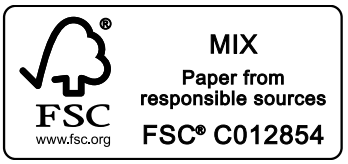
Transcription Factors, Chromatin Loops & Blood Cells

Ralph Stadhouders

Colofon

ISBN/EAN: 9789461087300

Cover illustration: Kristel Met
Lay-out: Ralph Stadhouders
Printed by: Gildeprint Drukkerijen - Enschede



The studies presented in this thesis were mainly performed at the department of Cell Biology of the Erasmus University Medical Center, Rotterdam, The Netherlands.

Transcription Factors, Chromatin Loops & Blood Cells

Transcriptiefactoren, chromatine loops & bloedcellen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 12 september 2014 om 13.30 uur

door

Ralph Stadhouders

geboren te Bergen op Zoom

Promotiecommissie

Promotor:	Prof.dr. F.G. Grosveld
Overige leden:	Prof.dr. J.N.J. Philipsen Prof.dr. R.W. Hendriks Prof.dr. D. Huylebroeck
Copromotor:	Dr. E. Soler

Table of Contents

List of Abbreviations

3C	Chromosome Conformation Capture
3C-Seq/4C-Seq	3C coupled to next-generation Sequencing
ChIP	Chromatin Immunoprecipitation
ChIP-Seq	ChIP coupled to next-generation Sequencing
PCR	Polymerase Chain Reaction
FACS	Fluorescence-Activated Cell Sorting
FISH	Fluorescence In Situ Hybridization
BAC	Bacterial Artificial Chromosome
ICM	Inner Cell Mass
ES (cell)	Embryonic Stem (cell)
HSC	Hematopoietic Stem Cell
EPO	Erythropoietin
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
TSS	Transcription Start Site
GTF	General Transcription Factor
TF	Transcription Factor
PIC	Pre-Initiation Complex
RNAPII or polII	RNA Polymerase II
CTD	Carboxyl-Terminal Domain
GRE	Gene Regulatory Element
LCR	Locus Control Region
HDAC	Histone Deacetylase
HAT	Histone Acetyltransferase
BRD4	Bromodomain containing 4
RPM	Reads Per Million
MYB	Myeloblastosis oncogene
HBS1L	HBS1-like (S. cerevisiae)
RNAi	RNA interference
FL	Fetal Liver
FB	Fetal Brain
BM	Bone Marrow
MEL	Murine Erythroleukemia
HEP	Human Erythroid Progenitor
DNAseI-HS	DNAseI Hypersensitive
SNP	Single Nucleotide Polymorphism
GWAS	Genome-Wide Association Study
HMIP	HBS1L-MYB Intergenic Polymorphism
ACH	Active Chromatin Hub
HbA	Adult Hemoglobin
HbF	Fetal Hemoglobin
CTCF	CCCTC-binding factor
BCL11A	B-cell CLL/lymphoma 11A
LDB1	LIM Domain Binding 1
TAL1	T-cell acute lymphocytic leukemia 1
LMO2	LIM domain Only 2
GATA1	GATA binding protein 1
ETO2	Eight-Twenty-One 2 (MTG16)
IRF2BP2	Interferon Regulatory Factor 2 Binding Protein 2
KLF1	Kruppel-Like Factor 1 (Erythroid)
CDK9	Cyclin-Dependent Kinase 9
TIF1γ	Transcription Intermediary Factor 1-Gamma (TRIM33)
DRB	5,6-dichloro-1-β-D-ribofuranosylbenzimidazole
Ig	Immunoglobulin
(pre-)BCR	(pre-)B Cell Receptor
GLT	Germline Transcription
iEκ	Intronic κ enhancer
3'Eκ	3'κ enhancer
Sis	Silencer in Intervening Sequence
Btk	Bruton's tyrosine kinase
Slp65	SH2 domain-containing Leukocyte Protein of 65 kDa

Scope of this thesis

Animal development and life demands strict control over gene expression. Our genes need to be expressed at the correct level in certain tissues and at specific time points, so that the functional molecules they encode are present at the right place and time. The regulation of gene expression is a complicated process, and its perturbation often results in developmental defects or disease. The work described in this thesis aims to contribute to our understanding of gene regulatory mechanisms in mammals.

The thesis starts with an introductory Chapter (Chapter 1). The first half of this Chapter contains an illustrated introduction to the basic concepts underlying mammalian development, the function of our DNA and the genes within. The second half describes our current understanding of gene regulatory mechanisms and their relevance to human health.

Chapters 2 to 8 contain the experimental work performed during the course of the PhD studies. Herein, I focus on studying the control of gene expression during blood cell development in humans and mice. In Chapter 2 I describe the identification of novel regulatory proteins and mechanisms that repress the late erythroid-specific transcriptome in immature erythroid progenitor cells. Chapters 3 and 4 outline our efforts to adapt 3C/4C methodology, an increasingly popular method used to uncover functional connections between gene regulatory elements, to the current high-throughput sequencing technology. This included the development of a bioinformatics pipeline to facilitate subsequent data analysis (Chapter 4). Chapters 5 and 6 describe a detailed analysis of the regulatory mechanisms that control the expression of the *Myb*/*MYB* proto-oncogene during erythroid development. Initial studies in mouse model systems (Chapter 5) were followed up by analyses of primary human erythroid cells, through which we uncovered the molecular relationship between non-coding genetic variation, *MYB* regulation and clinically relevant human erythroid parameters (Chapter 6). Chapters 7 and 8 contain the experimental work performed on early B lymphocyte development *in vivo*. In particular, we have studied the role of the insulator protein CTCF on B cell development and V(D)J recombination (Chapter 7). In a second study (Chapter 8) we have focussed on pre-BCR signalling, and how B cell development, gene expression and Igκ locus recombination are influenced by these signals.

In the final Chapter of this thesis (Chapter 9) I summarize the results of the experimental research described in Chapters 2 to 8. In addition, I consider the implications of these results for our understanding of the specific hematopoietic developmental processes they describe, as well as for gene regulatory mechanisms in general. Directions for future research and preliminary results of several follow-up experiments are also provided.

Chapter 1

General Introduction

Cellular differentiation and mammalian development

Early ideas on embryonic development and the rise of cell theory

The development of a complex multicellular organism from a single fertilized egg is a spectacular event that has fascinated mankind for over 2000 years. Aristotle (384-322 B.C.) was the first to present a systematic theory on embryogenesis. In '*On the Generation of Animals*', the first known scientific work on embryology[1], he postulates that organisms develop in a gradual manner from a 'formless' egg. This process was referred to as epigenesis, which Aristotle believed was guided by a 'soul'. In the mid-17th century, resistance to the theory of epigenesis rose to prominence as microscopy pioneers refuted it in favor of a preformation theory[2]. Proponents of 'preformationism' claimed that all the adult parts of an organism were already present in the egg, which then merely increased in size or number during embryonic development. Iconic for the 17th century dominant preformation dogma became a sketch from the Dutch mathematician and physicist Nicolaas Hartsoeker (1656-1725), who postulated the existence of miniature humans in the heads of sperm cells[3] (Figure 1A).

As microscopy tools and technology rapidly advanced during the 18th and especially 19th century, scientists realized that 'the elementary parts of all tissues are formed of cells'[4]. This concept culminated in the establishment of classical cell theory by the late 19th century and the demise of preformationism. Cell theory, one of the foundations of modern biology, comprises of three fundamental properties[5]:

1) All living organisms are composed of one or more cells
2) The cell is the most basic unit of life
3) All cells arise from pre-existing, living cells

Research throughout the 20th century has yielded tremendous insight into the biology of cells and how development, including that of vertebrate animal models, is accomplished. We have categorized and catalogued much of the living natural world around us, have developed a detailed understanding of the inner workings of a cell and have carefully dissected the function of most types of cells. The mechanisms underlying 'pillar three' of classical cell theory listed above have been particularly challenging to uncover for many developmental processes. This is especially true for embryogenesis, as it can only be studied within a limited timeframe. Until this day, exactly how immature cells progressively transform into specialized ones - a stepwise or gradual process referred to as 'differentiation' - remains only partially understood and is

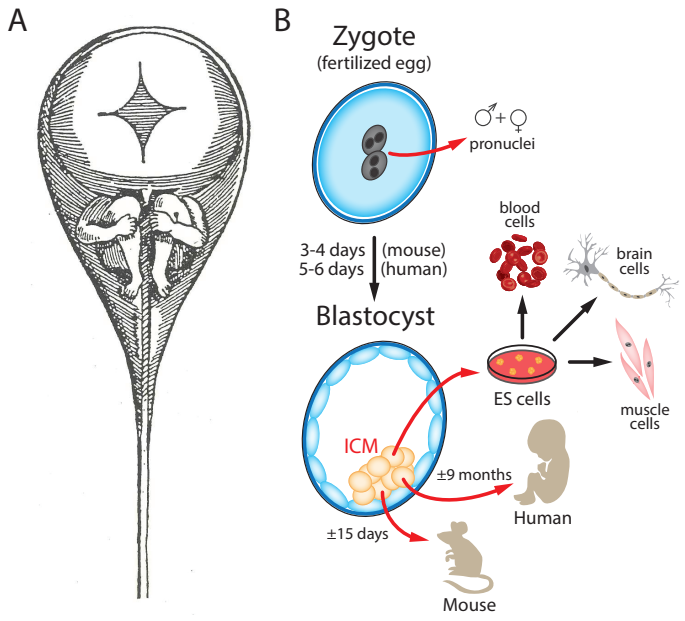


Figure 1. Mammalian development: from preformation theory to modern embryogenesis. (A) Illustration drawn by Nicolaas Hartsoeker in 1694 showing a small human (a 'homunculus') within the head of a sperm cell. **(B)** A summary of our current understanding of mammalian embryonic development. The first days after fertilization, the zygote (harbouring a male and female pronucleus) develops into a blastocyst containing the inner cell mass (ICM). The cells of the ICM will give rise to all cells and tissues of the developing animal. Embryonic stem (ES) cells can be derived from the ICM and cultured *in vitro*. Like the ICM, ES cells can give rise to all different mature cell types, a trait called 'pluripotency'.

intensively studied.

From stem cells to tissues

Shortly after fertilization, the mammalian zygote develops into a structure called the blastocyst. Within the blastocyst a small cluster of cells, the inner cell mass (ICM), arises. Through a process called gastrulation, these cells reorganize to form the three germ layers of cells laid down in a primitive body plan: the embryonic or primitive endoderm, the mesoderm and the ectoderm. Subsequently, during organogenesis, each of the three definitive layers (definitive endoderm, mesoderm and ectoderm, including its derivative neuroectoderm) and a transient cell population called the neural crest cell lineage give rise to specific sets of cell types and hence tissues. In a simplified summary, our lungs, liver and digestive tract are derived from definitive endoderm; the skeletal, muscular and circulatory systems (including blood) from mesoderm; and most of our skin and the entire nervous system from ectoderm[6]. Thus, the few cells forming the ICM are able to divide (or 'proliferate') and differentiate into all types of cells and tissues present in the adult (Figure 1B). The latter property is referred to as 'pluripotency'[6].

In 1981, Gail Martin and Martin Evans/Matthew Kaufman described a technique to isolate and culture ICM-derived cells *in vitro* (i.e. in a petri dish) from 3.5 days old (E3.5) mouse blastocysts[7,8]. These cells are referred to as embryonic stem (ES) cells, which are pluripotent (like the ICM) and can be expanded indefinitely in an undifferentiated state[9]. In 1998, James Thompson and Jeffrey Jones established the first human pluripotent ES cell lines from donated human embryos produced by *in vitro* fertilization[10]. Scientists have long since shown great interest in the conversion of differentiated cells back into pluripotent stem cells, a process called 'reprogramming' (see historical overview by Graf[11]). Using a technique called somatic cell nuclear transfer, and following landmark experiments in green frogs by Robert Briggs and Thomas King in the 50s[12], John Gurdon (in 1962, using African clawed frogs[13]) and later Ian Wilmut (in 1996, resulting in the first cloned mammal: 'Dolly' the sheep[14]) made the key discovery that somatic cell nuclei when transferred into enucleated or irradiated oocytes could sometimes result in the generation of an early embryo and even developing animal. Thus, somatic cells retain the potential to generate all three embryonic germ layers[11]. In 2006, a team led by Shinya Yamanaka demonstrated that fully differentiated cells grown in a culture dish could be directly reprogrammed into pluripotent stem cells using a defined set of factors[15]. He called these reprogrammed cells 'induced pluripotent stem' (or iPS) cells.

The importance of these landmark discoveries can hardly be overstated and is underscored by the 2007 and 2012 Nobel Prizes in Physiology or Medicine. The establishment of mouse ES cell lines paved the way for the generation of genetically modified mice, which have provided unprecedented insight into mammalian gene function and resulted in the generation of numerous mouse models of human disorders[16-18]. Because ES and iPS cells can in principle be differentiated into virtually any adult cell type and are amenable to genetic manipulation, they offer great therapeutic promise for patient-specific cell replacement/supplementation therapies as well as disease-specific drug screening[19]. As progress in this field has been and still is extremely rapid, exciting new discoveries are bound to emerge the coming years.

