunity. *J. Allergy Clin. Immunol.*, 125(2 Suppl 2):33–40, Feb 2010.

S. Bonn, R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczyński, A. Riddell, and E. E. Furlong. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.*, 44(2):148–156, Feb 2012.

J. C. Bories, D. M. Willerford, D. Grevin, L. Davidson, A. Camus, P. Martin, D. Stehelin, and F. W. Alt. Increased T-cell apoptosis and terminal B-cell differentiation induced by inactivation of the Ets-1 proto-oncogene. *Nature*, 377(6550):635–638, Oct 1995.

G. B. Bradford, B. Williams, R. Rossi, and I. Bertoncello. Quiescence, cycling, and turnover in the primitive hematopoietic stem cell compartment. *Exp. Hematol.*, 25(5):445–453, May 1997.

P. F. Bray, S. E. McKenzie, L. C. Edelstein, S. Nagalla, K. Delgrosso, A. Ertel, J. Kupper, Y. Jing, E. Londin, P. Loher, H.-W.

Chen, P. Fortina, and I. Rigoutsos. The complex transcriptional landscape of the anucleate human platelet. *BMC genomics*, 14(1):1, Jan. 2013.

S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, 18(6):630–634, Jun 2000a.

S. Brenner, S. R. Williams, E. H. Vermaas, T. Storck, K. Moon, C. McCollum, J. I. Mao, S. Luo, J. J. Kirchner, S. Eletr, R. B. DuBridge, T. Burcham, and G. Albrecht. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. U.S.A.*, 97(4):1665–1670, Feb 2000b.

J. Brown, M. F. Greaves, and H. V. Molgaard. The gene encoding the stem cell antigen, CD34, is conserved in mouse and expressed in haemopoietic progenitor cell lines, brain, and embryonic fibroblasts. *Int. Immunol.*, 3(2):175–184, Feb 1991.

A. C. Brun, J. M. Bjornsson, M. Magnusson, N. Larsson, P. Leveen, M. Ehinger, E. Nilsson, and S. Karlsson. Hoxb4-deficient mice undergo normal hematopoietic development but exhibit a mild proliferation defect in hematopoietic stem cells. *Blood*, 103(11):4126–4133, Jun 2004.

A. L. Brunner, D. S. Johnson, S. W. Kim, A. Valouev, T. E. Reddy, N. F. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao, C. B. Oyolu, G. P. Schroth, D. M. Absher, J. C. Baker, and R. M. Myers. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.*, 19(6):1044–1056, Jun 2009.

J. C. Bryne, E. Valen, M. H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, 36(Database issue):D102–106, Jan 2008.

H. Busch, R. Reddy, L. Rothblum, and Y. C. Choi. SnRNAs, SnRNPs, and RNA processing. *Annu. Rev. Biochem.*, 51:617–654, 1982.

M. Cai, E. M. Langer, J. G. Gill, A. T. Satpathy, J. C. Albring, W. KC, T. L. Murphy, and K. M. Murphy. Dual actions of Meis1 inhibit erythroid progenitor development and sustain general hematopoietic cell proliferation. *Blood*, 120(2):335–346, Jul 2012.

C. E. Campbell, M. Piper, C. Plachez, Y.-T. Yeh, J. S. Baizer, J. M. Osinski, E. D. Litwack, L. J. Richards, and R. M. Gronostajski. The transcription factor Nfix is essential for normal brain development. *BMC developmental biology*, 8:52, 2008.

G. A. Challen, N. C. Boles, S. M. Chambers, and M. A. Goodell. Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. *Cell Stem Cell*, 6(3):265–278, Mar 2010.

S. M. Chambers, N. C. Boles, K. Y. Lin, M. P. Tierney, T. V. Bowman, S. B. Bradfute, A. J. Chen, A. A. Merchant, O. Sirin, D. C. Weksberg, M. G. Merchant, C. J. Fisk, C. A. Shaw, and M. A. Goodell. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell*, 1(5):578–591, Nov 2007.

A. N. Chang, A. B. Cantor, Y. Fujiwara, M. B. Lodish, S. Droho, J. D. Crispino, and S. H. Orkin. GATA-factor dependence of the multitype zinc-finger protein FOG-1 for its essential role in megakaryopoiesis. *Proc. Natl. Acad. Sci. U.S.A.*, 99(14):9237–9242, Jul 2002.

C. P. Chang, L. Brocchieri, W. F. Shen, C. Largman, and M. L. Cleary. Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. *Mol. Cell. Biol.*, 16(4):1734–1745, Apr 1996.

C. P. Chang, Y. Jacobs, T. Nakamura, N. A. Jenkins, N. G. Copeland, and M. L. Cleary. Meis proteins are major in vivo DNA binding partners for wild-type but not chimeric Pbx proteins. *Mol. Cell. Biol.*, 17(10):5679–5687, Oct 1997.

C.-Y. Chang, H. A. Pasolli, E. G. Giannopoulou, G. Guasch, R. M. Gronostajski, O. Elemento, and E. Fuchs. NFIB is a governor of epithelial-melanocyte stem cell behaviour in a shared niche. *Nature*, 495(7439):98–102, Mar. 2013.

A. Z. Chaudhry, G. E. Lyons, and R. M. Gronostajski. Expression patterns of the four nuclear factor I genes during mouse embryogenesis indicate a potential role in development. *Developmental dynamics : an official publication of the American Association of Anatomists*, 208(3):313–325, Mar. 1997.

M. Chen, N. Mermod, and M. S. Horwitz. Protein-protein interactions between adenovirus DNA polymerase and nuclear factor I mediate formation of the DNA replication preinitiation complex. *The Journal of biological chemistry*, 265(30): 18634–18642, Oct. 1990.

J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725):1149–1154, May 2005.

S. H. Cheshier, S. J. Morrison, X. Liao, and I. L. Weissman. In vivo proliferation and cell cycle kinetics of long-term self-

renewing hematopoietic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):3120–3125, Mar 1999.

S. Chiba, K. Takeshita, Y. Imai, K. Kumano, M. Kurokawa, S. Masuda, K. Shimizu, S. Nakamura, F. H. Ruddle, and H. Hirai. Homeoprotein DLX-1 interacts with Smad4 and blocks a signaling pathway from activin A in hematopoietic cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15577–15582, Dec. 2003.

V. Chitu and E. R. Stanley. Colony-stimulating factor-1 in immunity and inflammation. *Curr. Opin. Immunol.*, 18(1):39–48, Feb 2006.

L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, Sep 1977.

J. L. Christensen and I. L. Weissman. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 98(25): 14541–14546, Dec 2001.

S. Chuikov, B. P. Levi, M. L. Smith, and S. J. Morrison. Prdm16 promotes stem cell maintenance in multiple tissues, partly by regulating oxidative stress. *Nat. Cell Biol.*, 12(10):999–1006, Oct 2010.

C. I. Civin, M. L. Banquerigo, L. C. Strauss, and M. R. Loken. Antigenic analysis of hematopoiesis. VI. Flow cytometric characterization of My-10-positive progenitor cells in normal human bone marrow. *Exp. Hematol.*, 15(1):10–17, Jan 1987.

T. A. Clark, C. W. Sugnet, and M. Ares. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, 296(5569):907–910, May 2002.

N. Cloonan, A. R. Forrest, G. Kolle, B. B. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani,

G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, 5(7):613–619, Jul 2008.

T. Clouaire and I. Stancheva. Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? *Cell. Mol. Life Sci.*, 65(10):1509–1522, May 2008.

P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, 38(6):1767–1771, Apr 2010.

F. S. Collins, M. Morgan, and A. Patrinos. The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617): 286–290, Apr 2003.

E. Conneally, J. Cashman, A. Petzer, and C. Eaves. Expansion in vitro of transplantable human cord blood stem cells demonstrated using a quantitative assay of their lympho-myeloid repopulating activity in nonobese diabetic-scid/scid mice. *Proc. Natl. Acad. Sci. U.S.A.*, 94(18):9836–9841, Sep 1997.

A. E. Corcoran, F. M. Smart, R. J. Cowling, T. Crompton, M. J. Owen, and A. R. Venkitaraman. The interleukin-7 receptor alpha chain transmits distinct signals for proliferation and differentiation during B lymphopoiesis. *EMBO J.*, 15(8):1924–1932, Apr 1996.

V. Costa, C. Angelini, I. De Feis, and A. Ciccodicola. Uncovering the complexity of transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.*, 2010:853916, 2010.

W. Craig, R. Kay, R. L. Cutler, and P. M. Lansdorp. Expression of Thy-1 on human hematopoietic progenitor cells. *J. Exp. Med.*, 177(5):1331–1342, May 1993.

M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, 107 (50):21931–21936, Dec 2010.

F. H. Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163, 1958.

A. Cvejic, J. Serbanovic-Canic, D. Stemple, and W. Ouwehand. The role of meis1 in primitive and definitive hematopoiesis during zebrafish development. *Haematologica*, 96(2):190, Feb. 2011.

A. Cvejic, L. Haer-Wigman, J. C. Stephens, M. Kostadima, P. A. Smethurst, M. Frontini, E. van den Akker, P. Bertone, E. Bielczyk-Maczy/'nska, S. Farrow, R. S. Fehrmann, A. Gray, M. de Haas, V. G. Haver, G. Jordan, J. Karjalainen, H. H. Kerstens, G. Kiddle, H. Lloyd-Jones, M. Needs, J. Poole, A. A. Soussan, A. Rendon, K. Rieneck, J. G. Sambrook, H. Schepers, H. H. Sillje, B. Sipos, D. Swinkels, A. U. Tamuri, N. Verweij, N. A. Watkins, H. J. Westra, D. Stemple, L. Franke, N. Soranzo, H. G. Stunnenberg, N. Goldman, P. van der Harst, C. E. van der Schoot, W. H. Ouwehand, and C. A. Albers. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.*, 45(5):542–545, May 2013.

R. Dahl, J. C. Walsh, D. Lancki, P. Laslo, S. R. Iyer, H. Singh, and M. C. Simon. Regulation of macrophage and neutrophil cell fates by the PU.1:C/EBPalpha ratio and granulocyte colony-stimulating factor. *Nat. Immunol.*, 4(10):1029–1036, Oct 2003.

G. F. Dalton. The tradition of blood sacrifice to the goddess Éire. *Studies: An Irish Quarterly Review*, 63(252):pp. 343–354, 1974. ISSN 00393495. URL http://www.jstor.org/stable/30088757.

L. das Neves, C. S. Duchala, F. Tolentino-Silva, M. A. Haxhiu, C. Colmenares, W. B. Macklin, C. E. Campbell, K. G. Butz, R. M. Gronostajski, and F. Godinho. Disruption of the murine nuclear factor I-A gene (Nfia) results in perinatal lethality, hydrocephalus, and agenesis of the corpus callosum. *Proceedings of the National Academy of Sciences of the United States of America*, 96(21):11946–11951, Oct. 1999.