Stem cells: an operational definition

Stem cells are by no means restricted to embryonic development, as they also have important functions in the adult animal. In fact, the field of stem cell biology emerged from the identification of an adult stem cell population. In the early 1960s, Till and McCullough identified hematopoietic stem cells (HSCs) in the bone marrow of adult mice[20-22], a stem cell population responsible for the continuous generation of all mature blood cell types. Since then, HSCs have been intensively studied and many of the basic principles of stem cell biology have been derived from studies of the hematopoietic system[23,24]. Perhaps not surprising, HSCs were the first stem cells routinely used in clinical practice, with HSC-containing grafts being transplanted to treat various blood cell disorders and leukemias[24]. At present, adult stem cells have been identified for many tissues[25-27].

Formally, for a cell to be considered a *bona fide* stem cell, it needs to satisfy the following three criteria[6]:

1) The cell has the ability to self-renew, meaning that after cell division at least one of the daughter cells maintains a stem cell identity.
2) The cell is not itself a terminally differentiated cell and can divide indefinitely (or at least for the organism's entire lifespan).
3) The cell has the ability to differentiate into one or more differentiated cell types *in vivo*.

Adult tissue somatic stem cells, unlike ES cells, are not pluripotent. Their differentiation potential is generally restricted to generating the specific cell types of a given tissue, although controversial findings of adult stem cell differentiation across tissue-specific lineage boundaries have been reported[26,28]. Hence, adult stem cells are therefore referred to as multipotent. They play an essential role in sustaining tissue homeostasis throughout life by generating new cells to compensate for tissue or cell loss. Tissue types that display a high cell turnover rate (e.g. epidermis and intestinal epithelial cells) need their stem cells to continuously divide[27], while more static tissues (e.g. kidney[29] or liver[30]) exhibit very low stem cell activity under normal conditions. An important aspect of how stem cells or specific subsets of a larger stem cell pool maintain tissue integrity is their ability to rapidly respond to tissue damage. Slow-dividing or even dormant ('quiescent') stem cells, such as those found in the liver, respond to tissue injury by rapidly undergoing cell divisions to replace the lost cells[31].

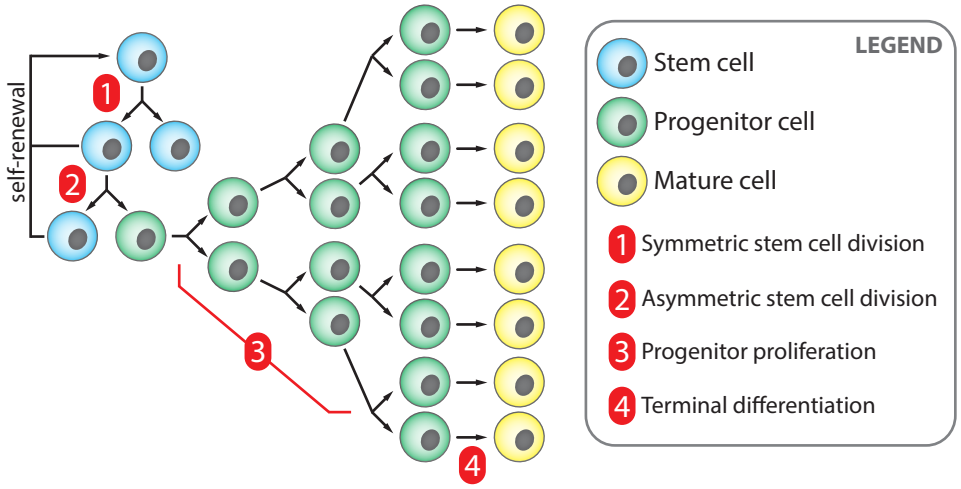


Figure 2. Basic concepts of stem cell division and differentiation. Stem cells are able to maintain their numbers through an unlimited capacity for self-renewal (1 and 2, [a]symmetric cell divisions). Progenitor cells (sometimes also referred to as 'transit amplifying cells') are responsible for the bulk of cell proliferation by going through a limited series of rapid divisions (3), after which they terminally differentiate (4).

Figure 2 depicts the basic principles through which stem cells are able to generate new progeny cells and maintain tissue function, independent as to whether they divide symmetrically or asymmetrically. Cell types within tissues can be divided into three broad categories: stem cells, progenitor cells and mature cells. The relationship between these groups is hierarchical[6]. Stem cells, usually present in low numbers only, reside at the apex of the hierarchy. Stem cells divide, yet maintain themselves as an undifferentiated population. This is achieved through their ability to self-renew, a classical feature of stem cells[32]. Without self-renewal, a stem cell population could over time become exhausted. Adult stem cell systems often rely on progenitor cells (sometimes also referred to as 'precursor cells' or 'transit amplifying cells') for the bulk of cell proliferation[6,27]. However, these progenitors are usually short-lived, since they do not possess the self-renewal ability of a true stem cell[32].

By making progenitor cells responsible for most of the proliferation, adult stem cells can generate a plentiful supply of new cells without dividing very frequently themselves (Figure 2). Extensive cell division of a long-lived stem cell brings along significant risk for the integrity of its genome, which jeopardizes stem

cell function and can ultimately result in cancer[32]. Mammalian tissue systems therefore often keep (a subset of) their stem cells in a quiescent state[33], only activating them when strictly required and/or in case of an acute need for new cells (e.g. injury[34] or infection[35]).

The research described in this thesis involves several differentiation processes within the hematopoietic system. I will therefore describe the generation of the different blood cell types from HSCs in greater detail in the next section.

Hematopoiesis

Hematopoietic development, or hematopoiesis, describes the continuous generation of all mature blood cell lineages from HSCs[23]. This extraordinary ability of the HSC is illustrated by the following experiment. When a single HSC is transplanted into an animal in which the endogenous hematopoietic system has been completely destroyed by for example irradiation or cytotoxic drugs, this HSC can reconstitute the recipient's entire blood system for the rest of its life[36]. In addition to generating such cellular diversity, HSCs need to produce millions of blood cells per second to sustain blood homeostasis in an adult human. Despite this high production demand, true HSCs (in the field called long-term repopulating HSCs) only rarely divide and usually reside in a state of low metabolic activity. The major proliferative burden within the hematopoietic system lies with the HSC's progenitor progeny[24].

During development, hematopoiesis occurs in two phases. The first wave of hematopoiesis is initiated in the yolk sac (around day E7.5 in the mouse embryo) and is referred to as 'primitive' hematopoiesis[23]. Its main function is to produce enough red blood cells to provide the necessary oxygen to sustain the rapidly growing embryo. Primitive hematopoiesis is transient, and the first definitive HSCs emerge primarily in the dorsal aorta within a region of the embryo called the aorta-gonad-mesonephros[37] (AGM; at E10.5 during mouse development – see article by Robin and colleagues for a movie of the 'birth' of an HSC[38]). Shortly after their appearance in the AGM, HSCs can also be found in the yolk sac and placenta (at E11)[23]. HSCs then migrate to the fetal liver (FL), where they undergo massive expansion (from ±10 HSCs at E11 to >1000 HSCs at E14, a phenomenon that is not well understood)[39,40]. The FL is the main HSC reservoir at E14. HSCs then complete their developmental journey by migrating to the thymus, spleen and finally the bone marrow (E17). The HSC potential within the FL is then lost, while the bulk of HSC activity will remain in the bone marrow for the organism's entire lifespan[23,41].

Definitive HSCs in the FL and (adult) bone marrow give rise to a multitude of different blood cell lineages. The different cell types generated by HSCs are depicted in Figure 3. In the broadest sense HSCs are able to generate cells of the myeloid and lymphoid lineages, which involve several increasingly committed progenitor stages[24,42]. These progenitors exhibit a high proliferative index[24] and will eventually give rise to mature, fully differentiated cells. The laboratory of Irving Weissman put forward the classical model of hematopoietic differentiation as shown in Figure 3[24]. A slightly altered scheme (involving a so-called lymphoid-primed multipotent progenitor) has been proposed by Adolfsson et al.[43] (dotted lines, Figure 3). I will briefly introduce the different types of mature hematopoietic cells and then focus on the two specific branches of the hematopoietic system that play a central role in this thesis: the development of erythroid cells and B lymphocytes.

Myeloid cells

Commitment to the myeloid lineage is initiated at the level of the common myeloid progenitor (CMP, Figure 3). CMPs in turn give rise to granulocyte-macrophage progenitors (GMP) and megakaryocyte-erythroid progenitors (MEP). Alternatively, MEPs may originate from a multipotent progenitor without passing through a CMP intermediate. GMPs differentiate into mast cells, monocytes/macrophages and granulocytes. These cell types fulfil essential roles in our cellular immune system, such as the phagocytosis of invading pathogens[6]. Dendritic cells, the body's main antigen-presenting cells, have been proposed to originate from the CMP, although they can also be derived from a lymphoid progenitor (see below)[44,45]. MEPs generate megakaryocytes, large polyploid cells that produce the platelets responsible for blood clotting, and red blood cells or 'erythrocytes'. The latter are the most common type of blood cell and the principle means of oxygen transport throughout the body[42]. Mature erythrocytes are formed through a differentiation process called erythropoiesis.

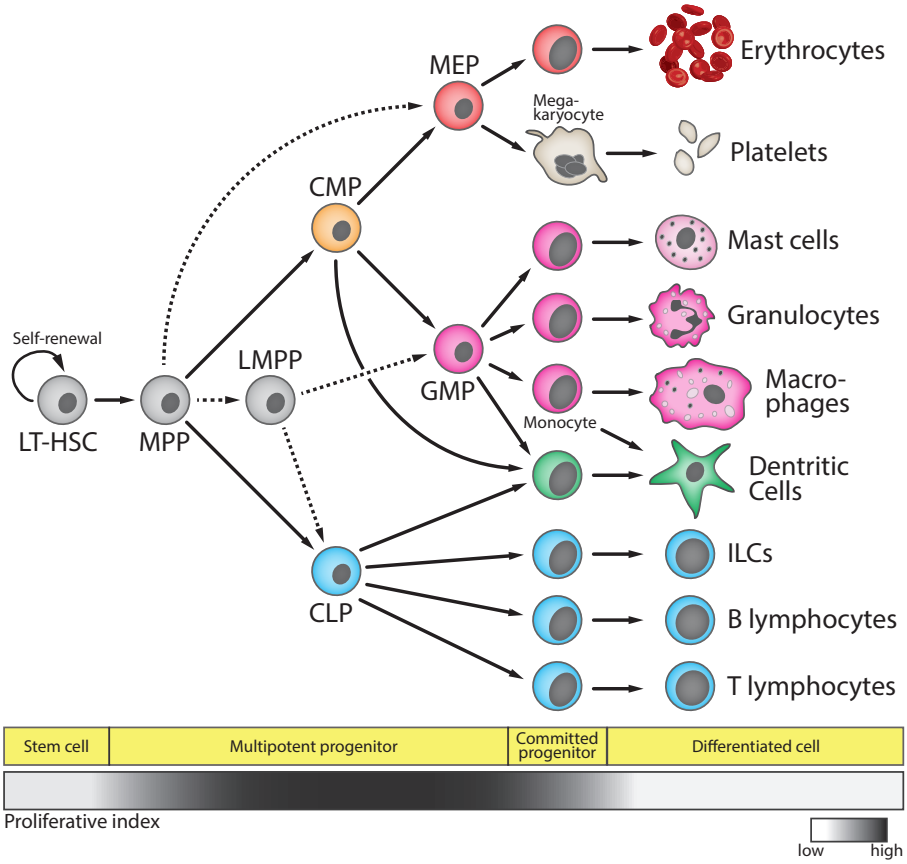


Figure 3. Model of hematopoietic development. A schematic representation of definitive hematopoiesis as it occurs in the fetal liver or (adult) bone marrow (although terminal differentiation of T lymphocytes takes place in the thymus) as proposed by the Weissman laboratory. Long-term hematopoietic stem cells (LT-HSCs) give rise to all mature blood cell lineages. Stem cell differentiation proceeds via several progenitor stages that become progressively restricted towards a specific lineage. The cellular origin of dendritic cells is complicated and still incompletely understood, but CMPs, GMPs, CLPs and monocytes have all been reported to give rise to dendritic cells. As LT-HSCs are fairly quiescent, the vast majority of cell proliferation is usually achieved by the different progenitors (as indicated by the proliferation index, which represents a general trend). The dotted arrows represent an alternative model postulated by Adolfsson et al.[43], which involves MEP generation directly from MPPs and the existence of an LMPP population that gives rise to both CLPs and GMPs. Figure was adapted from published reviews[24,42]. MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte-macrophage progenitor; MEP, megakaryocyte-erythroid progenitor; ILCs, innate lymphoid cells.

Erythropoiesis

In 1658, using an early microscope, the Dutch biologist Jan Swammerdam was the first person to describe red blood cells[46]. Several years later and unaware of Swammerdam's work, the famous Dutch pioneer of microscopy Antonie van Leeuwenhoek also provided a detailed description of red blood cells. Van Leeuwenhoek even made a first attempt at estimating their size: '25,000 times smaller than a fine grain of sand'[47].

To fully appreciate the abundance and significance of the erythroid system, one only has to look at the numbers in Table 1. This illustrates how the erythroid system is committed to satisfy the continuous demand for oxygen. Gas transport by erythrocytes is achieved through the erythroid-specific production

Table 1. Notable parameters of human erythroid cells and erythropoiesis[42,48]

Duration erythropoiesis (adult BM)	7 days
RBC Output (adult BM)	2.4 million/second
Lifespan (circulation)	3-4 months
% of total blood volume (hematocrit)	40%-50% (vs. 1% WBCs)
Quantity (steady-state)	20-30 trillion RBCs
% of total cellular content human body	~25%
Hemoglobin (Hb) content	270 million molecules/cell
% dry-weight RBC composed of Hb	96%
Erythrocyte size	6-8 μm

BM: bone marrow; WBCs: white blood cells; RBCs: red blood cells

of a molecule called hemoglobin[6]. In vertebrates, a hemoglobin molecule is a tetramer of four globular protein subunits. Each of these subunits binds a heme group, which is comprised of a charged iron atom held within a ring structure called a porphyrin ring. This iron ion is the actual site of oxygen binding and allows each hemoglobin molecule to carry four oxygen molecules. The spectral properties of hemoglobin are responsible for our blood's red color.

During development many animals, including mice and humans, produce different types of hemoglobin. Hemoglobin tetramers are predominantly composed of two 'α-like' (in humans and mice either ζ or α) and two 'β-like' globin subunits (in humans ε, γ, δ and β; in mice εy, βh1, βmajor and βminor)[6,42]. In adult humans, >98% of the hemoglobin pool consists of hemoglobin A (HbA, a tetramer of two α- and two β-globin protein subunits). However, during development the human fetus mainly produces fetal hemoglobin (or HbF) assembled from two α and two γ subunits[42]. HbF allows the fetus to more efficiently extract oxygen from maternal blood, as the affinity of HbF for oxygen is slightly higher than that of adult HbA. Around birth, production of the γ subunit stops and β subunit production is strongly increased[42]. This phenomenon is called 'hemoglobin switching' (Figure 4A) and is regulated at the gene expression level (further discussed at the end of this chapter).

In addition to producing massive amounts of hemoglobin, erythrocytes have evolved a unique morphology and physiology that is perfectly adapted to their function. Mature erythrocytes adopt an oval biconcave shape to maximize their surface area/volume ratio, allowing for rapid oxygen diffusion in and out of the cell[42]. Due to the synthesis of special membrane proteins, erythrocytes are extremely flexible and can squeeze themselves through even the tiniest capillaries of our circulatory system[49]. In mammals, erythrocytes even extrude their nucleus (a process called 'enucleation') and lose their organelles[50]. These events facilitate the extreme morphological changes erythrocytes undergo.

From the MEP, the first erythroid-restricted immature progenitors arise: the burst forming unit-erythroid (BFU-e) and subsequently the colony forming unit-erythroid (CFU-e) (Figure 4B). These give rise to proerythroblasts that further differentiate towards definitive red blood cells in a structure called the erythroblastic island[51]. This structure is composed of a central macrophage surrounded by layers of progressively differentiated erythrocytes. The central macrophages provide iron and developmental signals to the maturing red blood cells and are therefore sometimes referred to as 'nurse' cells. Within these erythroblastic islands, proerythroblasts differentiate via a series of erythroblast stages (basophilic, polychromatic and orthochromatic erythroblasts) into reticulocytes. During this process, the cells undergo a limited number of symmetric cell divisions, accumulate hemoglobin, decrease in cell size and finally enucleation (Figure 4B)[42,49,51]. The ejected nuclei are engulfed by macrophages[50,51] and the reticulocytes are released in the bloodstream where they mature into erythrocytes.

Lymphoid cells

Commitment to the lymphoid lineage is initiated at the level of the common lymphoid progenitor (CLP, Figure 3). These cells generate dendritic cells (which can also arise from the GMP[45]) and lymphocytes. There are three types of lymphocytes: innate lymphoid cells[52] (or ILCs; a multifunctional group of innate immune cells), T lymphocytes (from *thymus*, where they differentiate) and B lymphocytes (from *bone marrow*)[24]. Very recently, the ILC branch was expanded with the discovery of several new ILC subsets[52]. All ILCs were shown to originate from the CLP[53,54] and they fulfil a surprisingly diverse set of functions, ranging from immunity to viruses and tumour surveillance by natural killer (NK) cells (the prototypic ILC) to lymphoid organogenesis during embryonic development by lymphoid tissue-inducer (LTi) cells[55]. The T and B lymphocytes, often simply referred to as 'T' and 'B' cells, represent the effector cells of the adaptive immune system and are

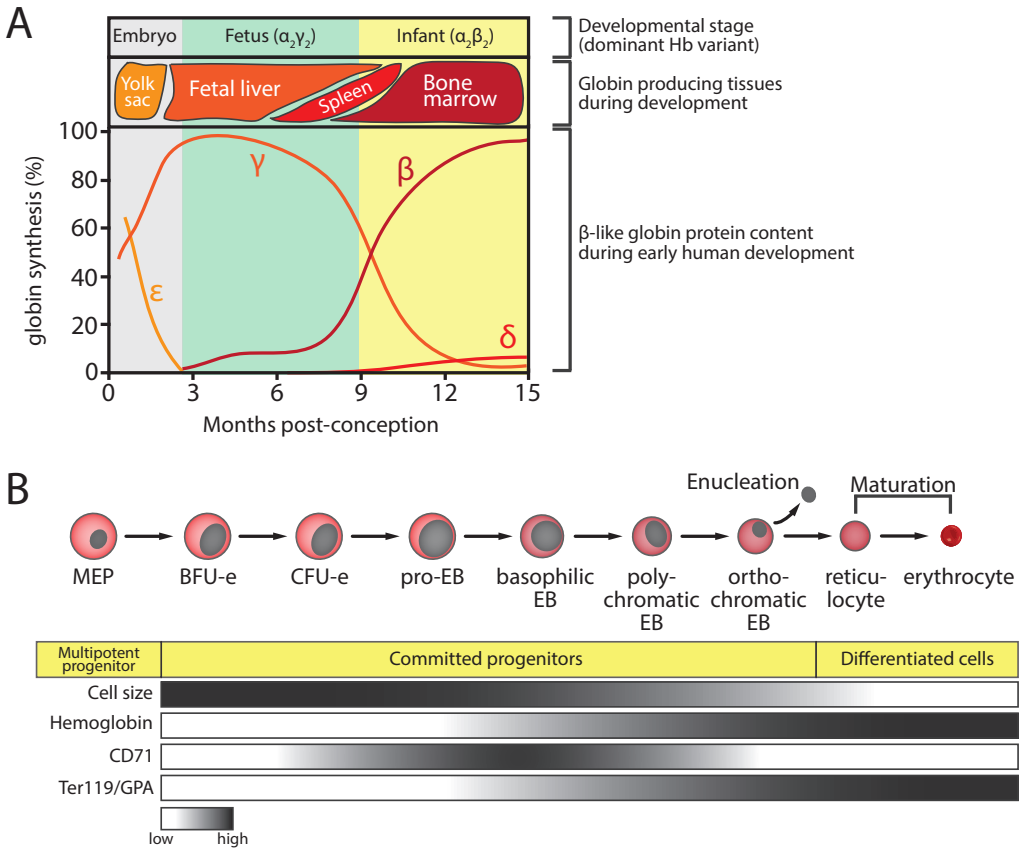


Figure 4. Erythropoiesis and β-globin switching. (**A**) Expression levels and location of the human β-like globins during early human development. Note that two switches take place: first the ε-to-γ switch during embryonic development, which is followed by the γ-to-β/δ switch around birth. Although $\alpha_2\beta_2$ hemoglobin (HbA) will remain the dominant species throughout adult life (>98% of total Hb synthesis on average), low levels of $\alpha_2\gamma_2$ hemoglobin (HbF) continue to be synthesized. Adapted from ref[193]. (**B**) Schematic representation of erythroid development from the megakaryocyte-erythroid progenitor (MEP) via several committed progenitor stages to an enucleated erythrocyte. CFU-e and pro-EB stage cells are the most sensitive to EPO signalling. Several characteristic markers (i.e. expression of CD71, Ter119 and Glycophorin A [GPA]) and cellular attributes (i.e. hemoglobin synthesis and cell size) that accompany differentiation are depicted as gradients. BFU-e, burst-forming unit erythroid; CFU-e, colony-forming unit erythroid; EB, erythroblast. Adapted from ref[42].

absolutely essential for pathogen elimination[56]. The difference with cells from the innate immune system (e.g. neutrophils, macrophages, ILCs) lies in the way the two systems recognize foreign invaders. Innate immune cells are equipped with germline-encoded receptors ('pattern recognition receptor', or PRR) that can identify a wide range of common pathogen constituents[56]. B and T cells use an enormous repertoire of specialized receptors, referred to as 'B cell receptors' (BCR) or 'T cell receptors' (TCR), to gradually develop a highly specific and effective immune response against any (foreign) substance (called 'antigens')[56]. T cells rely on innate immune cells, in particular dendritic cells, to present these antigens to them. Through such antigen presentation, T cells expressing a compatible antigen receptor on their cell surface will be selected to participate in mounting a cellular immune response against the antigen[56]. This involves the production of signal molecules (cytokines) to communicate with other immune cells or cytotoxins to eliminate infected/dysfunctional cells. B cells can also recognize free antigens in circulation. Once a B cell encounters its

cognate antigen and receives the appropriate co-stimulation (often from a mature T cell) it can differentiate into a plasma cell that is able to produce (pathogen-neutralizing) antibodies, a secreted form of the BCR[56]. Following pathogen elimination, specialized memory lymphocytes are able to persist for years to ensure an even faster and stronger response when challenged with the same antigen again. Such 'immunological memory' is the reason vaccines are so effective at providing long-term protection to a virus or bacterium, or why some pathogens often only cause (severe) disease symptoms once[56].

B cell development

Lymphocytes are not nearly as abundant as red blood cells: about 200.000 B cells can be found in one ml of human blood (compared to 5 billion erythrocytes), representing 10-25% of the total circulating lymphocyte population[57]. During development, early B cell precursors first arise in the FL and are later produced in the bone marrow[58]. Instrumental for our understanding of B cell function and antibody production within the immune system has been the 'clonal selection theory' first postulated by Burnet and Talmage in 1959 (and later proven to be correct)[56,59]. The essence of their theory stated that every individual B cell produces a unique antibody expressed on its surface as a receptor, which allows for the selection of B cells through antigen binding and the subsequent production of secreted antibodies against the antigen. BCR diversity is enormous: around 10^{11} different BCRs are estimated to occur within the B cell population at any given time, providing our immune system with a virtually unlimited capacity to detect foreign invaders[56].

It has become apparent that the differentiation of B cells is intimately connected to the genetic events responsible for generating BCR diversity (see below). B cell development represents a complicated developmental system, giving rise to several different types of mature effector cells[56]. Research described in this thesis is focused on early B cell development in the bone marrow, and therefore I will focus only on these aspects of B cell differentiation.

Bone marrow CLPs give rise to progenitor B cells, also referred to as pre-pro B cells. In mice, these cells start to express the B220 isoform of the CD45 protein, which will remain expressed on the surface of all B cells during differentiation[60]. Pre-pro-B cells will further differentiate into pro-B cells, concomitant with the production of the key B cell identity protein Pax5[61]. Pro-B cells go through several pre-B cell intermediates before reaching the immature B cell stage (see Figure 5A)[60]. At that time, they express a functional BCR on their surface and will migrate to the periphery (the blood circulation and lymphatic system) to further mature. During these early steps of B cell development, ongoing differentiation depends on the successful stepwise assembly of the BCR[60,62]. The BCR is a Y-shaped protein composed of two identical subunits that each consists of an immunoglobulin (Ig) heavy (IgH) and light (IgL; in mice and humans 2 subtypes exist, κ and λ) chain (Figure 5B)[56]. The IgH and IgL chain subunits are produced from the *Igh* and *Igκ*/*Igλ* loci respectively. These loci span very large genomic distances and contain various kinds of gene segments: variable (V), diversity (D, only found in the *Igh* locus) and joining (J) segments. Each type of gene segment is present multiple times in the genome, with V segment diversity being the largest (both the human and mouse *Igh* loci contain >100 V genes)[63]. The heavy and light chain BCR subunits are assembled by combining V, D (only for the heavy chain) and J segments through a process called V(D)J recombination (Figure 5B)[56,62]. This recombination process occurs to a large extent in a random fashion and is the most important determinant of antibody diversity. Essential for this process are the recombination activating genes (*Rag* genes, encoding the Rag1 and Rag2 proteins), which are essential for V(D)J recombination and therefore also for B cell development (see below).