J. Dausset. [Iso-leuko-antibodies]. *Acta Haematol.*, 20(1-4):156–166, 1958.

J. Dausset and A. Nenna. [Presence of leuko-agglutinin in the serum of a case of chronic agranulocytosis]. *C. R. Seances Soc. Biol. Fil.*, 146(19-20):1539–1541, Oct 1952.

R. N. de Jong and P. C. van der Vliet. Mechanism of DNA replication in eukaryotic cells: cellular host factors stimulating adenovirus DNA replication. *Gene*, 236(1):1–12, Aug. 1999.

J. Dekker, J. A. van Oosterhout, and P. C. van der Vliet. Two regions within the DNA binding domain of nuclear factor I interact with DNA and stimulate adenovirus DNA replication independently. *Mol. Cell. Biol.*, 16(8):4073–4080, Aug 1996.

E. J. Dettman and M. J. Justice. The zinc finger SET domain gene Prdm14 is overexpressed in lymphoblastic lymphomas with retroviral insertions at Evi32. *PLoS ONE*, 3(11):e3823, 2008.

E. J. Dettman, S. J. Simko, B. Ayanga, B. L. Carofino, J. F. Margolin, H. C. Morse, and M. J. Justice. Prdm14 initiates lymphoblastic leukemia after expanding a population of cells resembling common lymphoid progenitors. *Oncogene*, 30(25):2859–2873, Jun 2011.

A. Diaz, A. Nellore, and J. S. Song. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol.*, 13(10):R98, Oct 2012.

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

T. A. Down, V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Graf, N. Johnson, J. Herrero, E. M. Tomazou, N. P. Thorne, L. Backdahl, M. Herberth, K. L. Howe, D. K. Jackson, M. M. Miretti, J. C. Marioni, E. Birney, T. J. Hubbard, R. Durbin, S. Tavare, and S. Beck. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, 26(7):779–785, Jul 2008.

D. Duboule. The rise and fall of Hox gene clusters. *Development*, 134(14):2549–2560, Jul 2007.

B. Dykstra, D. Kent, M. Bowie, L. McCaffrey, M. Hamilton, K. Lyons, S. J. Lee, R. Brinkman, and C. Eaves. Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell*, 1(2):218–229, Aug 2007.

E. Dzierzak and N. A. Speck. Of lineage and legacy: the development of mammalian hematopoietic stem cells. *Nat. Immunol.*, 9(2):129–136, Feb 2008.

S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2(12):919–929, Dec 2001.

A. Efstratiadis, J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. L. Slightom, A. E. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot. The structure and evolution of the human beta-globin gene family. *Cell*, 21(3): 653–668, Oct. 1980.

G. L. Eliceiri. Small nucleolar RNAs. *Cell. Mol. Life Sci.*, 56(1-2): 22–31, Oct 1999.

H. Ellis. James Blundell, pioneer of blood transfusion. *Br J Hosp Med*, 68(8):447, Aug 2007.

ENCODE Project Consortium. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

ENCYCLOPAEDIA BRITANNICA. ENCYCLOPAEDIA BRITANNICA: Blooc cell formation. http://www.britannica.com/EBchecked/topic/69747/blood-cell-formation/. Accessed: 2013-11-08.

P. G. Engstrom, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, T. Alioto, J. Behr, P. Bertone, R. Bohnert, D. Campagna, C. A. Davis, A. Dobin, P. G. Engstrom, T. R. Gingeras, N. Goldman, G. R. Grant, R. Guigo, J. Harrow, T. J. Hubbard, G. Jean, A. Kahles, P. Kosarev, S. Li, J. Liu, C. E. Mason, V. Molodtsov, Z. Ning, H. Ponstingl, J. F. Prins, G. Ratsch, P. Ribeca, I. Seledtsov, B. Sipos, V. Solovyev, T. Steijger, G. Valle, N. Vitulo, K. Wang, T. D. Wu, G. Zeller, G. Ratsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigo, and P. Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, Nov 2013.

P. Ernst, J. K. Fisher, W. Avery, S. Wade, D. Foy, and S. J. Korsmeyer. Definitive hematopoiesis requires the mixed-lineage leukemia gene. *Dev. Cell*, 6(3):437–443, Mar 2004.

S. Eyquem, K. Chemin, M. Fasseu, M. Chopin, F. Sigaux, A. Cumano, and J. C. Bories. The development of early and mature B cells is impaired in mice deficient for the Ets-1 transcription factor. *Eur. J. Immunol.*, 34(11):3187–3196, Nov 2004.

P. J. Farnham. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, 10(9):605–616, Sep 2009.

A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. Jones. FindPeaks 3.1: a tool for identifying areas

of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, Aug 2008.

E. Ferretti, H. Marshall, H. Popperl, M. Maconochie, R. Krumlauf, and F. Blasi. Segmental expression of Hoxb2 in r4 requires two separate sites that integrate cooperative interactions between Prep1, Pbx and Hox proteins. *Development*, 127 (1):155–166, Jan 2000.

F. Ficara, M. J. Murphy, M. Lin, and M. L. Cleary. Pbx1 regulates self-renewal of long-term hematopoietic stem cells by maintaining their quiescence. *Cell Stem Cell*, 2(5):484–496, May 2008.

G. J. Filion, S. Zhenilo, S. Salozhin, D. Yamada, E. Prokhortchouk, and P. A. Defossez. A family of human zinc finger proteins that bind methylated DNA and repress transcription. *Mol. Cell. Biol.*, 26(1):169–181, Jan 2006.

R. C. Fisher and E. W. Scott. Role of PU.1 in hematopoiesis. *Stem Cells*, 16(1):25–37, 1998.

A. H. Fox, C. Liew, M. Holmes, K. Kowalski, J. Mackay, and M. Crossley. Transcriptional cofactors of the FOG family interact with GATA proteins by means of multiple zinc fingers. *EMBO J.*, 18(10):2812–2822, May 1999.

G. Fritsch, P. Buchinger, D. Printz, F. M. Fink, G. Mann, C. Peters, T. Wagner, A. Adler, and H. Gadner. Rapid discrimination of early CD34+ myeloid progenitors using CD45-RA analysis. *Blood*, 81(9):2301–2309, May 1993.

A. Fuchs, M. Cella, E. Giurisato, A. S. Shaw, and M. Colonna. Cutting edge: CD96 (tactile) promotes NK cell-target cell adhesion by interacting with the poliovirus receptor (CD155). *J. Immunol.*, 172(7):3994–3998, Apr 2004.

P. G. Fuhrken, C. Chen, P. A. Apostolidis, M. Wang, W. M. Miller, and E. T. Papoutsakis. Gene Ontology-driven transcriptional analysis of CD34+ cell-initiated megakaryocytic

cultures identifies new transcriptional regulators of megakaryopoiesis. *Physiological genomics*, 33(2):159–169, Apr. 2008.

Y. Fujiwara, C. P. Browne, K. Cunniff, S. C. Goff, and S. H. Orkin. Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. *Proc. Natl. Acad. Sci. U.S.A.*, 93(22):12355–12358, Oct 1996.

A. Galy, M. Travis, D. Cen, and B. Chen. Human T, B, natural killer, and dendritic cells arise from a common bone marrow progenitor cell subset. *Immunity*, 3(4):459–473, Oct 1995.

A. M. Gannon and B. T. Kinsella. Regulation of the human thromboxane A2 receptor gene by Sp1, Egr1, NF-E2, GATA-1, and Ets-1 in megakaryocytes. *J. Lipid Res.*, 49(12):2590–2604, Dec 2008.

K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer. A map of open chromatin in human pancreatic islets. *Nat. Genet.*, 42(3):255–259, Mar 2010.

H. Geiger and G. Van Zant. The aging of lympho-hematopoietic stem cells. *Nat. Immunol.*, 3(4):329–333, Apr 2002.

D. S. Gerhard, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brown-stein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei,

S. Mathavan, P. H. Gunaratne, J. Wu, A. M. Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G. Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morin, R. Morrin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak, and J. Malek. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, 14(10B):2121–2127, Oct 2004.

M. Gering, A. R. Rodaway, B. Gottgens, R. K. Patient, and A. R. Green. The SCL gene specifies haemangioblast development from early mesoderm. *EMBO J.*, 17(14):4029–4045, Jul 1998.

A. M. Gewirtz, B. Calabretta, B. Rucinski, S. Niewiarowski, and W. Y. Xu. Inhibition of human megakaryocytopoiesis in vitro by platelet factor 4 (PF4) and a synthetic COOH-terminal PF4 peptide. *The Journal of clinical investigation*, 83(5):1477–1486, May 1989.

A. Giampaolo, P. Sterpetti, D. Bulgarini, P. Samoggia, E. Pelosi, M. Valtieri, and C. Peschle. Key functional role and lineage-specific expression of selected HOXB genes in purified hematopoietic progenitor differentiation. *Blood*, 84(11):3637–3647, Dec 1994.

B. Giebel, T. Zhang, J. Beckmann, J. Spanholtz, P. Wernet, A. D. Ho, and M. Punzel. Primitive human hematopoietic cells give rise to differentially specified daughter cells upon their initial cell division. *Blood*, 107(5):2146–2152, Mar 2006.

C. Gieger, A. Radhakrishnan, A. Cvejic, W. Tang, E. Porcu, G. Pistis, J. Serbanovic-Canic, U. Elling, A. H. Goodall, Y. Labrune, L. M. Lopez, R. Magi, S. Meacham, Y. Okada, N. Pirastu, R. Sorice, A. Teumer, K. Voss, W. Zhang, R. Ramirez-Solis, J. C. Bis, D. Ellinghaus, M. Gogele, J. J. Hottenga, C. Langenberg, P. Kovacs, P. F. O'Reilly, S. Y. Shin, T. Esko, J. Hartiala, S. Kanoni, F. Murgia, A. Parsa, J. Stephens, P. van der Harst, C. Ellen van der Schoot, H. Allayee, A. Attwood, B. Balkau, F. Bastardot, S. Basu, S. E. Baumeister, G. Biino, L. Bomba, A. Bonnefond, F. Cambien, J. C. Chambers, F. Cucca, P. D'Adamo, G. Davies, R. A. de Boer, E. J. de Geus, A. Doring, P. Elliott, J. Erdmann, D. M. Evans, M. Falchi, W. Feng, A. R. Folsom, I. H. Frazer, Q. D. Gibson, N. L. Glazer, C. Hammond, A. L. Hartikainen, S. R. Heckbert, C. Hengstenberg, M. Hersch, T. Illig, R. J. Loos, J. Jolley, K. T. Khaw, B. Kuhnel, M. C. Kyrtsonis, V. Lagou, H. Lloyd-Jones, T. Lumley, M. Mangino, A. Maschio, I. Mateo Leach, B. McKnight, Y. Memari, B. D. Mitchell, G. W. Montgomery, Y. Nakamura, M. Nauck, G. Navis, U. Nothlings, I. M. Nolte, D. J. Porteous, A. Pouta, P. P. Pramstaller, J. Pullat, S. M. Ring, J. I. Rotter, D. Ruggiero, A. Ruokonen, C. Sala, N. J. Samani, J. Sambrook, D. Schlessinger, S. Schreiber, H. Schunkert, J. Scott, N. L. Smith, H. Snieder, J. M. Starr, M. Stumvoll, A. Takahashi, W. H. Tang, K. Taylor, A. Tenesa, S. Lay Thein, A. Tonjes, M. Uda, S. Ulivi, D. J. van Veldhuisen, P. M. Visscher, U. Volker, H. E. Wichmann, K. L. Wiggins, G. Willemsen, T. P. Yang, J. Hua Zhao, P. Zitting, J. R. Bradley, G. V. Dedoussis, P. Gasparini, S. L. Hazen, A. Metspalu, M. Pirastu, A. R. Shuldiner, L. Joost van Pelt, J. J. Zwaginga, D. I. Boomsma, I. J. Deary, A. Franke, P. Froguel, S. K. Ganesh, M. R. Jarvelin, N. G. Martin, C. Meisinger, B. M. Psaty, T. D. Spector, N. J. Wareham, J. W. Akkerman, M. Ciullo, P. Deloukas, A. Greinacher, S. Jupe, N. Kamatani, J. Khadake, J. S. Kooner, J. Penninger, I. Prokopenko, D. Stemple, D. Toniolo, L. Wernisch, S. Sanna, A. A. Hicks, A. Rendon, M. A. Ferreira, W. H. Ouwe-

hand, and N. Soranzo. New gene functions in megakaryopoiesis and platelet formation. *Nature*, 480(7376):201–208, Dec 2011.