V(D)J recombination is a tightly regulated process, which follows a precise order of events[56,63]. Very early in B-cell development, at the progenitor B and early pro-B cell stages, *Igh* D and J segments are first rearranged (D_H to J_H), after which the resulting D_HJ_H segment is coupled to an *Igh* V gene (V_H to D_HJ_H) in pro-B cells. This is also where the first developmental checkpoint is introduced: a pro-B cell will only be allowed to move on to the pre-B cell stage upon successful *Igh* rearrangement and production of the IgH protein (referred to as 'μ'). Pro-B cells have an ingenious way of checking for successful *Igh* rearrangement: they use the μ protein, together with an invariant surrogate light chain (composed of the λ5 and VpreB proteins), to build a pre-BCR. This pre-BCR is expressed on the cell surface and, like the mature BCR, has the capacity to send intracellular signals through its association with the Igα and Igβ proteins[64,65]. Further downstream, pre-BCR signals are relayed through several kinases and adapter proteins, ultimately allowing differentiation into a pre-B cell[65]. After a short phase of proliferation, the *Igκ* or *Igλ* light chain loci will undergo V_L to J_L

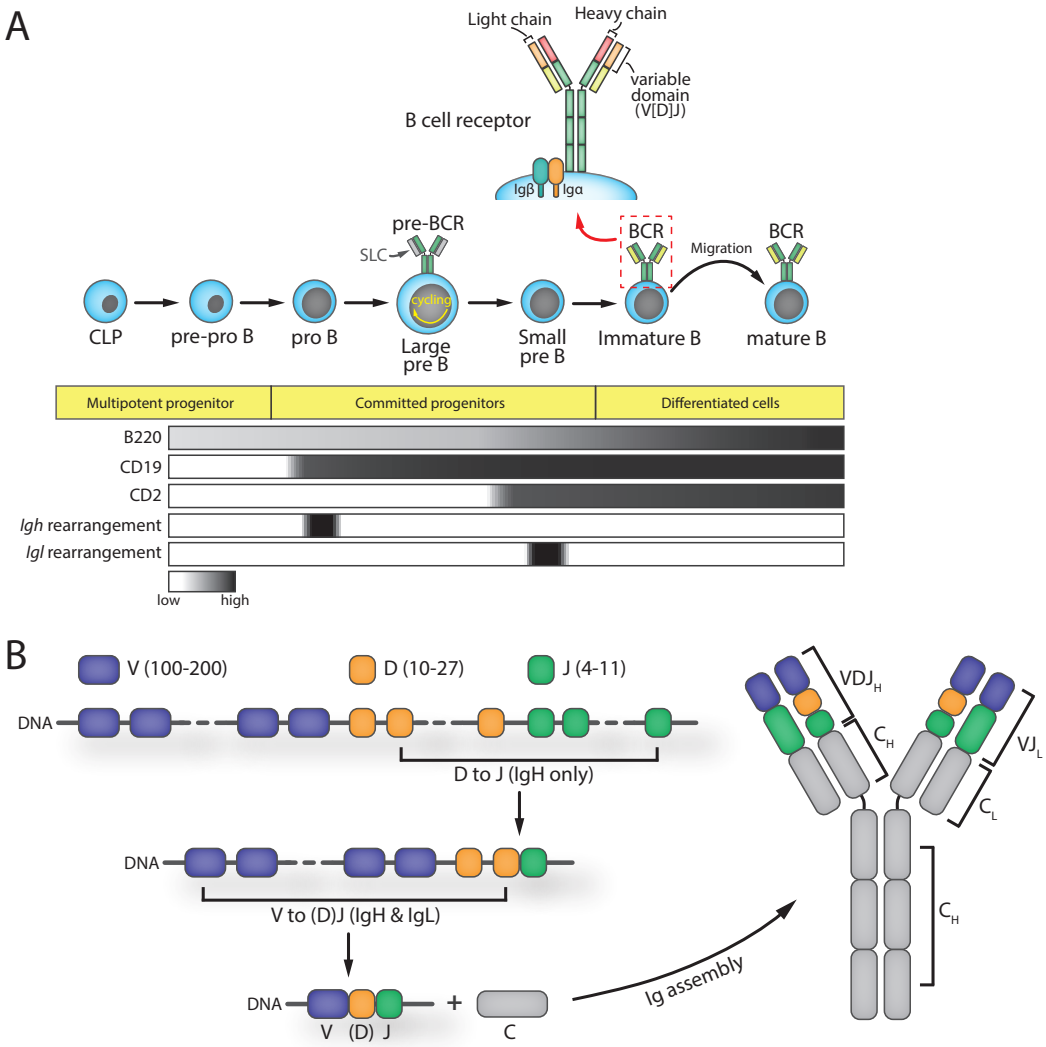


Figure 5. Early B cell development. (**A**) Schematic representation of early B cell development from the common lymphoid progenitor (CLP) via several committed progenitors to a mature B cell. Activation of the B cell transcriptional program starts at the pre-pro B stage; actual commitment to the B cell lineage occurs at the pro-B cell stage. Immunoglobulin (Ig) rearrangements take place at very specific stages of differentiation (*Igh*: pro B; *Igl*: small pre B) and result in the assembly of first the pre-BCR (using a SLC) and later the (mature) BCR. A zoom-in picture of a surface-bound BCR complex depicts the general composition of the BCR associated with the Igα and Igβ signal transducers. Note the transient increase in proliferation at the large pre B cell stage after successful *Igh* rearrangement. The expression of common surface markers (i.e. B220, CD19 and CD2) and the developmental window of Ig rearrangements are depicted as gradients. BCR, B cell receptor; SLC, surrogate light chain. (**B**) The process of V(D)J recombination and subsequent Ig assembly explained. Starting from an *Igh* locus in germline configuration (top DNA strand), a D segment is joined to a J segment (D-to-J). Any intervening sequences will be removed in the process. Next, a V segment is recombined with the DJ (*Igh*) segment (V-to-DJ). *Igl* rearrangements start with a V-to-J joining as they lack D segments. The new V(D)J gene is transcribed and spliced to an Ig Constant (C) segment. Rearranged heavy and light chains are then assembled into a functional Ig molecule, which can be expressed as a receptor at the cell surface (the BCR) or excreted as an antibody. Antigen recognition is achieved by the combined variable parts of the heavy and light chains (the non-gray V[D]J regions).

rearrangement. Here the second checkpoint is set: only pre-B cells that produce a functional IgL protein will be selected for further maturation. The checking mechanisms are very similar to those of the first checkpoint. The resulting IgL will be paired with the already present IgH protein; they will be assembled into the BCR and expressed at the cell surface. Signals from the BCR and its associated proteins will ensure the cell's survival, and after the BCR is checked for auto-reactivity, the immature B cells leave the bone marrow to further mature in secondary lymphoid organs such as the spleen[56]. T cell lineage specification and commitment in the thymus proceeds in an analogous fashion to the early B cell pathway described above, including similar V(D)J recombination processes that act as developmental checkpoints and result in TCR assembly[56].

DNA, chromatin and gene expression

The above part of this *Introduction* illustrates the importance of developmental processes for the proper growth of an embryo and for maintaining tissue homeostasis during adult life. But how does a(n) (embryonic) stem cell give rise to all those mature cell types? How do progenitor cells enter a proliferative state, to later exit it and differentiate into mature cells? Even more fundamentally, one could ask: what is the underlying mechanism that makes a neuron and a B cell so different? An important part of the answer to these questions resides in the nucleus of a cell, where the genetic material is stored as DNA.

Deoxyribonucleic acid or 'DNA': structure and content

DNA can be viewed as a simple code composed of four 'letters' represented by the four types of nucleotides (also referred to as bases: adenine [A], guanine [G], cytosine [C] and thymine [T]) it is built from[6]. Due to selective pairing of the nucleotides (A pairs only with T and G only with C, a phenomenon called basepairing), DNA usually exists as a double-stranded molecule composed of two anti-parallel complementary strands. The fact that DNA exists as two complementary strands provides a straightforward way of copying it: a single DNA strand can serve as a template for the synthesis of the complementary strand. The entire human DNA code is approximately 3 billion nucleotides in length, which is referred to as the human genome[66]. The genome contains all the hereditary information required for the development and function of an organism. The Swiss chemist Friedrich Miescher discovered DNA in 1869[67], and its 3D-structure, the characteristic double helix, was resolved in 1953 by the famous duo James Watson and Francis Crick (with help from Rosalind Franklin and Maurice Wilkins)[68]. DNA is further organised in structural units called chromosomes (Figure 6A). Humans have 46 chromosomes: two copies (1 from the father, 1 from the mother) of 22 autosomes and 2 sex chromosomes, XX in females and XY in males[6].

Using electron microscopy, scientists have discovered that the chromosomal structure of the genome shows several intricate layers of packaging[69]. At the molecular level, our genome appeared to be organized into structures called the 30nm and 10nm fibers (Figure 6A). The latter, when visualized using an electron microscope, resembles 'beads on a string'. The beads seen along the 10nm fiber are actually the basic structural units of DNA packaging: the nucleosomes (Figure 6A)[70]. Nucleosomal packaging of DNA is a common feature shared by all multicellular organisms. A nucleosome consists of 146 base pairs (bp) of DNA wrapped around an octamer of histone proteins. The histone octamer is comprised of two copies of the H2A, H2B, H3 and H4 histone proteins[70]. The combination of DNA and packaging histones is referred to as chromatin' DNA packaging into chromatin is an amazing compaction process, which is still poorly understood. Without packaging, the naked, uncoiled DNA that needs to be stored in a nucleus (with an average diameter of 6µm) would measure up to about 2 meters[6]. Besides compacting the genome, chromatin has several other functions. Important among these are the impact of chromatin structure and folding on the interpretation of genome-encoded information, which I will further highlight in the next sections of this chapter.

Every cell containing a nucleus carries an essentially identical copy of the genome. The most important information encoded by our genome comes in the form of genes. Genes present themselves as defined regions of the genome that encode a functional molecule[6]. The most important and best characterized of these gene-encoded molecules are proteins. Humans have approximately 22.000 protein-coding genes scattered across their genome[66], of which the majority is highly conserved throughout evolution. Although this number might sound impressive, it is actually not: a 'simple' roundworm has a

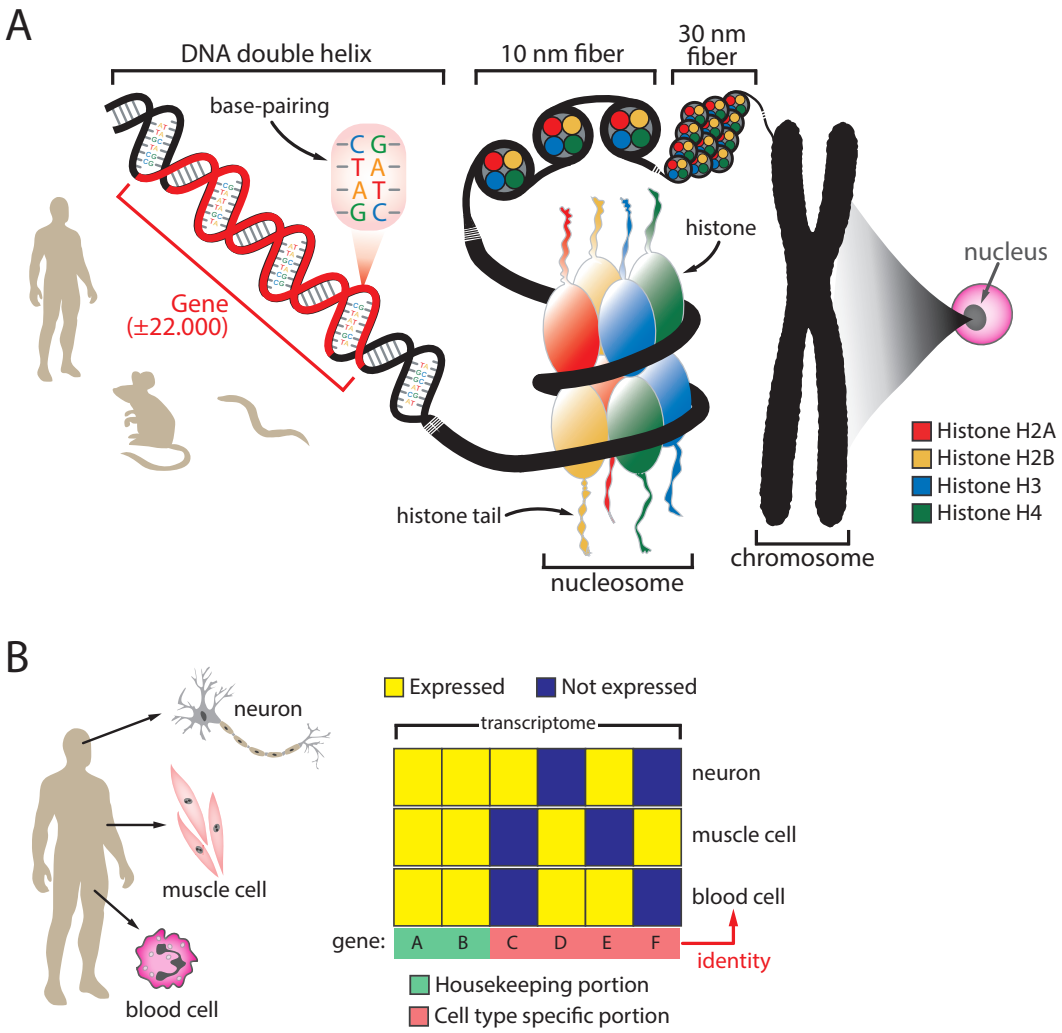


Figure 6. DNA, chromatin and genes. (**A**) Our genetic material is stored in the form of DNA and resides within the nucleus of a cell. Defined regions of the genome called 'genes' encode for functional proteins or non-coding RNA molecules. Humans carry approximately 22.000 genes, and similar gene numbers have been detected in other frequently studied eukaryotes such as the mouse or roundworm. Due to selective basepairing (A:T and C:G), DNA molecules adopt an antiparallel double helix structure. In the nucleus DNA is tightly associated with histone proteins, forming a protein-DNA complex referred to as 'chromatin'. The basic unit of chromatin is the nucleosome, which consists of 147 base pairs of DNA wrapped around a histone octamer (consisting of 2 copies of the 4 core histones). Nucleosomes are further packaged in 10 and 30 nm fiber structures, resulting in a remarkable compaction of the eukaryotic genome. Collectively, these chromatin fibers are organized in large structures called chromosomes. Image was adapted from the PhD thesis of dr. D. Noordermeer (Erasmus University Rotterdam, 2009) with permission from the author. (**B**) The different cell types in our body display diverse functions and morphologies (shown are muscle cells, blood cells and a neuron with myelinating cells around the axon). Responsible for conferring cellular identity is the unique combination of genes expressed in a given cell (its 'transcriptome'). Some genes are ubiquitously expressed ('housekeeping portion'), while others show more restricted activity ('cell type specific portion'). Combined, they compose a specific gene expression signature that provides the molecules (e.g. proteins) required for cellular identity and function.

similar number of genes[71].

The synthesis of protein from a gene's DNA sequence proceeds via a nucleic acid intermediate called RNA[6]. From the DNA, RNA copies are produced via a process called transcription. These RNA transcripts are then processed into messenger RNA (mRNA), which involves the removal of non-coding parts of the genes (introns) via a mechanism called splicing. As a result, only the coding parts or exons are present in the mRNA, which is then transported out of the nucleus. By including or excluding specific exons through a process called alternative splicing, individual genes are often able to produce multiple kinds of mRNAs - thereby greatly increasing gene product diversity. In the cytoplasm, mRNAs are translated into proteins by the ribosome: every three bases of mRNA (a 'codon') encode an amino acid, which are chained together to form a specific protein. Proteins are the workhorse molecules of the cell, performing a vast array of functions ranging from regulating cell shape to actually performing gene transcription. Our genome also contains many genes that encode an RNA molecule that is not translated into a protein ('non-coding RNA genes', such as the ribosomal and transfer RNAs involved in translation[6]). Several new classes of non-coding RNAs have been identified the past decade, and they appear to fulfil important functions within the cell[72]. Currently about as many human non-coding RNA genes have been discovered as there are coding genes, and it is anticipated that many more will follow[72,73].

Reading the DNA code: gene expression

When a gene is actively transcribed it is referred to as 'expressed'. Not all genes are transcribed in a cell at a given time; only a specific fraction is expressed. This key observation explains the fundamental difference between two cell types: they express a different set of genes and therefore produce a different set of proteins and functional RNA molecules (Figure 6B). Thus, cellular identity is a direct consequence of differences in the complete set of expressed genes (called the 'transcriptome' or 'transcriptional program')[6]. Studies have indicated that about 8000 (protein-coding) genes are expressed in every type of cell[74]. These genes, the so-called housekeeping genes, are required to fulfil general functions necessary for any type of cell (e.g. forming a membrane). However, another several thousand genes show a much more restricted pattern of expression and a small subset are only expressed in one specific cell type[74]. Together, these genes form a 'molecular signature' that defines cellular identity and behaviour (Figure 6B). An example that illustrates this concept is the expression pattern of the previously mentioned globin genes. They are uniquely expressed in late erythroblasts, granting specifically these cells the ability to produce Hb for the transport of O_2. Also more general processes, such as proliferation, are ultimately a consequence of gene expression: cells can adjust their transcriptional program to alter the production levels of proteins involved in initiating or terminating cell division.

The major impact of transcriptome differences on how cells function and behave immediately implies that gene expression needs to be appropriately regulated. This is indeed the case: gene expression levels are constantly adjusted and controlled to allow cells to adapt to a changing environment, to enter a state of differentiation/proliferation or to simply ensure that a cells maintains its identity[6]. Important to note is that both the activation and the repression of gene expression are critical ways of altering a cell's transcriptome. Stochastic transcriptional output may also have a role in creating (subtle) differences in gene expression, which can influence cellular decision-making[75].

In general, cells rely on environmental cues to instruct them which changes in gene expression to implement and when to do so. This phenomenon is referred to as 'signalling' and is achieved through specialized signalling molecules (such as hormones, cytokines and growth factors) or through cell-to-cell contact[6]. These signals bind receptors at the surface of their target cells and trigger a complex intracellular cascade of events that eventually leads to changes in gene expression. This process is called signal transduction. A well-studied example of such a signalling molecule is erythropoietin or EPO. This cytokine is produced in the kidney and is absolutely essential for erythropoiesis. EPO is released in the bloodstream to reach the bone marrow, where it can bind to the EPO receptors present on the surface of erythroid precursors. The intracellular signalling cascade triggered by EPO binding to its receptor leads to specific changes in gene expression that ultimately result in increased proliferation, survival and subsequent differentiation of erythroid progenitors[76]. Under normal conditions, small amounts of EPO are sufficient to maintain red blood cell homeostasis. However, when adequate O_2 supply throughout the body is compromised, for example by excessive blood loss or staying at high altitude regions, EPO production is immediately increased (up to

1000 fold) to stimulate red blood cell production[76]. This example of cellular communication illustrates the key role of signalling in influencing a particular cell's gene expression program when required.

Within a cell, the output of a specific gene (i.e. the amount of RNA or protein produced) can be regulated at several different levels[6]:

- Transcription (e.g. adjusting the rate of transcription initiation or elongation)
- RNA transcript processing (e.g. alternative splicing)
- Post-transcriptional (e.g. degradation of mRNA molecules)
- Translation (e.g. preventing or interfering with translation into protein)

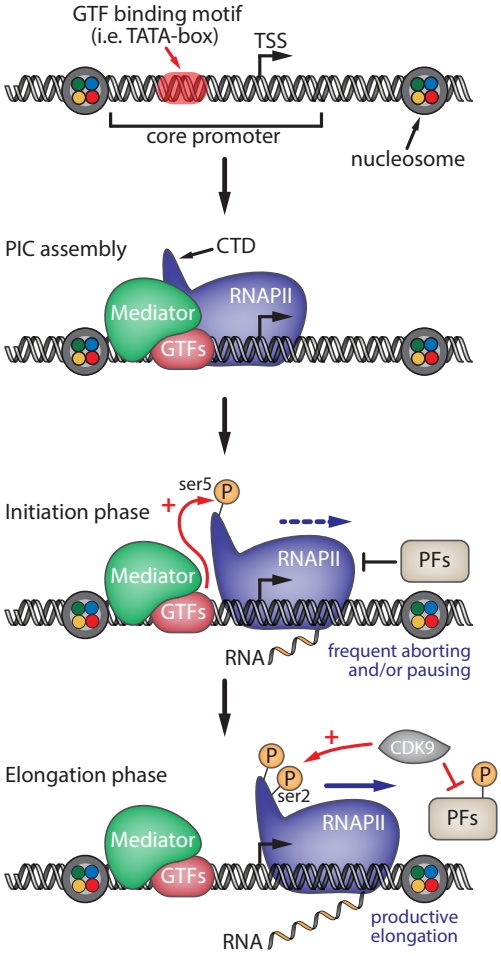


Figure 7. Eukaryotic transcription by RNA Polymerase II. Binding of the basal transcription machinery to initiate gene transcription occurs at the core promoter region, in the direct vicinity of the transcription start site (TSS). General transcription factors (GTFs) recognize and bind core promoter elements, after which they recruit RNA polymerase II (RNAPII) and the Mediator co-activator complex to form the pre-initiation complex (PIC). Phosphorylation of the RNAPII carboxyl terminal domain (CTD) at serine 5 (ser5) by the GTFs allows RNAPII to escape the promoter and initiate transcription. Shortly after initiation, RNAPII is frequently paused due to the actions of pausing factors (PFs). The transition from initiation to productive elongation involves the recruitment of the CDK9 kinase. CDK9 phosphorylates serine 2 (ser2) of the RNAPII CTD, as well as several PFs, resulting in pause release and/or progression into the elongation phase to complete RNA transcript synthesis. See text for more details.

Some of the most important and intensively studied gene regulatory mechanisms take place at the level of transcription itself. The different ways employed by cells to regulate the transcriptional output of their genome play a central role in this thesis and will be further discussed below.

Managing transcription: chromatin, transcription factors and gene regulatory elements

Before I address how transcription is regulated and controlled, I will first review the process of transcription itself.

Transcription: the essentials

Transcription is a complex process. Much of our current knowledge of the mechanisms of transcription is derived from studies attempting to purify and crystallize the responsible proteins. Leading among these efforts where those of Roger Kornberg[77], whose laboratory resolved the molecular basis of several key steps of transcription (yielding him the 2006 Nobel Prize in Chemistry). Responsible for transcribing the DNA template into an RNA copy is a class of enzymes called the DNA-dependent RNA polymerases (RNA polymerases in brief). In eukaryotes, the type II RNA polymerase (also known as RNA polymerase II, RNAPII or polII) catalyses the synthesis of primary RNA transcripts from all protein-coding genes and many non-coding RNA genes[78]. Kornberg's studies have been instrumental in our understanding of RNAPII structure. In fact, the RNAPII enzyme consists of many proteins interacting together: the RNAPII complex purified from yeast, mice and humans consists of 12 protein subunits[78].

Gene transcription starts at a region called the promoter. Promoters are located at the beginning of the transcribed genic DNA region and therefore encompass the transcription start site (TSS) of the gene. Promoter regions are usually short in length (<1kb) and contain key sequence elements required for the initiation of transcription by RNAPII[79]. Specialized proteins, called transcription factors (TFs), are able to recognize these sequence elements (or 'motifs') and bind to them (Figure 7). For transcription to initiate a specialized group of TFs, the 'general' TFs (or GTFs), need to be recruited to sequence elements near the TSS in a region of the promoter called the core promoter[79,80]. After their assembly on the DNA, the GTFs will recruit RNAPII to the gene and form the pre-initiation complex or PIC (Figure 7)[78,81]. PIC assembly is facilitated by another multiprotein complex called Mediator, which in doing so also resides at gene promoters and is therefore sometimes viewed as part of the PIC[82]. Despite the importance of Mediator for general RNAPII-mediated transcription, its role as a bona fide GTF is still under debate[83].

PIC subunits are able to melt the double-stranded promoter DNA to provide RNAPII with access to a single-stranded template. RNAPII will then make several attempts at initiating transcription, which will be aborted after only a few nucleotides ('abortive initiation')[84]. Critical for RNAPII to escape the promoter are the actions of the multifunctional GTF TFIIH[80]. An important aspect of TFIIH function is its ability to add a phosphate group to a specific amino acid residue (the serine at position 5; ser5) of the RNAPII carboxyl terminal domain (CTD), which consists of several repeats of 7 amino acids and resembles a tail-like structure[85,86]. This modification is thought to be functionally important for uncoupling RNAPII from the promoter-bound GTFs and for attracting proteins required for the subsequent part of transcription (Figure 7)[85,87]. After RNAPII escapes the promoter, most GTFs and Mediator are released and the transcription process enters the elongation phase. Important to note is that pausing of RNAPII during this early elongation phase (this predominantly occurs just after promoter escape, but has also been observed during later stages of elongation) is a widespread phenomenon. RNAPII pausing has only recently been recognized to occur at such a large scale[88]. Pausing is a consequence of the action of proteins called pausing factors, and current data indicate that RNAPII pausing provides an important manner of regulating transcriptional output (see below).