W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. *Proc. Natl. Acad. Sci. U.S.A.*, 70(12):3581–3584, Dec 1973.

A. Gorgens, S. Radtke, M. Mollmann, M. Cross, J. Durig, P. A. Horn, and B. Giebel. Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. *Cell Rep*, 3(5):1539–1552, May 2013.

F. Gounari, R. De Francesco, J. Schmitt, P. van der Vliet, R. Cortese, and H. Stunnenberg. Amino-terminal domain of NF1 binds to DNA as a dimer and activates adenovirus DNA replication. *EMBO J.*, 9(2):559–566, Feb 1990.

T. Graf and T. Enver. Forcing cells to change lineages. *Nature*, 462(7273):587–594, Dec 2009.

M. Grant. *Galen on Food and Diet*. Routledge, 1 edition, 10 2000. ISBN 9780415232326. URL http://amazon.com/o/ASIN/0415232325/.

J. A. Grass, M. E. Boyer, S. Pal, J. Wu, M. J. Weiss, and E. H. Bresnick. GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8811–8816, Jul 2003.

R. M. Gronostajski. Roles of the NFI/CTF gene family in transcription and development. *Gene*, 249(1-2):31–45, May 2000.

J. D. Growney, H. Shigematsu, Z. Li, B. H. Lee, J. Adelsperger, R. Rowan, D. P. Curley, J. L. Kutok, K. Akashi, I. R. Williams, N. A. Speck, and D. G. Gilliland. Loss of Runx1 perturbs adult hematopoiesis and is associated with a myeloproliferative phenotype. *Blood*, 106(2):494–504, Jul 2005.

S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble. Quantifying similarity between motifs. *Genome Biol.*, 8(2):R24, 2007.

A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952, Oct 1999.

P. J. Handagama, J. N. George, M. A. Shuman, R. P. McEver, and D. F. Bainton. Incorporation of a circulating protein into megakaryocyte and platelet granules. *Proc. Natl. Acad. Sci. U.S.A.*, 84(3):861–865, Feb 1987.

Q. L. Hao, J. Zhu, M. A. Price, K. J. Payne, L. W. Barsky, and G. M. Crooks. Identification of a novel, human multilymphoid progenitor in cord blood. *Blood*, 97(12):3683–3690, Jun 2001.

M. Harbers and P. Carninci. Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods*, 2 (7):495–502, Jul 2005.

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 hemizygously deleted in patients with thrombocytopenia. *Immunity*, 13(2):167–177, Aug 2000.

R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11(7):476–486, Jul 2010.

N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39(3):311–318, Mar 2007.

L. Hennighausen, U. Siebenlist, D. Danner, P. Leder, D. Rawlins, P. Rosenfeld, and T. Kelly. High-affinity binding site for a specific nuclear protein in the human IgM gene. *Nature*, 314 (6008):289–292, Mar. 1985.

T. Hisa, S. E. Spence, R. A. Rachel, M. Fujita, T. Nakamura, J. M. Ward, D. E. Devor-Henneman, Y. Saiki, H. Kutsuna, L. Tessarollo, N. A. Jenkins, and N. G. Copeland. Hematopoietic, angiogenic and eye defects in Meis1 mutant animals. *EMBO Journal*, 23(2):450–459, Jan. 2004.

P. Holmfeldt, J. Pardieck, A. C. Saulsberry, S. K. Nandakumar, D. Finkelstein, J. T. Gray, D. A. Persons, and S. McKinney-Freeman. Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood*, Sep 2013.

J. Howell. {SUCCESSFUL} {CASE} {OF} transfusion. *The Lancet*, 9(232):698 – 699, 1828. ISSN 0140-6736. doi: http://dx.doi.org/10.1016/S0140-6736(02)92791-1. URL http://www.sciencedirect.com/science/article/pii/S0140673602927911. <ce:title>Originally published as Volume 1, Issue 232</ce:title>.

M. Hu, D. Krause, M. Greaves, S. Sharkis, M. Dexter, C. Heyworth, and T. Enver. Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.*, 11(6): 774–785, Mar 1997.

M. Ichikawa, T. Asai, T. Saito, S. Seo, I. Yamazaki, T. Yamagata, K. Mitani, S. Chiba, S. Ogawa, M. Kurokawa, and H. Hirai. AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nat. Med.*, 10(3): 299–304, Mar 2004.

H. Igarashi, S. C. Gregory, T. Yokota, N. Sakaguchi, and P. W. Kincade. Transcription from the RAG1 locus marks the earli-

est lymphocyte progenitors in bone marrow. *Immunity*, 17(2): 117–130, Aug 2002.

K. Igarashi, K. Kataoka, K. Itoh, N. Hayashi, M. Nishizawa, and M. Yamamoto. Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins. *Nature*, 367(6463):568–572, Feb 1994.

P. Ikonomi, C. E. Rivera, M. Riordan, G. Washington, A. N. Schechter, and C. T. Noguchi. Overexpression of GATA-2 inhibits erythroid and promotes megakaryocyte differentiation. *Exp. Hematol.*, 28(12):1423–1431, Dec 2000.

N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924): 218–223, Apr 2009.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822):860–921, Feb 2001.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

J. E. Italiano. Unraveling mechanisms that control platelet production. *Semin. Thromb. Hemost.*, 39(1):15–24, Feb 2013.

J. E. Italiano, P. Lecine, R. A. Shivdasani, and J. H. Hartwig. Blood platelets are assembled principally at the ends of proplatelet processes produced by differentiated megakaryocytes. *J. Cell Biol.*, 147(6):1299–1312, Dec 1999.

H. Iwasaki, S. Mizuno, R. A. Wells, A. B. Cantor, S. Watanabe, and K. Akashi. GATA-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte lineages. *Immunity*, 19(3):451–462, Sep 2003.

H. Iwasaki, C. Somoza, H. Shigematsu, E. A. Duprez, J. Iwasaki-Arai, S. Mizuno, Y. Arinobu, K. Geary, P. Zhang, T. Dayaram, M. L. Fenyus, S. Elf, S. Chan, P. Kastner, C. S. Huettner, R. Murray, D. G. Tenen, and K. Akashi. Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood*, 106(5):1590–1600, Sep 2005.

H. Iwasaki, S. Mizuno, Y. Arinobu, H. Ozawa, Y. Mori, H. Shigematsu, K. Takatsu, D. G. Tenen, and K. Akashi. The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes Dev.*, 20(21):3010–3021, Nov 2006.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, Jun 2007.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152 (1-2):327–339, Jan. 2013.

K. A. Jones, J. T. Kadonaga, P. J. Rosenfeld, T. J. Kelly, and R. Tjian. A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication. *Cell*, 48(1):79–89, Jan. 1987.

R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36(16):5221–5231, Sep 2008.

M. S. Jurica, D. Sousa, M. J. Moore, and N. Grigorieff. Three-dimensional structure of C complex spliceosomes by electron microscopy. *Nat. Struct. Mol. Biol.*, 11(3):265–269, Mar 2004.

F. E. Katz, R. Tindle, D. R. Sutherland, and M. F. Greaves. Identification of a membrane glycoprotein associated with haemopoietic progenitor cells. *Leuk. Res.*, 9(2):191–198, 1985.

L. M. Kelly, U. Englmeier, I. Lafon, M. H. Sieweke, and T. Graf. MafB is an inducer of monocytic differentiation. *EMBO J.*, 19 (9):1987–1997, May 2000.

I. A. Khan, S. K. Daya, and R. M. Gowda. Evolution of the theory of circulation. *Int. J. Cardiol.*, 98(3):519–521, Feb 2005.

I. Kim, T. L. Saunders, and S. J. Morrison. Sox17 dependence distinguishes the transcriptional regulation of fetal from adult hematopoietic stem cells. *Cell*, 130(3):470–483, Aug 2007.

J.-A. Kim, Y.-J. Jung, J.-Y. Seoh, S.-Y. Woo, J.-S. Seo, and H.-L. Kim. Gene Expression Profile of Megakaryocytes from Human Cord Blood CD34 +Cells Ex Vivo Expanded by Thrombopoietin. *Stem Cells*, 20(5):402–416, Sept. 2002.

K. Kirito, N. Fox, and K. Kaushansky. Thrombopoietin stimulates Hoxb4 expression: an explanation for the favorable effects of TPO on hematopoietic stem cells. *Blood*, 102(9):3172–3178, Nov. 2003.

T. Kitaguchi, K. Kawakami, and A. Kawahara. Transcriptional regulation of a myeloid-lineage specific gene lysozyme C during zebrafish myelopoiesis. *Mech. Dev.*, 126(5-6):314–323, 2009.

R. J. Klose and A. P. Bird. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.*, 31(2):89–97, Feb 2006.

J. C. Knight. Resolving the variable genome and epigenome in human disease. *J. Intern. Med.*, 271(4):379–391, Apr 2012.

U. Koehler, E. Holinski-Feder, B. Ertl-Wagner, J. Kunz, A. von Moers, H. von Voss, and C. Schell-Apacik. A novel 1p31.3p32.2 deletion involving the NFIA gene detected by array CGH in a patient with macrocephaly and hypoplasia of

the corpus callosum. *European journal of pediatrics*, 169(4):463–468, Apr. 2010.

R. D. Kornberg. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871, May 1974.

T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, Feb 2007.

D. S. Krause, M. J. Fackler, C. I. Civin, and W. S. May. CD34: structure, biology, and clinical utility. *Blood*, 87(1):1–13, Jan 1996.