The initiation-to-elongation switch also involves a transition in RNAPII CTD phosphorylation: mainly through the actions of the CDK9 protein kinase (although CDK12 also appears to participate in this process), the serine residue at position 2 (ser2) of the CTD is also phosphorylated[85,89]. Phosphorylated Ser2 (Ser2P) becomes increasingly abundant during the course of elongation, while Ser5 phosphorylation (Ser5P) is progressively diminished. Ser2P is required to recruit proteins essential for the final stages of transcription to the elongating polymerase (e.g. the splicing machinery)[89]. Importantly, CDK9 also phosphorylates certain pausing factors, resulting in the release of paused RNAPII complexes (Figure 7)[88]. During the elongation phase, RNAPII will transcribe the full-length RNA molecule at an average speed of 3-4 kb/min[90]. Near the end of the gene, RNAPII will eventually transcribe through a specific adenine-rich sequence (the 'poly(A)' site). This sequence is recognized by specialized termination proteins that will pause RNAPII and cleave the RNA transcript to release it as part of the 3'-end mRNA maturation process, effectively ending the transcription cycle[91].

The above description summarizes the essential steps of the transcription process. For a more detailed description I refer to several excellent reviews[78,80,88,89], one of which even provides a lively animated molecular movie of RNAPII transcription[84].

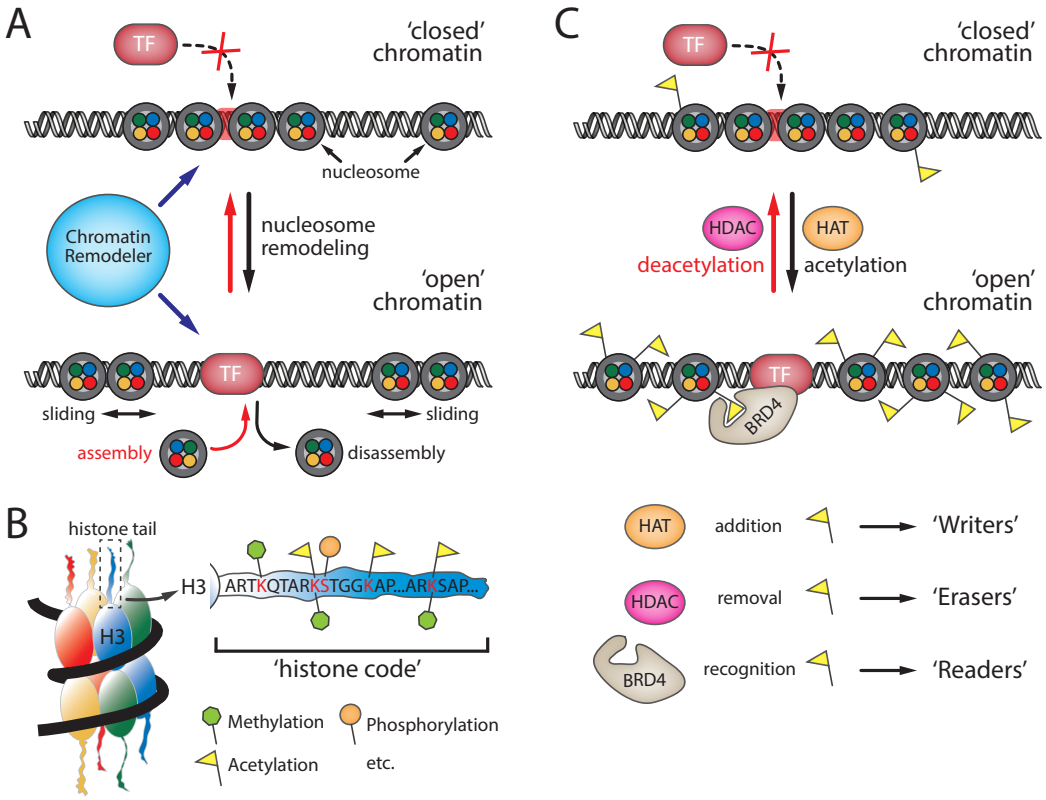


Figure 8. Chromatin remodelling and histone modifications. (**A**) Chromatin remodelling complexes can shape the nucleosome landscape through repositioning (or 'sliding') and (dis)assembly of nucleosomes. As a consequence, they can establish regions of open or closed chromatin, affecting local genome accessibility to transcription factor (TF) binding. (**B**) Histone tails protruding from the nucleosome are subjected to extensive and diverse post-translational modification at specific amino acid residues (exemplified here by the histone H3 tail). Collectively, these histone modifications are known as the 'histone code'. (**C**) An example of the functional consequences of local changes in histone modification patterns. Low levels of histone acetylation (i.e. through the actions of histone deacetylases [HDACs]) result in a closed chromatin structure inaccessible to TF binding. Histone acetylation by histone acetyltransferases (HATs) creates an open chromatin domain that facilitates TF binding. Histone modifications can also be themselves recognized and bound by regulatory proteins (as shown for the acetyl-binding BRD4 protein). The maintenance, modulation and interpretation of the histone code are realised by three categories of proteins: histone writers (e.g. HATs), erasers (e.g. HDACs) and readers (e.g. BRD4).

Chromatin: more than a simple 'wrapper'

As mentioned above, DNA resides in the nucleus within a protein-DNA complex called chromatin. It is important to realize that chromatin is not just required for DNA packaging and condensation; it also provides a way to control how that DNA is used. Chromatin has a dynamic structure, which exerts a significant impact on virtually all DNA-related cellular processes[92]. For the purpose of this thesis, I will specifically focus on the role of chromatin in modulating transcription.

In general terms, chromatin can influence transcription in a direct or indirect manner[93]. The direct mechanism is based on genome accessibility, which is dictated by chromatin organisation. When DNA is tightly wrapped around the histones and nucleosome density is very high, the DNA strands become difficult to access for TFs and the RNAPII machinery. Chromatin can therefore act as a barrier that needs to be actively overcome for transcriptional activation to occur[93]. Conversely when DNA wrapping around the histones is less tight and nucleosome density is low, TF and RNAPII binding to their target DNA sequences is

greatly facilitated. The latter, accessible form of chromatin is referred to as 'open chromatin' or 'euchromatin', while the inaccessible variant is called 'closed chromatin' or 'heterochromatin' (Figure 8)[6]. The inability of proteins to access the DNA in regions of closed chromatin can also be exploited experimentally. The classical assay for determining chromatin accessibility uses nucleases such as DNAseI, which is able to randomly cut DNA strands. When applied on a sample of chromatin, DNAseI will preferentially cut accessible sites of low nucleosome density (called DNAseI hypersensitive sites). By measuring the cutting-frequency across the genome, researchers can construct a map of open chromatin regions across the genome. Cells can actively remodel the accessibility of their genome locally to either prevent or promote gene activation[94]. Two classes of specialized proteins are involved in adapting chromatin structure to a transcription permissive or non-permissive environment: nucleosome remodelling and histone modifying proteins.

Nucleosome (or chromatin) remodellers are multiprotein complexes that use a chemical reaction called ATP hydrolysis to slide or disassemble histone octamers[95]. Through their actions, promoters of repressed genes can be actively remodelled from a nucleosome-dense and inaccessible to a nucleosome-depleted, highly accessible region and vice versa (Figure 8A). Chromatin remodelling is of crucial importance to many DNA-related cellular activities, not only for transcriptional regulation[95]. For example, remodelling is also required for the DNA repair and replication machineries to gain access to the DNA template. Several different classes of chromatin remodelling complexes exist, often performing specialized functions. More information on the types of chromatin remodelling complexes, their functions and modes of action can be obtained from several informative reviews[95,96].

Our genome encodes over 150 proteins involved in the modification of histones[97], which are currently intensively studied by numerous laboratories around the world. Histone modifying proteins target a specific part of the histone proteins referred to as the 'histone tail'. These histone tails are amino acid stretches that protrude from the nucleosome and can contact adjacent nucleosomes (Figure 8B). Similar to the RNAPII CTD, histone tails are subjected to post-translational modifications (PTMs, occurring *after* the protein has been translated from mRNA). The abundance and diversity among the different histone modifications is staggering: histones can for example be acetylated, methylated, phosphorylated and ubiquitinated (Figure 8B)[98]. To further complicate matters, histone octamers can also be modified through the incorporation of variant histone proteins (e.g. replacement of H2A with the H2AZ or H2A.X variants). These histone variants can be (subtly) different from canonical histones in their amino acid composition, structure and residues available for PTMs[99]. In 2000, David Allis postulated the existence of a 'histone code', referring to the complex patterns of histone tail modifications and histone variants that appeared to correlate with very specific features of the underlying chromatin[100]. For example, the addition of three methyl groups (Me3) to the lysine residue on position 4 (K4) of histone 3, referred to as 'H3K4Me3', occurs specifically at promoter regions[101]. This specificity makes the mapping of histone modifications and/or variants an excellent tool for predicting the function of certain genomic regions. In the context of chromatin accessibility, the acetylation and deacetylation of lysine residues on the H3 and H4 tails in particular provides an important regulatory mechanism. Histone acetylation is able to weaken histone-DNA interactions by neutralizing the positive charge of the lysine residue that attracts the negatively charged DNA strand[98]. Therefore, adjusting histone acetylation levels provides another mechanism, complementary to nucleosome remodelling, to regulate chromatin accessibility and consequently TF binding (Figure 8C). When considering the histone code as a language, enzymes that deposit certain modifications (such as the histone acetyltransferases or HATs that acetylate histone tails) are often called 'histone writers'; those that remove the marks (for example the histone deacetylases or HDACs that remove the acetyl mark) are referred to as 'histone erasers'[102].

In addition to writers and erasers, a third class of proteins vital for the biological functions of the histone code exists: the histone readers. These proteins have domains that allow them to specifically bind modified histone tails[103]. In fact, the second, indirect mechanism employed by chromatin to regulate gene expression operates through the attraction of these histone readers, which are then able to modulate the transcription process in various ways[103]. An example of a well-characterized histone reader is BRD4. The BRD4 protein has two tandem bromodomains that bind preferentially to acetylated histones (Figure 8C). Active promoter regions for example show high levels of histone acetylation, which are therefore recognized by BRD4. Bound at these sites, BRD4 is able to facilitate the activation of transcription by recruiting additional cofactors such as CDK9 and Mediator (Figure 7). Thus, chromatin can operate as a scaffold that allows other proteins to dock and regulate transcription. It is important to note that DNA itself can also be modified and that these modifications can have a profound impact on transcription. More specifically, DNA can be

methylated, which is generally associated with gene silencing[104]. Mechanistically, DNA methylation impacts transcription in a similar fashion to modifications of the histone octamer: direct by precluding TFs from binding to their target DNA sequence[105] or indirect by recruiting proteins that specifically bind methylated DNA[104]. The chromatin modifications described above, encompassing modifications of the DNA itself, the

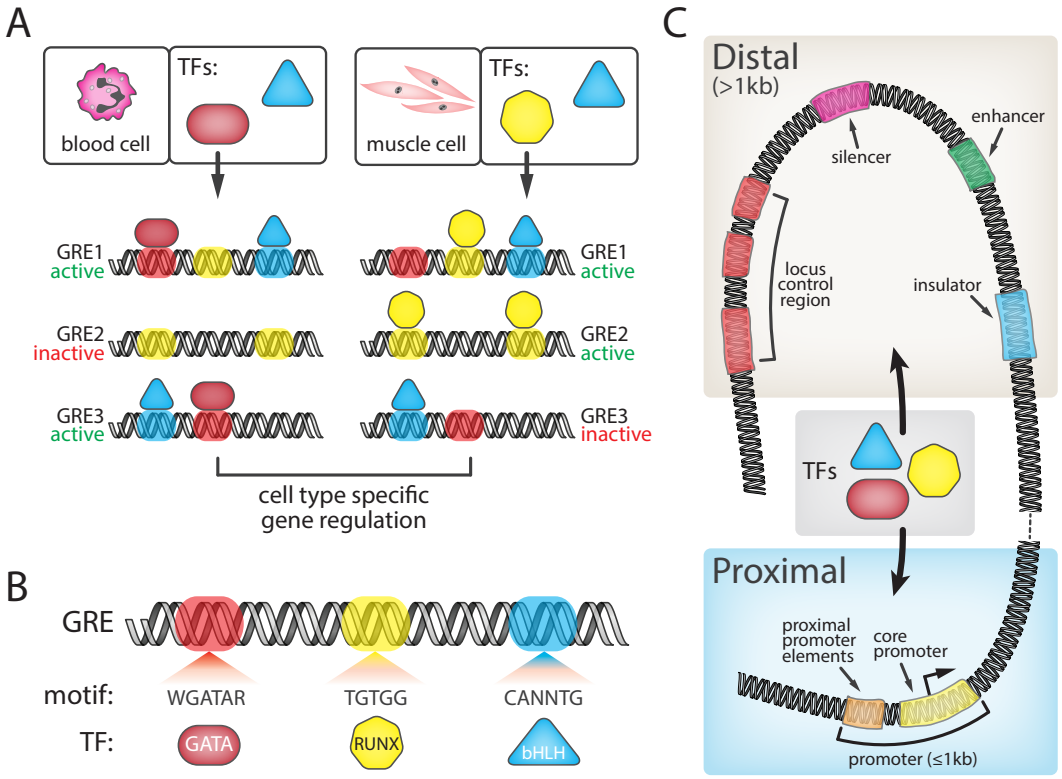


Figure 9: Transcription factors bind gene regulatory elements to regulate gene expression. (**A**) Due to their unique transcriptional programs, each cell type produces a specific combination of transcription factors (TFs). TFs bind gene regulatory elements (GREs) in a combinatorial fashion, resulting in a tissue-specific activity pattern of GREs and, as a consequence, a unique gene expression profile. (**B**) TFs recognize short degenerate sequences called motifs that cluster in GREs. Here, examples of 3 classes of TFs (GATA, RUNX and basic helix-loop-helix [bHLH] factors) and their core DNA binding motifs are shown (W=A or T, R=A or G, N=A, T, C or G). (**C**) The types of GREs to which TFs are recruited can be divided into 2 broad categories: proximal (<1kb from the transcription start site [TSS], denote by the arrow in the core promoter region) and distal (>1kb from the TSS). Proximal GREs are located within the promoter region. See Figure 10 for a detailed description of the different classes of distal GREs.

histone octamer and nucleosome positioning, are collectively referred to as epigenetic changes. Unlike genetic changes, which involve modification of the actual DNA sequence, epigenetic modifications are functionally relevant changes to the genome that do not entail changes in DNA sequence[106].