E. A. Kruse, S. J. Loughran, T. M. Baldwin, E. C. Josefsson, S. Ellis, D. K. Watson, P. Nurden, D. Metcalf, D. J. Hilton, W. S. Alexander, and B. T. Kile. Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. *Proc. Natl. Acad. Sci. U.S.A.*, 106 (33):13814–13819, Aug 2009.

U. Kruse, F. Qian, and A. E. Sippel. Identification of a fourth nuclear factor I gene in chicken by cDNA cloning: NFI-X. *Nucleic acids research*, 19(23):6641, Dec. 1991.

G. Lacaud, S. Robertson, J. Palis, M. Kennedy, and G. Keller. Regulation of hemangioblast development. *Ann. N. Y. Acad. Sci.*, 938:96–107, Jun 2001.

A. H. Lagrue-Lak-Hal, N. Debili, G. Kingbury, C. Lecut, J. P. Le Couedic, J. L. Villeval, M. Jandrot-Perrus, and W. Vainchenker. Expression and function of the collagen receptor GPVI during megakaryocyte maturation. *The Journal of biological chemistry*, 276(18):15316–15325, May 2001.

C. V. Laiosa, M. Stadtfeld, and T. Graf. Determinants of lymphoid-myeloid lineage diversification. *Annu. Rev. Immunol.*, 24:705–738, 2006.

K. Landsteiner. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Zentralblatt Bakteriologie*, 27:357–362, 1900.

S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831, Sep 2012.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862, Oct 2001.

E. Laurenti, S. Doulatov, S. Zandi, I. Plumb, J. Chen, C. April, J. B. Fan, and J. E. Dick. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat. Immunol.*, 14(7):756–763, Jul 2013.

H. J. Lawrence, C. D. Helgason, G. Sauvageau, S. Fong, D. J. Izon, R. K. Humphries, and C. Largman. Mice bearing a targeted interruption of the homeobox gene HOXA9 have defects in myeloid, erythroid, and lymphoid hematopoiesis. *Blood*, 89(6):1922–1930, Mar 1997.

P. Lecine, J. L. Villeval, P. Vyas, B. Swencki, Y. Xu, and R. A. Shivdasani. Mice lacking transcription factor NF-E2 provide in vivo validation of the proplatelet model of thrombocytopoiesis and show a platelet production defect that is intrinsic to megakaryocytes. *Blood*, 92(5):1608–1616, Sep 1998.

S. LeClerc, R. Palaniswami, B. X. Xie, and M. V. Govindan. Molecular cloning and characterization of a factor that binds the human glucocorticoid receptor gene and represses its expression. *J. Biol. Chem.*, 266(26):17333–17340, Sep 1991.

R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864, Oct 2001.

P. A. Leegwater, W. van Driel, and P. C. van der Vliet. Recognition site of nuclear factor I, a sequence-specific DNA-binding protein from HeLa cells that stimulates adenovirus DNA replication. *EMBO Journal*, 4(6):1515–1521, June 1985.

H. G. Leitch, K. R. McEwen, A. Turp, V. Encheva, T. Carroll, N. Grabole, W. Mansfield, B. Nashun, J. G. Knezovich, A. Smith, M. A. Surani, and P. Hajkova. Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.*, 20(3):311–316, Mar 2013.

P. Lengyel and D. Soll. Mechanism of protein biosynthesis. *Bacteriol Rev*, 33(2):264–301, Jun 1969.

B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, 13(4):233–245, Apr 2012.

A. Lepage, M. Leboeuf, J. P. Cazenave, C. de la Salle, F. Lanza, and G. Uzan. The alpha(IIb)beta(3) integrin and GPIb-V-IX complex identify distinct stages in the maturation of CD34(+) cord blood cells to megakaryocytes. *Blood*, 96(13):4169–4177, Dec 2000.

J. E. Levy, O. Jin, Y. Fujiwara, F. Kuo, and N. C. Andrews. Transferrin receptor is necessary for development of erythrocytes and the nervous system. *Nature genetics*, 21(4):396–399, Apr. 1999.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

Y. Liang, G. Van Zant, and S. J. Szilvassy. Effects of aging on the homing and engraftment of murine hematopoietic stem and progenitor cells. *Blood*, 106(4):1479–1487, Aug 2005.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

C. K. Lim, W. Y. K. Hwang, S. E. Aw, and L. Sun. Study of gene expression profile during cord blood-associated megakaryopoiesis. *European journal of haematology*, 81(3):196–208, Sept. 2008.

J. Lindberg and J. Lundeberg. The plasticity of the mammalian transcriptome. *Genomics*, 95(1):1–6, Jan 2010.

R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, May 2008.

G. E. R. Lloyd. *Methods and Problems in Greek Science: Selected Papers*. Cambridge University Press, 1 edition, 4 1991. ISBN 9780521374194. URL http://amazon.com/o/ASIN/0521374197/.

M. R. Loken, V. O. Shah, K. L. Dattilio, and C. I. Civin. Flow cytometric analysis of human bone marrow. II. Normal B lymphocyte development. *Blood*, 70(5):1316–1324, Nov 1987.

A. J. Lopez. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, 32:279–305, 1998.

R. Lower. An account of the experiment of transfusion, practiced upon a man in London. 1667. *Yale J Biol Med*, 75(5-6): 293–297, 2002.

W. Lu, F. Quintero-Rivera, Y. Fan, F. S. Alkuraya, D. J. Donovan, Q. Xi, A. Turbe-Doan, Q.-G. Li, C. G. Campbell, A. L. Shanske, E. H. Sherr, A. Ahmad, R. Peters, B. Rilliet, P. Parvex, A. G. Bassuk, D. J. Harris, H. Ferguson, C. Kelly, C. A. Walsh, R. M. Gronostajski, K. Devriendt, A. Higgins, A. H. Ligon, B. J. Quade, C. C. Morton, J. F. Gusella, and R. L. Maas. NFIA haploinsufficiency is associated with a CNS malformation syndrome and urinary tract defects. *PLoS genetics*, 3(5):e80, May 2007.

K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–260, Sep 1997.

V. Lulli, P. Romania, O. Morsilli, M. Gabbianelli, A. Pagliuca, S. Mazzeo, U. Testa, C. Peschle, and G. Marziali. Overexpression of Ets-1 in human hematopoietic progenitor cells blocks erythroid and promotes megakaryocytic differentiation. *Cell Death Differ.*, 13(7):1064–1074, Jul 2006.

K. W. Lynch. Consequences of regulated pre-mRNA splicing in the immune system. *Nat. Rev. Immunol.*, 4(12):931–940, Dec 2004.

Z. Ma, T. Swigut, A. Valouev, A. Rada-Iglesias, and J. Wysocka. Sequence-specific regulator Prdm14 safeguards mouse ESCs

from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.*, 18(2):120–127, Feb 2011.

I. C. Macaulay, M. R. Tijssen, D. C. Thijssen-Timmer, A. Gusnanto, M. Steward, P. Burns, C. F. Langford, P. D. Ellis, F. Dudbridge, J.-J. Zwaginga, N. A. Watkins, C. E. van der Schoot, and W. H. Ouwehand. Comparative gene expression profiling of in vitro differentiated megakaryocytes and erythroblasts identifies novel activatory and inhibitory platelet membrane proteins. *Blood*, 109(8):3260–3269, Apr. 2007.

R. Majeti, C. Y. Park, and I. L. Weissman. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell*, 1(6):635–645, Dec 2007.

R. Mansson, A. Hultquist, S. Luc, L. Yang, K. Anderson, S. Kharazi, S. Al-Hashmi, K. Liuba, L. Thoren, J. Adolfsson, N. Buza-Vidas, H. Qian, S. Soneji, T. Enver, M. Sigvardsson, and S. E. Jacobsen. Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. *Immunity*, 26(4):407–419, Apr 2007.

M. G. Manz, D. Traver, K. Akashi, M. Merad, T. Miyamoto, E. G. Engleman, and I. L. Weissman. Dendritic cell development from common myeloid progenitors. *Ann. N. Y. Acad. Sci.*, 938: 167–173, Jun 2001a.

M. G. Manz, D. Traver, T. Miyamoto, I. L. Weissman, and K. Akashi. Dendritic cell potentials of early lymphoid and myeloid progenitors. *Blood*, 97(11):3333–3341, Jun 2001b.

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

S. Marco-Sola, M. Sammeth, R. Guigo, and P. Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9(12):1185–1188, Dec 2012.

E. R. Mardis. Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*, 6:287–303, 2013.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, Sept. 2008.

S. B. Marley, J. L. Lewis, R. J. Davidson, I. A. Roberts, I. Dokal, J. M. Goldman, and M. Y. Gordon. Evidence for a continuous decline in haemopoietic cell function from birth: application to evaluating bone marrow failure in children. *Br. J. Haematol.*, 106(1):162–166, Jul 1999.

F. Martin, J. M. van Deursen, R. A. Shivdasani, C. W. Jackson, A. G. Troutman, and P. A. Ney. Erythroid maturation and globin gene expression in mice with combined deficiency of NF-E2 and nrf-2. *Blood*, 91(9):3459–3466, May 1998.

J. F. Martin, S. D. Kristensen, A. Mathur, E. L. Grove, and F. A. Choudry. The causal role of megakaryocyteâ€“platelet hyperactivity in acute coronary syndromes. *Nat Rev Cardiol*, 9(11): 658–670, Nov 2012.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, 7(1):29–59, 2006a. doi: 10.1146/annurev.genom.7.080505.115623. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.7.080505.115623. PMID: 16719718.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006b.

A. J. Matlin and M. J. Moore. Spliceosome assembly and composition. *Adv. Exp. Med. Biol.*, 623:14–35, 2007.

A. J. Matlin, F. Clark, and C. W. Smith. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, 6(5):386–398, May 2005.

H. Mayani, P. Baines, A. Jones, T. Hoy, and A. Jacobs. Effects of recombinant human granulocyte-macrophage colony-stimulating factor (rhGM-CSF) on single CD34-positive hemopoietic progenitors from human bone marrow. *Int. J. Cell Cloning*, 7(1):30–36, Jan 1989.

H. Mayani, W. Dragowska, and P. M. Lansdorp. Characterization of functionally distinct subpopulations of CD34+ cord blood cells in serum-free long-term cultures supplemented with hematopoietic cytokines. *Blood*, 82(9):2664–2672, Nov 1993.

K. B. McCredie, E. M. Hersh, and E. J. Freireich. Cells capable of colony formation in the peripheral blood of man. *Science*, 171(3968):293–294, Jan 1971.

S. McKinney-Freeman, P. Cahan, H. Li, S. A. Lacadie, H. T. Huang, M. Curran, S. Loewer, O. Naveiras, K. L. Kathrein, M. Konantz, E. M. Langdon, C. Lengerke, L. I. Zon, J. J. Collins, and G. Q. Daley. The transcriptional landscape of

hematopoietic stem cell ontogeny. *Cell Stem Cell*, 11(5):701–714, Nov 2012.

R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, W. K. Yung, O. Bogler, J. N. Weinstein, S. VandenBerg, M. Berger, M. Prados, D. Muzny, M. Morgan, S. Scherer, A. Sabo, L. Nazareth, L. Lewis, O. Hall, Y. Zhu, Y. Ren, O. Alvi, J. Yao, A. Hawes, S. Jhangiani, G. Fowler, A. San Lucas, C. Kovar, A. Cree, H. Dinh, J. Santibanez, V. Joshi, M. L. Gonzalez-Garay, C. A. Miller, A. Milosavljevic, L. Donehower, D. A. Wheeler, R. A. Gibbs, K. Cibulskis, C. Sougnez, T. Fennell, S. Mahan, J. Wilkinson, L. Ziaugra, R. Onofrio, T. Bloom, R. Nicol, K. Ardlie, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, M. D. McLellan, J. Wallis, D. E. Larson, X. Shi, R. Abbott, L. Fulton, K. Chen, D. C. Koboldt, M. C. Wendl, R. Meyer, Y. Tang, L. Lin, J. R. Osborne, B. H. Dunford-Shore, T. L. Miner, K. Delehaunty, C. Markovic, G. Swift, W. Courtney, C. Pohl, S. Abbott, A. Hawkins, S. Leong, C. Haipek, H. Schmidt, M. Wiechert, T. Vickery, S. Scott, D. J. Dooling, A. Chinwalla, G. M. Weinstock, E. R. Mardis, R. K. Wilson, G. Getz, W. Winckler, R. G. Verhaak, M. S. Lawrence, M. O'Kelly, J. Robinson, G. Alexe, R. Beroukhim, S. Carter, D. Chiang, J. Gould, S. Gupta, J. Korn, C. Mermel, J. Mesirov, S. Monti, H. Nguyen, M. Parkin, M. Reich, N. Stransky, B. A. Weir, L. Garraway, T. Golub, M. Meyerson, L. Chin, A. Protopopov, J. Zhang, I. Perna, S. Aronson, N. Sathiamoorthy, G. Ren, J. Yao, W. R. Wiedemeyer, H. Kim, S. W. Kong, Y. Xiao, I. S. Kohane, J. Seidman, P. J. Park, R. Kucherlapati, P. W. Laird, L. Cope, J. G. Herman, D. J. Weisenberger, F. Pan, D. Van den Berg, L. Van Neste, J. M. Yi, K. E. Schuebel, S. B. Baylin, D. M. Absher, J. Z. Li, A. Southwick, S. Brady, A. Aggarwal, T. Chung, G. Sherlock, J. D. Brooks, R. M. Myers, P. T. Spellman, E. Purdom, L. R. Jakkula, A. V. Lapuk, H. Marr, S. Dorton, Y. G. Choi, J. Han, A. Ray, V. Wang, S. Durinck, M. Robinson, N. J. Wang, K. Vran-

izan, V. Peng, E. Van Name, G. V. Fontenay, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, C. Brennan, N. D. Socci, A. Olshen, B. S. Taylor, A. Lash, N. Schultz, B. Reva, Y. Antipin, A. Stukalov, B. Gross, E. Cerami, W. Q. Wang, L. X. Qin, V. E. Seshan, L. Villafania, M. Cavatore, L. Borsu, A. Viale, W. Gerald, C. Sander, M. Ladanyi, C. M. Perou, D. N. Hayes, M. D. Topal, K. A. Hoadley, Y. Qi, S. Balu, Y. Shi, J. Wu, R. Penny, M. Bittner, T. Shelton, E. Lenkiewicz, S. Morris, D. Beasley, S. Sanders, A. Kahn, R. Sfeir, J. Chen, D. Nassau, L. Feng, E. Hickey, A. Barker, D. S. Gerhard, J. Vockley, C. Compton, J. Vaught, P. Fielding, M. L. Ferguson, C. Schaefer, J. Zhang, S. Madhavan, K. H. Buetow, F. Collins, P. Good, M. Guyer, B. Ozenberger, J. Peterson, and E. Thomson. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216): 1061–1068, Oct 2008.

K. A. McMahon, S. Y. Hiew, S. Hadjur, H. Veiga-Fernandes, U. Menzel, A. J. Price, D. Kioussis, O. Williams, and H. J. Brady. Mll has a critical role in fetal and adult hematopoietic stem cell self-renewal. *Cell Stem Cell*, 1(3):338–345, Sep 2007.

A. Meigs. Food as a cultural construction. *Food and Foodways*, 2(1):341–357, 1987. doi: 10.1080/07409710.1987.9961926. URL http://www.tandfonline.com/doi/abs/10.1080/07409710.1987.9961926.

M. Meisterernst, I. Gander, L. Rogge, and E. L. Winnacker. A quantitative analysis of nuclear factor I/DNA interactions. *Nucleic acids research*, 16(10):4419–4435, May 1988.

J. Meletis and K. Konstantopoulos. The beliefs, myths, and reality surrounding the word hema (blood) from homer to the present. *Anemia*, 2010:857657, 2010.

M. Merad, P. Sathe, J. Helft, J. Miller, and A. Mortha. The dendritic cell lineage: ontogeny and function of dendritic cells

and their subsets in the steady state and the inflamed setting. *Annu. Rev. Immunol.*, 31:563–604, 2013.

N. Mermod, E. A. O'Neill, T. J. Kelly, and R. Tjian. The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain. *Cell*, 58(4):741–753, Aug. 1989.

T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448 (7153):553–560, Aug 2007.

H. K. Mikkola, J. Klintman, H. Yang, H. Hock, T. M. Schlaeger, Y. Fujiwara, and S. H. Orkin. Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. *Nature*, 421 (6922):547–551, Jan 2003.

A. Miller. British Red Cross: 10 things you didn't know about the red cross. http://blogs.redcross.org.uk/emergencies/2010/09/10-things-you-didnt-know-about-the-red-cross/, 2010. Accessed: 2013-11-02.

K. Miyake, H. Ogata, Y. Nagai, S. Akashi, and M. Kimoto. Innate recognition of lipopolysaccharide by Toll-like receptor 4/MD-2 and RP105/MD-1. *J. Endotoxin Res.*, 6(5):389–391, 2000.

P. Moretti, P. Simmons, P. Thomas, D. Haylock, P. Rathjen, M. Vadas, and R. D'Andrea. Identification of homeobox genes expressed in human haemopoietic progenitor cells. *Gene*, 144 (2):213–219, Jul 1994.

W. D. Morgan, G. T. Williams, R. I. Morimoto, J. Greene, R. E. Kingston, and R. Tjian. Two transcriptional activators, CCAAT-box-binding transcription factor and heat shock transcription factor, interact with a human hsp70 gene promoter. *Molecular and cellular biology*, 7(3):1129–1138, Mar. 1987.

R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, 45(1):81–94, Jul 2008.

S. J. Morrison, A. M. Wandycz, K. Akashi, A. Globerson, and I. L. Weissman. The aging of hematopoietic stem cells. *Nat. Med.*, 2(9):1011–1016, Sep 1996.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

M. R. Muller, Y. Sasaki, I. Stevanovic, E. D. Lamperti, S. Ghosh, S. Sharma, C. Gelinas, D. J. Rossi, M. E. Pipkin, K. Rajewsky, P. G. Hogan, and A. Rao. Requirement for balanced Ca/NFAT signaling in hematopoietic and embryonic development. *Proc. Natl. Acad. Sci. U.S.A.*, 106(17):7034–7039, Apr 2009.

C. E. Muller-Sieburg, R. H. Cho, M. Thoman, B. Adkins, and H. B. Sieburg. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood*, 100(4):1302–1309, Aug 2002.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, Jun 2008.

K. Nagata, R. A. Guggenheimer, T. Enomoto, J. H. Lichy, and J. Hurwitz. Adenovirus DNA replication in vitro: identification of a host factor that stimulates synthesis of the pre-

terminal protein-dCMP complex. *Proceedings of the National Academy of Sciences of the United States of America*, 79(21):6438–6442, Nov. 1982.

S. Nakamura, E. Gohda, T. Matsunaga, I. Yamamoto, and J. Minowada. Production of hepatocyte growth factor by human haematopoietic cell lines. *Cytokine*, 6(3):285–294, May 1994.

M. Namihira, J. Kohyama, K. Semi, T. Sanosaka, B. Deneen, T. Taga, and K. Nakashima. Committed neuronal precursors confer astrocytic potential on residual neural precursor cells. *Dev Cell*, 16(2):245–255, Feb. 2009.

J. P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Paabo, J. K. Pritchard, and E. M. Rubin. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, Nov 2006.

F. Notta, S. Doulatov, E. Laurenti, A. Poeppl, I. Jurisica, and J. E. Dick. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science*, 333 (6039):218–221, Jul 2011.

N. Novershtern, A. Subramanian, L. N. Lawton, R. H. Mak, W. N. Haining, M. E. McConkey, N. Habib, N. Yosef, C. Y. Chang, T. Shay, G. M. Frampton, A. C. B. Drake, I. Leskov, B. Nilsson, F. Preffer, D. Dombkowski, J. W. Evans, T. Liefeld, J. S. Smutko, J. Chen, N. Friedman, R. A. Young, T. R. Golub, A. Regev, and B. L. Ebert. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144(2):296–309, Jan. 2011.

J. Nowock, U. Borgmeyer, A. W. Püschel, R. A. Rupp, and A. E. Sippel. The TGGCA protein binds to the MMTV-LTR, the adenovirus origin of replication, and the BK virus enhancer. *Nucleic acids research*, 13(6):2045–2061, Mar. 1985.

B. Nuez, D. Michalovich, A. Bygrave, R. Ploemacher, and F. Grosveld. Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. *Nature*, 375(6529): 316–318, May 1995.

S. T. Nürnberg, A. Rendon, P. A. Smethurst, D. S. Paul, K. Voss, J. N. Thon, H. Lloyd-Jones, J. G. Sambrook, M. R. Tijssen, HaemGen Consortium, J. E. Italiano, P. Deloukas, B. Göttgens, N. Soranzo, and W. H. Ouwehand. A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood*, 120(24): 4859–4868, Dec. 2012.

T. T. Odell, C. W. Jackson, and D. G. Gosslee. Maturation of rat megakaryocytes studied by microspectrophotometric measurement of DNA. *Proc. Soc. Exp. Biol. Med.*, 119(4):1194–1199, 1965.

Y. Okada, E. Matsuura, Z. Tozuka, R. Nagai, A. Watanabe, K. Matsumoto, K. Yasui, R. W. Jackman, T. Nakano, and T. Doi. Upstream stimulatory factors stimulate transcription through E-box motifs in the PF4 gene in megakaryocytes. *Blood*, 104(7):2027–2034, Oct. 2004.

T. Okuda, J. van Deursen, S. W. Hiebert, G. Grosveld, and J. R. Downing. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell*, 84(2):321–330, Jan. 1996.