Transcription factors: executive managers of gene expression

Nucleosome remodellers, histone modifiers, Mediator and the basal transcription machinery are all indispensable for a controlled transcriptional output of our genome. However, these proteins act in a very general manner. In contrast, cellular differentiation is a dynamic process that requires very specific alterations to be made to a cell's transcriptional program over time. To direct the gene regulatory

machinery to the right locations at the right moment, a large class of specialized proteins has evolved. These sequence-specific DNA binding proteins are called transcription factors (TFs) and recognize specific short DNA sequences present in gene regulatory regions such as promoters[107]. The key concepts of how a TF controls gene transcription were first established in bacterial systems. François Jacob and Jacques Monod, two French pioneers of transcriptional regulation, were the first to propose a gene regulatory mechanism involving a protein that repressed transcription by directly binding its target gene[108]. These discoveries laid the foundation for our current understanding of eukaryotic transcriptional regulation, which follows many of the principles identified in bacteria. Jacob and Monod were awarded the 1965 Nobel Prize in Physiology and Medicine for their seminal contributions to our understanding of transcriptional regulation.

Unlike the ubiquitous general TFs involved in recruiting RNAPII to the genome, most TFs have a limited set of target genes they control, ranging from a few hundred to several thousand genes. The human genome contains approximately 1400 TF genes[109]. As combinations of TFs are expressed in a highly tissue- or stage-specific manner, the activity of the gene regulatory regions they occupy is often restricted to certain cell types or developmental stages (Figure 9)[107]. This concept represents the fundamental principle behind the precise spatiotemporal expression of genes.

As I will discuss in the next section, TFs use different types of regulatory regions at which they control gene expression. However, the mechanisms they employ are often very similar among the different classes of gene regulatory sites. As stated before, TFs target the different protein complexes involved in modifying chromatin structure and the initiation of transcription to the DNA. Whether TF binding will result in gene activation or repression depends on the combination of cofactors it recruits. Several strategies can be used by TFs to modulate transcriptional output[73,87,107,110,111]:

- Directly interact with components of the general transcription machinery (e.g. GTFs, Mediator) to promote or disrupt PIC formation.
- Promoting or inhibiting different steps in the transcription process (i.e. initiation, pause-release, elongation). This can occur through direct interactions with the transcription machinery or via cofactor recruitment.
- Recruitment of chromatin modifying complexes, either directly or via intermediate cofactors. The type of chromatin modification (e.g. histone acetylation or nucleosome remodeling) determines the effect on transcription.
- Facilitating the binding of other TFs (either directly or via cofactor recruitment). Prime examples are the 'pioneer' TFs. These proteins are able to bind DNA within regions of inaccessible chromatin and establish a permissive chromatin environment that allows other factors to bind.
- Competition with other, or displacement of already bound, TFs.

The presence or absence of a TF can literally be a transforming event. The power of TFs in shaping a cell's transcriptome and identity is illustrated by early studies on the lineage-instructive role of TFs. In a classic study from the Graf laboratory[112], the role of the GATA1 TF in hematopoietic cells was investigated. GATA1 is an important transcriptional regulator expressed in erythroid cells, megakaryocytes, eosinophils and dendritic cells, but is absent in granulocytes and macrophages[113]. Surprisingly, when they introduced GATA1 expression into macrophage-granulocyte precursor cells, the cells began to switch lineage (or 'transdifferentiate'): their cellular identity changed to that of cells resembling eosinophils and MEPs. Introduction of GATA1 in the macrophage-granulocyte progenitors turned out to repress the macrophage-granulocyte transcriptional program (via the repression of the myeloid master TF PU.1) and simultaneously activate genes characteristic for eosinophils and MEPs[114]. Thus, a single TF can determine cellular fate by rewiring the regulatory connections that confer cell type-specific gene expression.

It is important to realize is that the actions of TFs are often endpoints of an upstream signal transduction pathway[87]. As previously mentioned, cells need to interpret extracellular signals and respond to these by implementing the required changes to their transcriptional program. Signalling cascades activated upon binding of signalling molecules to cell surface receptors ultimately result in the activation of one or more TFs. In the case of the EPO receptor signalling pathway discussed before, binding of EPO to its receptor results in the activation of the JAK2 kinase (as well as other signaling cascades)[115]. Activated JAK2 will phosphorylate a latent cytoplasmic TF called STAT5, which then translocates to the nucleus and

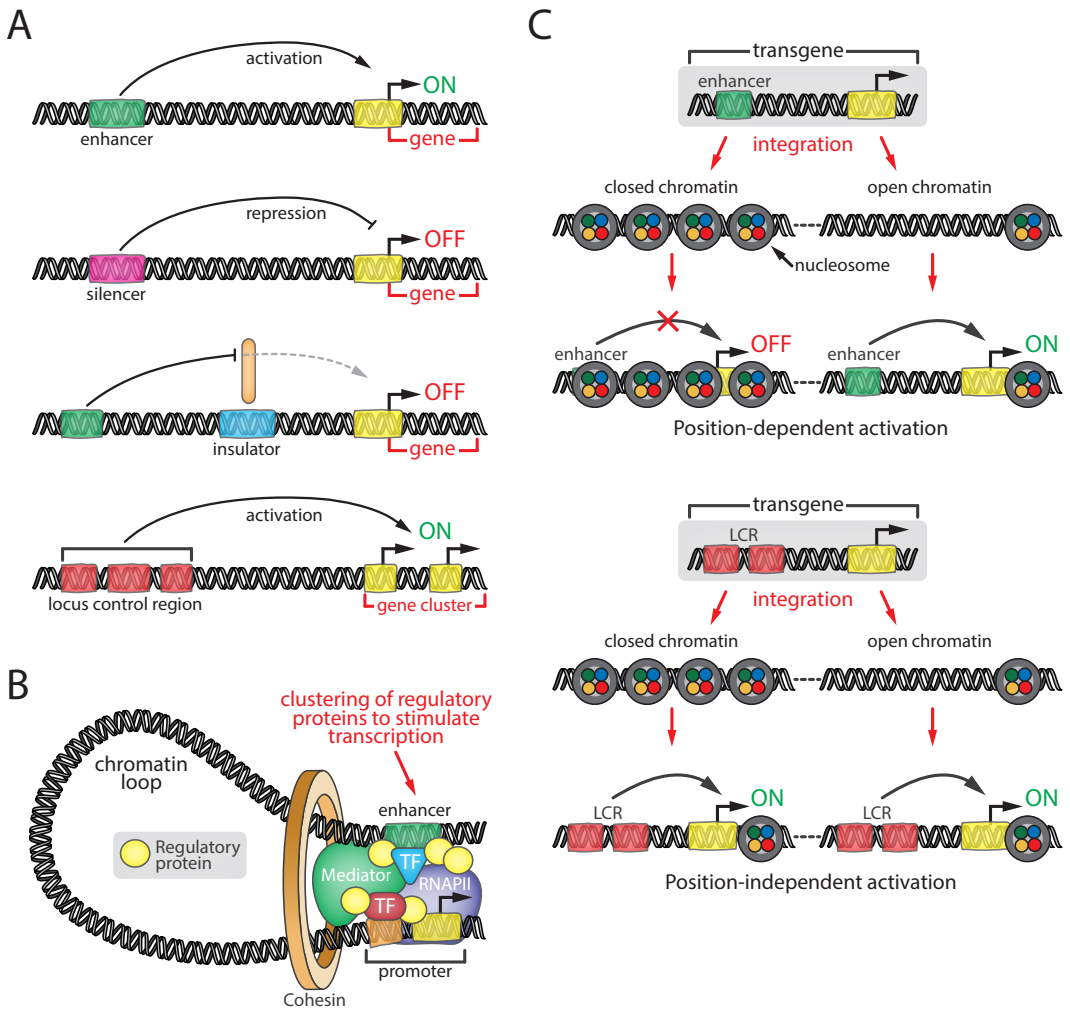


Figure 10. Functions of the different classes of distal gene regulatory elements. (**A**) Schematic representation of the four common types of distal gene regulatory elements (GREs) and their actions. Enhancers and silencers can activate or silence a gene from a distance. Insulators can act as barriers to prevent long-range GRE-to-gene interactions from inappropriately regulating non-target genes. Locus control regions (LCRs) are composed of combinations of (different) distal GREs that cooperate to regulate the expression of a locus or gene cluster. (**B**) Speculative model of how chromatin loop formation can bring distal GREs (an enhancer in this case) in close physical proximity to a gene. This way, TFs and regulatory complexes (e.g. chromatin modifiers, co-activators/repressors) are clustered around the transcription start site of a gene to regulate (in the case of enhancers: to stimulate) its transcription. Recent evidence suggests that the Cohesin complex, which forms a ring-like structure, in involved in maintaining these chromatin loops. (**C**) Unlike a conventional enhancer, LCRs can activate gene expression in a position-independent fashion at ectopic sites. Upon random integration in the genome, enhancers are unable to establish gene regulation when positioned in a closed chromatin environment. However when an LCR integrates into such an inhibitory region of chromatin, it has the remarkable ability to create an open chromatin domain and establish gene activation. This phenomenon of position-independent gene activation is unique to LCRs and is used to operationally define these GREs.

regulates genes important for erythroid survival and differentiation. Alternatively, in nuclear hormone receptor signalling pathways (e.g. estrogen signalling), the activated receptors themselves act as TFs[116].

Regulatory elements: docking sites for transcription factors

TFs are modular proteins that typically recognize small (6-12 bp) degenerate sequences called motifs. TFs have a DNA binding domain (DBD) responsible for binding a specific DNA motif and are categorized into different TF families based on the type of DBD they possess[117]. Well-known DBDs include the homeodomain, the basic helix-loop-helix (bHLH) domain and the zinc finger. Gene regulatory regions typically contain clusters of different TF binding motifs, immediately suggesting that TFs operate in a combinatorial fashion. Indeed, TFs often cooperate or even compete to bind to motifs closely clustered in regulatory regions (Figure 9A-B)[107]. The relative position and orientation of the different motifs, collectively named 'motif grammar', can have important implications for the recruitment of TFs to a particular site (comprehensively reviewed elsewhere[107]). For example, the spacing between two individual binding motifs can be important for cooperative DNA binding by TFs.

The non-random distribution of TF binding motifs within the genome determines the positions to which a specific TF can bind. So where in the genome are these TF binding sequences located? One obvious location is the gene promoter region to which the basal transcription machinery is recruited to initiate transcription. It has been well established in different species, ranging from unicellular yeast to humans, that TFs commonly bind near or within promoter regions to modulate gene expression (Figure 9C)[110]. Such binding sites are often referred to as 'proximal promoter elements' to distinguish them from the core promoter region bound by the basal transcription machinery[79,118]. However, it appears that higher organisms have evolved more elaborate ways of controlling gene transcription.

In a simple eukaryote like yeast, transcriptional control is a local, promoter-centered, affair. In contrast, metazoans like the fruit fly, mouse and human rely predominantly on gene regulation via a combination of promoters and multiple distal gene regulatory regions (Figure 9C)[110]. The most prominent and intensively studied distal regulatory elements are the enhancers (Figure 10A), which are usually found in non-coding regions of the genome (such as introns and intergenic regions). Enhancers are small (±500bp[110]) regions littered with TF binding motifs occurring at various distances from their target genes (between 1 to >1000 kb, with a median distance of 120kb[119]). They display a characteristic chromatin signature (high H3K4Me1 and H3K27Ac levels, low levels of H3K4Me3 and low nucleosome density), which can be used to accurately predict their location[101]. Enhancers are important for the activation of gene expression and multiple enhancer elements often regulate the expression of a single target gene[81,119]. Characteristically, most enhancers function independent of both the distance and orientation relative to their target gene[81].