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

S. H. Orkin and L. I. Zon. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644, Feb 2008.

J. Palis. Developmental biology: no red cell is an island. *Nature*, 432(7020):964–965, Dec 2004.

L. Pang, H.-H. Xue, G. Szalai, X. Wang, Y. Wang, D. K. Watson, W. J. Leonard, G. A. Blobel, and M. Poncz. Maturation stage-specific regulation of megakaryopoiesis by pointed-domain Ets proteins. *Blood*, 108(7):2198–2206, Oct. 2006.

W. W. Pang, E. A. Price, D. Sahoo, I. Beerman, W. J. Maloney, D. J. Rossi, S. L. Schrier, and I. L. Weissman. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc. Natl. Acad. Sci. U.S.A.*, 108 (50):20012–20017, Dec 2011.

G. Paonessa, F. Gounari, R. Frank, and R. Cortese. Purification of a NF1-like DNA-binding protein from rat liver and cloning of the corresponding cDNA. *EMBO J.*, 7(10):3115–3123, Oct 1988.

V. R. Paralkar and M. J. Weiss. Long noncoding RNAs in biology and hematopoiesis. *Blood*, 121(24):4842–4846, Jun 2013.

C. S. Pareek, R. Smoczynski, and A. Tretyn. Sequencing technologies and genome sequencing. *J. Appl. Genet.*, 52(4):413–435, Nov 2011.

S. R. Patel, J. H. Hartwig, and J. E. Italiano. The biogenesis of platelets from megakaryocyte proplatelets. *J. Clin. Invest.*, 115 (12):3348–3354, Dec 2005.

L. J. Patterson, M. Gering, C. E. Eckfeldt, A. R. Green, C. M. Verfaillie, S. C. Ekker, and R. Patient. The transcription factors Scl and Lmo2 act together during development of the hemangioblast in zebrafish. *Blood*, 109(6):2389–2398, Mar 2007.

D. S. Paul, C. A. Albers, A. Rendon, K. Voss, J. Stephens, P. van der Harst, J. C. Chambers, N. Soranzo, W. H. Ouwehand, P. Deloukas, J. W. Akkerman, C. A. Albers, A. Algra, A. Al-Hussani, H. Allayee, F. Anni, F. W. Asselbergs, A. Attwood, B. Balkau, S. Bandinelli, F. Bastardot, S. Basu, S. E. Baumeister, J. Beckmann, B. Benyamin, G. Biino, J. C.

Bis, L. Bomba, A. Bonnefond, D. I. Boomsma, J. R. Bradley, F. Cambien, J. C. Chambers, M. Ciullo, W. O. Cookson, F. Cucca, A. Cvejic, A. P. D'Adamo, J. Danesh, F. Danjou, D. Das, G. Davies, P. I. de Bakker, R. A. de Boer, E. J. de Geus, I. J. Deary, G. V. Dedoussis, P. Deloukas, M. Dimitriou, C. Dina, A. Doring, U. Elling, D. Ellinghaus, P. Elliott, G. Engstrom, J. Erdmann, T. Esko, D. M. Evans, G. I. Eyjolfsson, M. Falchi, W. Feng, M. A. Ferreira, L. Ferrucci, K. Fischer, A. R. Folsom, P. Fortina, A. Franke, L. Franke, I. H. Frazer, P. Froguel, R. Galanello, S. K. Ganesh, S. F. Garner, P. Gasparini, B. Genser, Q. D. Gibson, C. Gieger, G. Girotto, N. L. Glazer, M. Gogele, A. H. Goodall, A. Greinacher, D. F. Gudbjartsson, C. Hammond, S. E. Harris, J. Hartiala, A. L. Hartikainen, S. L. Hazen, S. R. Heckbert, B. Hedblad, C. Hengstenberg, M. Hersch, A. A. Hicks, H. Holm, J. J. Hottenga, T. Illig, M. R. Jarvelin, J. Jolley, S. Jupe, M. Kahonen, N. Kamatani, S. Kanoni, I. P. Kema, J. P. Kemp, J. Khadake, K. T. Khaw, M. E. Kleber, J. S. Kooner, P. Kovacs, B. Kuhnel, M. C. Kyrtsonis, Y. Labrune, V. Lagou, C. Langenberg, T. Lehtimaki, X. Li, L. Liang, H. Lloyd-Jones, R. J. Loos, L. M. Lopez, T. Lumley, L. P. Lyytikainen, W. Maerz, R. Magi, M. Mangino, N. G. Martin, A. Maschio, I. Mateo Leach, B. McKnight, S. Meacham, S. E. Medland, C. Meisinger, O. Melander, Y. Memari, A. Metspalu, K. Miller, B. D. Mitchell, M. F. Moffatt, G. W. Montgomery, C. Moore, F. Murgia, Y. Nakamura, M. Nauck, G. Navis, I. M. Nolte, U. Nothlings, T. Nutile, Y. Okada, I. Olafsson, P. T. Onundarson, P. F. O'Reilly, W. H. Ouwehand, D. Parracciani, A. Parsa, D. S. Paul, J. M. Penninger, B. W. Penninx, M. Pirastu, N. Pirastu, G. Pistis, E. Porcu, L. Portas, D. Porteous, A. Pouta, P. P. Pramstaller, I. Prokopenko, B. M. Psaty, J. Pullat, A. Radhakrishnan, O. Raitakari, R. Ramirez-Solis, A. Rendon, J. S. Ried, S. M. Ring, A. Robino, J. I. Rotter, D. Ruggiero, A. Ruokonen, C. Sala, A. Saluments, N. J. Samani, J. Sambrook, S. Sanna, D. Schlessinger, C. O. Schmidt, S. Schreiber, H. Schunkert, J. Scott, J. Sehmi, J. Serbanovic-Canic, S. Y. Shin,

A. R. Shuldiner, R. Sladek, J. H. Smit, G. D. Smith, J. G. Smith, N. L. Smith, H. Snieder, N. Soranzo, R. Sorice, T. D. Spector, J. M. Starr, K. Stefansson, D. Stemple, J. Stephens, M. Stumvoll, P. Sulem, A. Takahashi, S. T. Tan, T. Tanaka, C. Tang, W. Tang, W. H. Tang, K. Taylor, A. Tenesa, A. Teumer, S. L. Thein, U. Thorsteinsdottir, D. Toniolo, A. Tonjes, M. Traglia, M. Uda, S. Ulivi, P. van der Harst, E. van der Schoot, W. H. van Gilst, L. J. van Pelt, D. J. van Veldhuisen, N. Verweij, P. M. Visscher, U. Volker, P. Vollenweider, K. Voss, N. J. Wareham, L. Wernisch, H. J. Westra, J. B. Whitfield, H. E. Wichmann, K. L. Wiggins, G. Willemsen, B. R. Winkelmann, G. Wirnsberger, B. H. Wolffenbuttel, J. Yang, T. P. Yang, J. H. Zhang, J. H. Zhao, P. Zitting, and J. J. Zwaginga. Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res.*, 23(7):1130–1141, Jul 2013.

D. Penkov, D. Mateos San Martin, L. C. Fernandez-Diaz, C. A. Rossello, C. Torroja, F. Sanchez-Cabo, H. J. Warnatz, M. Sultan, M. L. Yaspo, A. Gabrieli, V. Tkachuk, A. Brendolan, F. Blasi, and M. Torres. Analysis of the DNA-binding profile and function of TALE homeoproteins reveals their specialization and specific interactions with Hox genes/proteins. *Cell Rep*, 3(4):1321–1333, Apr 2013.

D. A. Persons, J. A. Allay, E. R. Allay, R. A. Ashmun, D. Orlic, S. M. Jane, J. M. Cunningham, and A. W. Nienhuis. Enforced expression of the GATA-2 transcription factor blocks normal hematopoiesis. *Blood*, 93(2):488–499, Jan 1999.

L. Pevny, M. C. Simon, E. Robertson, W. H. Klein, S. F. Tsai, V. D'Agati, S. H. Orkin, and F. Costantini. Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature*, 349(6306):257–260, Jan 1991.

N. Pineault, C. D. Helgason, H. J. Lawrence, and R. K. Humphries. Differential expression of Hox, Meis1, and Pbx1 genes in primitive cells throughout murine hematopoietic ontogeny. *Exp. Hematol.*, 30(1):49–57, Jan 2002.

S. D. Power. *Harvey, William, 1578-1657; Harvey, William, 1578-1657; Physicians*. London : T. Fisher Unwin, 1897.

A. Prokhortchouk, B. Hendrich, H. Jørgensen, A. Ruzov, M. Wilm, G. Georgiev, A. Bird, and E. Prokhortchouk. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.*, 15(13):1613–1618, Jul 2001.

E. Prokhortchouk and P. A. Defossez. The cell biology of DNA methylation in mammals. *Biochim. Biophys. Acta*, 1783(11): 2167–2173, Nov 2008.

F. Qian, U. Kruse, P. Lichter, and A. E. Sippel. Chromosomal localization of the four genes (NFIA, B, C, and X) for the human transcription factor nuclear factor I by FISH. *Genomics*, 28(1):66–73, July 1995.

A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470 (7333):279–283, Feb 2011.

V. Randrianarison-Huetz, B. Laurent, V. Bardet, G. C. Blobe, F. Huetz, and D. Dumenil. Gfi-1B controls human erythroid and megakaryocytic differentiation by regulating TGF-beta signaling at the bipotent erythro-megakaryocytic progenitor stage. *Blood*, 115(14):2784–2795, Apr 2010.

H. Raslova, A. Kauffmann, D. Sekkaï, H. Ripoche, F. Larbret, T. Robert, D. T. Le Roux, G. Kroemer, N. Debili, P. Dessen, V. Lazar, and W. Vainchenker. Interrelation between polyploidization and megakaryocyte differentiation: a gene profiling approach. *Blood*, 109(8):3225–3234, Apr. 2007.

K. Ravid, J. Lu, J. M. Zimmet, and M. R. Jones. Roads to polyploidy: the megakaryocyte example. *J. Cell. Physiol.*, 190(1): 7–20, Jan 2002.

V. A. Reddy, A. Iwama, G. Iotzova, M. Schulz, A. Elsasser, R. K. Vangala, D. G. Tenen, W. Hiddemann, and G. Behre. Granulocyte inducer C/EBPalpha inactivates the myeloid master regulator PU.1: possible role in lineage commitment decisions. *Blood*, 100(2):483–490, Jul 2002.

J. L. Richardson, R. A. Shivdasani, C. Boers, J. H. Hartwig, and J. E. Italiano. Mechanisms of organelle transport and capture along proplatelets during platelet production. *Blood*, 106(13): 4066–4075, Dec 2005.

K. Robasky and M. L. Bulyk. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, 39(Database issue):D124–128, Jan 2011.

L. Robb, I. Lyons, R. Li, L. Hartley, F. Köntgen, R. P. Harvey, D. Metcalf, and C. G. Begley. Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proceedings of the National Academy of Sciences of the United States of America*, 92(15):7075–7079, July 1995.

G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657, Aug 2007.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

J. Rocca and Galen. *Galen on the Brain: Anatomical Knowledge and Physiological Speculation in the Second Century Ad (Studies in Ancient Medicine) (Multilingual Edition)*. Brill Academic Pub, 1 edition, 2 2003. ISBN 9789004125124. URL http://amazon.com/o/ASIN/9004125124/.

K. Rogers. *New Thinking About Genetics*. 21st Century Science. Britannica Educational Publishing, 2010. ISBN 9781615301690. URL http://books.google.co.uk/books?id=Pcu4KXUI4wcC.

E. Roulet, M. T. Armentero, G. Krey, B. Corthesy, C. Dreyer, N. Mermod, and W. Wahli. Regulation of the DNA-binding and transcriptional activities of Xenopus laevis NFI-X by a novel C-terminal domain. *Mol. Cell. Biol.*, 15(10):5552–5562, Oct 1995.

P. Rous. Karl landsteiner. 1868-1943. *Obituary Notices of Fellows of the Royal Society*, 5(15):294–324, 1947. doi: 10.1098/rsbm.1947.0002. URL http://rsbm.royalsocietypublishing.org/content/obits/5/15/294.short.

J. W. Rowley, A. J. Oler, N. D. Tolley, B. N. Hunter, E. N. Low, D. A. Nix, C. C. Yost, G. A. Zimmerman, and A. S. Weyrich. Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. *Blood*, 118(14):e101–11, Oct. 2011.

Y. Ruan, P. Le Ber, H. H. Ng, and E. T. Liu. Interrogating the transcriptome. *Trends Biotechnol.*, 22(1):23–30, Jan 2004.

R. A. Rupp, U. Kruse, G. Multhaup, U. Göbel, K. Beyreuther, and A. E. Sippel. Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B and NFI-C with homologues in mammalian genomes. *Nucleic acids research*, 18(9):2607–2616, May 1990.

F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74 (12):5463–5467, Dec 1977.

A. Sanjuan-Pla, I. C. Macaulay, C. T. Jensen, P. S. Woll, T. C. Luis, A. Mead, S. Moore, C. Carella, S. Matsuoka, T. B. Jones, O. Chowdhury, L. Stenson, M. Lutteropp, J. C. Green, R. Facchini, H. Boukarabila, A. Grover, A. Gambardella, S. Thongjuea, J. Carrelha, P. Tarrant, D. Atkinson, S. A. Clark, C. Nerlov, and S. E. Jacobsen. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*, Aug 2013.

C. Santoro, N. Mermod, P. C. Andrews, and R. Tjian. A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs. *Nature*, 334(6179):218–224, July 1988.

M. Sanyal, J. W. Tung, H. Karsunky, H. Zeng, L. Selleri, I. L. Weissman, L. A. Herzenberg, and M. L. Cleary. B-cell development fails in the absence of the Pbx1 proto-oncogene. *Blood*, 109(10):4191–4199, May 2007.

N. Sasai and P. A. Defossez. Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. *Int. J. Dev. Biol.*, 53(2-3):323–334, 2009.

G. Sauvageau, P. M. Lansdorp, C. J. Eaves, D. E. Hogge, W. H. Dragowska, D. S. Reid, C. Largman, H. J. Lawrence, and R. K. Humphries. Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells. *Proc. Natl. Acad. Sci. U.S.A.*, 91(25):12223–12227, Dec 1994.

G. Sauvageau, U. Thorsteinsdottir, C. J. Eaves, H. J. Lawrence, C. Largman, P. M. Lansdorp, and R. K. Humphries. Overexpression of HOXB4 in hematopoietic cells causes the selective

expansion of more primitive populations in vitro and in vivo. *Genes Dev.*, 9(14):1753–1765, Jul 1995.

C. A. Schiffer. Diagnosis and management of refractoriness to platelet transfusion. *Blood Rev.*, 15(4):175–180, Dec 2001.

R. Schofield. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood cells*, 4 (1-2):7–25, 1978.

D. E. Schones, K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, Mar 2008.

H. D. Schwer, P. Lecine, S. Tiwari, J. E. Italiano, J. H. Hartwig, and R. A. Shivdasani. A lineage-restricted and divergent beta-tubulin isoform is essential for the biogenesis, structure and function of blood platelets. *Curr. Biol.*, 11(8):579–586, Apr 2001.

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

E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, Sep 2013.

J. A. Shavit, H. Motohashi, K. Onodera, J. Akasaka, M. Yamamoto, and J. D. Engel. Impaired megakaryopoiesis and behavioral defects in mafG-null mutant mice. *Genes Dev.*, 12(14):2164–2174, Jul 1998.

W. F. Shen, J. C. Montgomery, S. Rozenfeld, J. J. Moskow, H. J. Lawrence, A. M. Buchberg, and C. Largman. AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol. Cell. Biol.*, 17(11):6448–6458, Nov 1997.

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

J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145, Oct 2008.

J. Shendure and E. Lieberman Aiden. The expanding scope of DNA sequencing. *Nat. Biotechnol.*, 30(11):1084–1094, Nov 2012.

C. J. Sherr. Colony-stimulating factor-1 receptor. *Blood*, 75(1): 1–12, Jan 1990.

H. Shigematsu, B. Reizis, H. Iwasaki, S. Mizuno, D. Hu, D. Traver, P. Leder, N. Sakaguchi, and K. Akashi. Plasmacytoid dendritic cells activate lymphoid-specific genetic programs irrespective of their cellular origin. *Immunity*, 21 (1):43–53, Jul 2004.

M. H. Shim, A. Hoover, N. Blake, J. G. Drachman, and J. A. Reems. Gene expression profile of primary human CD34+CD38lo cells differentiating along the megakaryocyte lineage. *Exp. Hematol.*, 32(7):638–648, Jul 2004.

T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.*, 100(26):15776–15781, Dec 2003.

R. A. Shivdasani, E. L. Mayer, and S. H. Orkin. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature*, 373(6513):432–434, Feb. 1995a.

R. A. Shivdasani, M. F. Rosenblatt, D. Zucker-Franklin, C. W. Jackson, P. Hunt, C. J. Saris, and S. H. Orkin. Transcription factor NF-E2 is required for platelet formation independent

of the actions of thrombopoietin/MGDF in megakaryocyte development. *Cell*, 81(5):695–704, Jun 1995b.

R. A. Shivdasani, Y. Fujiwara, M. A. McDevitt, and S. H. Orkin. A lineage-selective knockout establishes the critical role of transcription factor GATA-1 in megakaryocyte growth and platelet development. *EMBO J.*, 16(13):3965–3973, Jul 1997.

T. Shu, K. G. Butz, C. Plachez, R. M. Gronostajski, and L. J. Richards. Abnormal development of forebrain midline glia and commissural projections in Nfia knock-out mice. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(1):203–212, Jan. 2003.

L. G. Smith, I. L. Weissman, and S. Heimfeld. Clonal analysis of hematopoietic stem-cell differentiation in vivo. *Proc. Natl. Acad. Sci. U.S.A.*, 88(7):2788–2792, Apr 1991.

J. Soler. The semiotics of food in the bible. *Food and culture: A reader*, pages 55–66, 1997.

M. J. Solomon, P. L. Larsen, and A. Varshavsky. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, Jun 1988.

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

F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, Sep 2012.

D. D. Spyropoulos, P. N. Pharr, K. R. Lavenburg, P. Jackers, T. S. Papas, M. Ogawa, and D. K. Watson. Hemorrhage, impaired hematopoiesis, and lethality in mouse embryos carrying a targeted disruption of the Fli1 transcription factor. *Mol. Cell. Biol.*, 20(15):5643–5652, Aug 2000.

L. M. Starnes, A. Sorrentino, M. Ferracin, M. Negrini, E. Pelosi, C. Nervi, and C. Peschle. A transcriptome-wide approach reveals the key contribution of NFI-A in promoting erythroid differentiation of human CD34(+) progenitors and CML cells. *Leukemia*, 24(6):1220–1223, June 2010.

T. K. Starr, S. C. Jameson, and K. A. Hogquist. Positive and negative selection of T cells. *Annu. Rev. Immunol.*, 21:139–176, 2003.

G. Steele-Perkins, K. G. Butz, G. E. Lyons, M. Zeichner-David, H.-J. Kim, M.-I. Cho, and R. M. Gronostajski. Essential role for NFI-C/CTF transcription-replication factor in tooth root development. *Molecular and cellular biology*, 23(3):1075–1084, Feb. 2003.

G. Steele-Perkins, C. Plachez, K. G. Butz, G. Yang, C. J. Bachurski, S. L. Kinsman, E. D. Litwack, L. J. Richards, and R. M. Gronostajski. The transcription factor gene Nfib is essential for both lung maturation and brain development. *Molecular and cellular biology*, 25(2):685–698, Jan. 2005.

D. Stites, A. Terr, and T. Parslow. *Medical Immunology*. Appleton & Lange, 9th edition edition, 4 1997. ISBN 9780838562789. URL http://amazon.co.uk/o/ASIN/0838562787/.

H. J. Sutherland, C. J. Eaves, A. C. Eaves, W. Dragowska, and P. M. Lansdorp. Characterization and partial purification of human marrow cells capable of initiating long-term hematopoiesis in vitro. *Blood*, 74(5):1563–1570, Oct 1989.

D. Sweetman and A. Munsterberg. The vertebrate spalt genes in development and disease. *Dev. Biol.*, 293(2):285–293, May 2006.

N. Takayama, S. Nishimura, S. Nakamura, T. Shimizu, R. Ohnishi, H. Endo, T. Yamaguchi, M. Otsu, K. Nishimura, M. Nakanishi, A. Sawaguchi, R. Nagai, K. Takahashi, S. Yamanaka, H. Nakauchi, and K. Eto. Transient activation

of c-MYC expression is critical for efficient platelet generation from human induced pluripotent stem cells. *J. Exp. Med.*, 207 (13):2817–2830, Dec 2010.

D. Talbot and F. Grosveld. The 5'HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 binding sites. *EMBO J.*, 10(6):1391–1398, Jun 1991.

M. J. Tanner. The major integral proteins of the human red cell. *Baillière's clinical haematology*, 6(2):333–356, June 1993.

E. Tenedini, M. E. Fagioli, N. Vianelli, P. L. Tazzari, F. Ricci, E. Tagliafico, P. Ricci, L. Gugliotta, G. Martinelli, S. Tura, M. Baccarani, S. Ferrari, and L. Catani. Gene expression profiling of normal and malignant CD34-derived megakaryocytic cells. *Blood*, 104(10):3126–3135, Nov 2004.

L. W. Terstappen, S. Huang, M. Safford, P. M. Lansdorp, and M. R. Loken. Sequential generations of hematopoietic colonies derived from single nonlineage-committed CD34+CD38- progenitor cells. *Blood*, 77(6):1218–1227, Mar 1991.

E. D. Thomas, H. L. Lochte, W. C. LU, and J. W. Ferrebee. Intravenous infusion of bone marrow in patients receiving radiation and chemotherapy. *N. Engl. J. Med.*, 257(11):491–496, Sep 1957.

M. R. Tijssen and C. Ghevaert. Transcription factors in late megakaryopoiesis and related platelet disorders. *J. Thromb. Haemost.*, 11(4):593–604, Apr 2013.

M. R. Tijssen, A. Cvejic, A. Joshi, R. L. Hannah, R. Ferreira, A. Forrai, D. C. Bellissimo, S. H. Oram, P. A. Smethurst, N. K. Wilson, X. Wang, K. Ottersbach, D. L. Stemple, A. R. Green, W. H. Ouwehand, and B. Gottgens. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev. Cell*, 20(5):597–609, May 2011.

J. E. Till and E. A. McCulloch. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.*, 14:213–222, Feb 1961.

A. Tomer. Human marrow megakaryocyte differentiation: multiparameter correlative analysis identifies von Willebrand factor as a sensitive and distinctive marker for early (2N and 4N) megakaryocytes. *Blood*, 104(9):2722–2727, Nov. 2004a.

A. Tomer. Human marrow megakaryocyte differentiation: multiparameter correlative analysis identifies von Willebrand factor as a sensitive and distinctive marker for early (2N and 4N) megakaryocytes. *Blood*, 104(9):2722–2727, Nov 2004b.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010a.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, May 2010b.

C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578, Mar 2012.

D. Traver, K. Akashi, M. Manz, M. Merad, T. Miyamoto, E. G. Engleman, and I. L. Weissman. Development of CD8alpha-positive dendritic cells from a common myeloid progenitor. *Science*, 290(5499):2152–2154, Dec 2000.

F. Y. Tsai, G. Keller, F. C. Kuo, M. Weiss, J. Chen, M. Rosenblatt, F. W. Alt, and S. H. Orkin. An early haematopoietic defect

in mice lacking the transcription factor GATA-2. *Nature*, 371 (6494):221–226, Sep 1994.

A. P. Tsang, Y. Fujiwara, D. B. Hom, and S. H. Orkin. Failure of megakaryopoiesis and arrested erythropoiesis in mice lacking the GATA-1 transcriptional cofactor FOG. *Genes Dev.*, 12(8): 1176–1188, Apr 1998.

N. Tsuneyoshi, T. Sumi, H. Onda, H. Nojima, N. Nakatsuji, and H. Suemori. PRDM14 suppresses expression of differentiation marker genes in human embryonic stem cells. *Biochem. Biophys. Res. Commun.*, 367(4):899–905, Mar 2008.

R. S. Tubbs, M. Loukas, M. M. Shoja, M. R. Ardalan, and W. J. Oakes. Richard lower (1631-1691) and his early contributions to cardiology. *International Journal of Cardiology*, 128(1):17 – 21, 2008. ISSN 0167-5273. doi: http://dx.doi.org/10.1016/j.ijcard.2007.11.069. URL http://www.sciencedirect.com/science/article/pii/S0167527307020840.

E. Turro, S. Y. Su, A. Goncalves, L. J. Coin, S. Richardson, and A. Lewin. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.*, 12 (2):R13, 2011.

H. Ueno, T. Sato, S. Yamamoto, K. Tanaka, S. Ohkawa, H. Takagi, O. Yokosuka, J. Furuse, H. Saito, A. Sawaki, H. Kasugai, Y. Osaki, S. Fujiyama, K. Sato, K. Wakabayashi, and T. Okusaka. Randomized, double-blind, placebo-controlled trial of bovine lactoferrin in patients with chronic hepatitis C. *Cancer Sci.*, 97(10):1105–1110, Oct 2006.

I. Ulitsky and D. P. Bartel. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46, Jul 2013.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–263, Apr. 2009.

V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.

A. Verger and M. Duterque-Coquillaud. When Ets transcription factors meet their partners. *Bioessays*, 24(4):362–370, Apr 2002.

E. Vivier, D. H. Raulet, A. Moretta, M. A. Caligiuri, L. Zitvogel, L. L. Lanier, W. M. Yokoyama, and S. Ugolini. Innate or adaptive immunity? The example of natural killer cells. *Science*, 331(6013):44–49, Jan 2011.

P. Volkel and P. O. Angrand. The control of histone lysine methylation in epigenetic regulation. *Biochimie*, 89(1):1–20, Jan 2007.

A. von Decastello and A. Sturli. Zur Ueber die Isoagglutinine im Serum gesunder und kranker Menschen. *Mfinch med Wschr*, 49:1090–1095, 1902.

M. T. Walton. The first blood transfusion: French of English? *Med Hist*, 18(4):360–364, Oct 1974.

G. G. Wang, M. P. Pasillas, and M. P. Kamps. Persistent transactivation by meis1 replaces hox function in myeloid leukemogenesis models: evidence for co-occupancy of meis1-pbx and hox-pbx complexes on promoters of leukemia-associated genes. *Mol. Cell. Biol.*, 26(10):3902–3916, May 2006.

J. C. Wang, M. Doedens, and J. E. Dick. Primitive human hematopoietic cells are enriched in cord blood compared with adult bone marrow or mobilized peripheral blood as measured by the quantitative in vivo SCID-repopulating cell assay. *Blood*, 89(11):3919–3924, Jun 1997.

K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, 38(18):e178, Oct 2010.

L. D. Wang and A. J. Wagers. Dynamic niches in the origination and differentiation of haematopoietic stem cells. *Nature reviews. Molecular cell biology*, 12(10):643–655, Oct. 2011.

Q. Wang, T. Stacy, J. D. Miller, A. F. Lewis, T. L. Gu, X. Huang, J. H. Bushweller, J. C. Bories, F. W. Alt, G. Ryan, P. P. Liu, A. Wynshaw-Boris, M. Binder, M. Marín-Padilla, A. H. Sharpe, and N. A. Speck. The CBFbeta subunit is essential for CBFalpha2 (AML1) function in vivo. *Cell*, 87(4):697–708, Nov. 1996.

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

A. J. Warren, W. H. Colledge, M. B. Carlton, M. J. Evans, A. J. Smith, and T. H. Rabbitts. The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. *Cell*, 78(1):45–57, Jul 1994.

N. A. Watkins, A. Gusnanto, and B. de Bono. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *e-Blood*, 2009.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171 (4356):737–738, Apr 1953.

J. B. West. Ibn al-Nafis, the pulmonary circulation, and the Islamic Golden Age. *J. Appl. Physiol.*, 105(6):1877–1880, Dec 2008.

I. Whitehouse, A. Flaus, B. R. Cairns, M. F. White, J. L. Workman, and T. Owen-Hughes. Nucleosome mobilization catalysed by the yeast SWI/SNF complex. *Nature*, 400(6746):784–787, Aug 1999.

WHO. World Health Organiazation: Blood safety and availability. http://www.who.int/mediacentre/factsheets/fs279/en/, 2013. Accessed: 2013-11-02.

B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bahler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, Jun 2008.

N. K. Wilson, S. D. Foster, X. Wang, K. Knezevic, J. Schutte, P. Kaimakis, P. M. Chilarska, S. Kinston, W. H. Ouwehand, E. Dzierzak, J. E. Pimanda, M. F. de Bruijn, and B. Gottgens. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, 7(4):532–544, Oct 2010.

T. D. Wu and S. Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7): 873–881, Apr 2010.

H. Xu, L. Handoko, X. Wei, C. Ye, J. Sheng, C. L. Wei, F. Lin, and W. K. Sung. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, 26(9):1199–1204, May 2010.

J. Xu, S. D. Pope, A. R. Jazirehi, J. L. Attema, P. Papathanasiou, J. A. Watts, K. S. Zaret, I. L. Weissman, and S. T. Smale. Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 104(30):12377–12382, Jul 2007.

J. Xu, Z. Shao, K. Glass, D. E. Bauer, L. Pinello, B. Van Handel, S. Hou, J. A. Stamatoyannopoulos, H. K. Mikkola, G. C. Yuan, and S. H. Orkin. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell*, 23(4):796–811, Oct 2012.

K. Yamada, J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H.

Chang, J. M. Lee, M. Toriumi, M. M. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, and J. R. Ecker. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, 302(5646):842–846, Oct 2003.

Y. Yamada, A. J. Warren, C. Dobson, A. Forster, R. Pannell, and T. H. Rabbitts. The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. *Proc. Natl. Acad. Sci. U.S.A.*, 95(7):3890–3895, Mar 1998.

Y. Yang, C. K. Hwang, U. M. D'Souza, S. H. Lee, E. Junn, and M. M. Mouradian. Three-amino acid extension loop homeodomain proteins Meis2 and TGIF differentially regulate transcription. *The Journal of biological chemistry*, 275(27): 20734–20741, July 2000.

Z. F. Yang, K. Drumea, J. Cormier, J. Wang, X. Zhu, and A. G. Rosmarin. GABP transcription factor is required for myeloid differentiation, in part, through its control of Gfi-1 expression. *Blood*, 118(8):2243–2253, Aug 2011.

G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X. D. Fu, and F. H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, 16(2):130–137, Feb 2009.

H. Yoshihara, F. Arai, K. Hosokawa, T. Hagiwara, K. Takubo, Y. Nakamura, Y. Gomei, H. Iwasaki, S. Matsuoka, K. Miyamoto, H. Miyazaki, T. Takahashi, and T. Suda. Thrombopoietin/MPL signaling regulates hematopoietic stem cell

quiescence and interaction with the osteoblastic niche. *Cell Stem Cell*, 1(6):685–697, Dec 2007.

C. Yu, A. B. Cantor, H. Yang, C. Browne, R. A. Wells, Y. Fujiwara, and S. H. Orkin. Targeted deletion of a high-affinity GATA-binding site in the GATA-1 promoter leads to selective loss of the eosinophil lineage in vivo. *J. Exp. Med.*, 195(11):1387–1395, Jun 2002.

S. Yu, D. M. Zhao, R. Jothi, and H. H. Xue. Critical requirement of GABPalpha for normal T cell development. *J. Biol. Chem.*, 285(14):10179–10188, Apr 2010.

S. Yu, K. Cui, R. Jothi, D. M. Zhao, X. Jing, K. Zhao, and H. H. Xue. GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells. *Blood*, 117(7):2166–2178, Feb 2011.

L. Zhang, A. Eddy, Y. T. Teng, M. Fritzler, M. Kluppel, F. Melet, and A. Bernstein. An immunological renal disease in transgenic mice that overexpress Fli-1, a member of the ets family of transcription factor genes. *Mol. Cell. Biol.*, 15(12):6961–6970, Dec 1995.

Y. Zhang and D. Reinberg. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.*, 15(18):2343–2360, Sep 2001.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.

Y. Zhao, Y. Zhang, S. Wang, Z. Hua, and J. Zhang. The clock gene Per2 is required for normal platelet formation and function. *Thromb. Res.*, 127(2):122–130, Feb 2011.

V. W. Zhou, A. Goren, and B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, 12(1):7–18, Jan 2011.

K. Århem. Maasai food symbolism: The cultural connotations of milk, meat, and blood in the pastoral maasai diet. *Anthropos*, 84(1/3):pp. 1–23, 1989. ISSN 02579774. URL http://www.jstor.org/stable/40461671.