Exactly how enhancer regions are able to influence the transcription of genes hundreds of kilobases away still poses an enigmatic aspect of their biology. Inspired by studies of several model loci, a general consensus has emerged[120]. The current dominant model for enhancer function is a direct interaction between the enhancer and its target gene via a process of 'chromatin looping'[118]. This model requires the intervening chromatin to be 'looped out' in order to permit enhancer-gene interactions. As a result, regulatory protein complexes occupying the enhancers are delivered to for example the gene promoter region, where they are able to stimulate transcription (Figure 10B). A growing body of evidence suggests that TFs are involved in stabilizing these interactions and loops[121]. Recently, the Mediator and Cohesin protein complexes have also been implicated in orchestrating chromatin looping[122-124]. Especially the involvement of the Cohesin complex, a protein ring structure first identified as a regulator of sister chromatid cohesion, provides an exciting mechanism for holding a chromatin loop in place (Figure 10B)[121]. Additionally, it has recently been proposed that non-coding RNAs, some even produced from the enhancer itself, play a role in enhancer-gene communication[125].

First discovered in SV40 tumor virus DNA (1981[126]), enhancers are essential activators of gene transcription that very often function in a highly tissue-specific manner. Operated by TFs, enhancers are the prime determinants of tissue-specific gene expression and therefore key components in controlling development and differentiation[118]. The first mammalian enhancer described, the intronic enhancer of the *Igh* locus[127,128], illustrates this concept. The powerful enhancers of the *Igh* and *Igl* loci regulate the V(D)J recombination process at multiple levels (see Chapters 7 and 8)[129]. The occurrence of the V(D)J recombination process needs to be flawlessly controlled: it needs to be activated only in subsets of early B cells (see previous section of this chapter). The presence of a B cell-specific TF complement ensures that the activity of these enhancers is indeed restricted to early B cells only[61].

In addition to enhancers and promoters, several other classes of gene regulatory elements (GREs)

are now known to be involved in metazoan transcriptional regulation. It turns out that the non-coding portion of our genome (>95% of the total genome size), not too long ago referred to as 'junk-DNA', is full of promoters, enhancers, silencers, insulators and locus control regions (LCRs)[81]. That is why the non-coding genome, now sometimes referred to as a 'regulatory treasure box', in principle controls transcription of the coding genome. The latest estimates even predict a staggering 40% of our genome carrying regulatory potential[119]. The function and genomic position of the different types of GREs is illustrated in Figure 10A. Through the recruitment of TFs and regulatory protein complexes, enhancers stimulate transcription; silencers do the opposite[81]. Insulators have two main properties: 1) they block enhancer-gene communication (Figure 10A), effectively restricting the action of regulatory elements to confined regions, and 2) they can function as a barrier to separate chromatin domains, e.g. to prevent heterochromatin from spreading into an area of active transcription[130]. They are often marked by binding of the CTCF TF and function in a position-dependent fashion[81].

LCRs were first described in 1987 by Grosveld and colleagues[131]. They showed that to obtain uniform, high-level expression of a human β-globin transgene in mice, several up- and downstream regulatory elements needed to be included in the transgene. This collection of regulatory elements was called the β-globin locus control region, and a number of other LCRs were identified subsequently[132]. LCRs exert a particularly potent transcription-enhancing activity and are frequently involved in regulating the tissue-specific expression of an entire locus or gene cluster. What sets them apart from traditional enhancers and is used to operationally define LCRs is their ability to create and maintain a region of open chromatin when integrated into ectopic genomic sites (Figure 10C)[132]. Since its discovery, the β-globin LCR has served as a paradigm for control of tissue-specific and developmentally regulated transcription. In fact, the first compelling evidence supporting a chromatin-looping model of enhancer function was deduced from studies of the β-globin LCR[133,134].

Complexity of transcriptional regulation: driving force behind biological complexity?

Neither the size of an organism's genome nor the number of protein-coding genes, is strongly correlated to the biological complexity of an organism – an observation that has puzzled researchers for quite some time[139]. Recent advances in our ability to study epigenetic modifications and RNA transcription on a genome-wide level have provided several possible explanations for this conundrum. In general, higher organisms have evolved more complex ways of regulating and diversifying the output of their genomes[119]. For example, mechanisms such as alternative splicing or the number of non-coding RNAs present in a genome appear to positively correlate with biological complexity[140,141]. The evolution of more elaborate ways of regulating gene expression also provides a plausible explanation for the increased complexity of mammals when compared to yeast or fruit flies[110,119], in agreement with the strong positive correlation that exists between the amount of non-coding genomic sequence and biological complexity[142]. Moreover, the proportion of total genes encoding TFs increases with complexity (from 3.4% in yeast, 4.2% in roundworms and 5.5% in fruit flies to 8-9% in humans[143]). Being able to tweak transcriptomes in more ways allows for the regulatory fine-tuning required for the development of uniquely complex organs such as the human brain.

In 2013, the Richard Young laboratory reported the identification of a novel class of enhancers they dubbed 'super-enhancers'[135] (also called 'stretch-enhancers'[136]). Super-enhancers form unusually large hotspots of TF binding (>5000 bp) with exceptionally high enrichments for many TFs and cofactors, in particular the Mediator complex. Interestingly, they were found to control the expression of key cellular identity genes (e.g. tissue-specific TFs) and represent prime targets of signalling pathways[137,138]. There appears to be a significant overlap between LCRs and super-enhancers, as computer algorithms used to identify the latter also recognized known LCRs, including the β-globin LCR[136]. Additional studies will have to reveal the possibly unique functional characteristics of super-enhancers.

Introduction to the transcription factors and signalling proteins studied in this thesis

How TFs (and the signal transduction pathways they respond to) orchestrate gene regulatory processes,

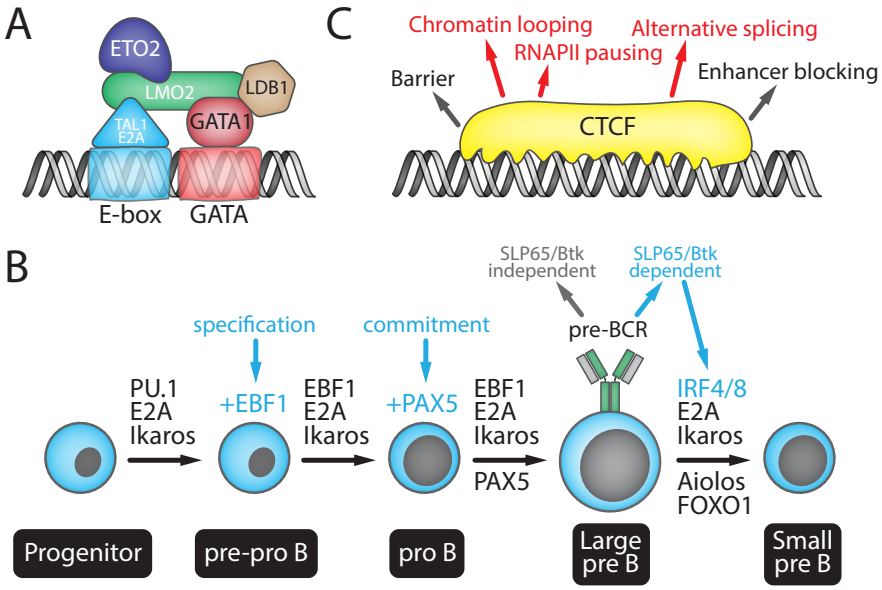


Figure 11: Erythroid transcription factors, key regulators of early B cell development and the CTCF protein. (A) Cartoon representing the core components of the erythroid LDB1-complex: the DNA-binding TFs GATA1 (recognizes a GATA motif) and a TAL1-containing bHLH heterodimer (recognizes an E-box motif; TAL1-E2A is depicted here), the LMO2 bridging protein, the ETO2 cofactor and the LDB1 adapter protein. The schematic representation does not take into account the right size dimensions of the double helix with regard to the schematized proteins. (B) Schematic of early B cell development with the key (transcriptional) regulators depicted for every step. Critical events/factors are shown in blue: EBF1 activation marks B cell specification (pre-pro B), PAX5 activation marks B cell commitment (pro B) and Slp65/Btk-mediated pre-BCR signaling activates IRF TFs to induce *Igl* rearrangements (pre B). (C) Cartoon depicting the large multi zinc finger CTCF protein bound to DNA. Several classic (grey) and novel (red) functions of this versatile protein are shown.

and as a consequence cellular development, is the central theme of investigation in this thesis. We chose hematopoietic differentiation as a model system, specifically focusing on definitive erythropoiesis and early B cell development. We focused our studies on several TFs and signalling proteins known to be essential for these differentiation processes, which are introduced below.

Key erythroid TFs: the LDB1-complex. The different TFs involved in red blood cell development have been extensively characterized over the past 25 years[144]. Undoubtedly the best studied erythroid TF is GATA1, the founding[145] member of a TF family that recognizes a core 'GATA' DNA motif. GATA1 is characterized by the presence of two conserved zinc finger domains that mediate its interaction with DNA and other transcriptional regulators[113]. GATA1 is absolutely essential for proerythroblast differentiation in both cell lines and animal models[113]. More recent studies have addressed the genome-wide chromatin occupancy of GATA1 in erythroid cells[146-148], providing unique insight into the genes it regulates. Perhaps not surprisingly, GATA1 was found to control virtually all genes critical for erythroid differentiation and identity. Numerous proteins have been found to interact with GATA1. Very prominent among these partners are the TAL1, LMO2 and LDB1 proteins, all of which are essential proteins for hematopoietic and/or erythroid development[144,149]. This multimeric complex, hereafter referred to as the 'LDB1-complex' (Figure 11A), appears to play a dominant role in erythroid gene regulation, especially in the context of gene activation[148]. The LDB1-complex is recruited to composite E-box/GATA motifs through the cooperative action of a TAL1-containing bHLH heterodimer (recognizing the E-box part) and GATA1 (binding the GATA part). LMO2 acts as a bridging molecule connecting the two DNA-binding modules (Figure 11A)[150], while LDB1 appears to act as a protein interaction interface for cofactor recruitment[151]. Transcriptional regulation by the LDB1-complex is primarily achieved through binding of distal regulatory elements, as exemplified by

its key role in long-range β-globin gene activation via the LCR[152]. Critical for establishing LCR-β-globin gene communication is the LDB1 subunit, which has the ability to dimerize through its self-association domain, enabling the juxtaposition of LDB1-complexes bound at the LCR and the β-globin gene promoter[153]. The LDB1-complex, its associated cofactors and its gene regulatory actions are the main subject of investigation in Chapters 2, 5 and 6.

The early B cell TF network and pre-BCR signalling. As is the case for red blood cell development, many TFs have been implicated in driving early B cell development. Mouse models in which TFs can be specifically deleted in early hematopoietic or B cell progenitors have identified key TFs indispensable for B cell development. These studies have resulted in an elegant multistep model of B cell development directed by TFs (Figure 11B)[61,154,155]. In early (lymphoid-primed) hematopoietic progenitors (i.e. LMPPs and CLPs, Figure 3) the PU.1, Ikaros and E2A TFs induce B cell specification by activating the expression of the EBF1 TF, resulting in differentiation towards the pre-pro B cell stage. EBF1 then induces expression of the B cell commitment TF PAX5, which 'locks in' the B cell expression program and silences any remnants of T cell or myeloid expression programs[156]. In these pro-B cells, *Igh* locus recombination is initiated and depends on the presence of several TFs including E2A, Ikaros and PAX5. At the subsequent pre-B cells stage, after succesfull *Igh* rearrangement and pre-BCR assembly, SLP65/Btk-mediated pre-BCR signalling actives IRF4/8[64]. IRF TFs then strongly induce Aiolos and Ikaros expression to downregulate pre-BCR signalling, promote cell-cycle withdrawal and initiate *Igl* locus recombination[157]. How exactly TFs control V(D)J recombination of the *Ig* loci upon is still not entirely clear, but it involves several enhancers and dynamic changes in non-coding transcription, chromatin structure and locus topology[63]. Studies on early B cell development presented in Chapter 8 address the role of E2A and Ikaros during *Igκ* locus recombination in the context of an *in vivo* pre-B cell signalling gradient.

The insulator protein CTCF. The mammalian insulator-binding protein CTCF is required for the development of a broad range of cell types[158]. It contains a highly conserved DBD consisting of 11 zinc fingers and occupies >50,000 sites in mammalian genomes, of which many are ultraconserved between tissues and species[159]. The ubiquitous requirement for CTCF is a consequence of its important general functions in transcriptional regulation (Figure 11C). In addition to its classical role as an enhancer-blocking protein and a chromatin barrier, CTCF also appears to play an important role in facilitating long-range enhancer-gene communication and chromatin folding (see article by Ong and Corces for an excellent review on CTCF function[159]). This 'chromatin organizing' function of CTCF has sparked an intense research effort, and CTCF, along with its interaction partner Cohesin and the Mediator complex, is now referred to as an architectural protein[124]. As mentioned above, antigen receptor loci undergo dramatic changes in local chromatin folding. Therefore, and because of CTCF's general importance for gene regulation and development, we attempted to comprehensively address CTCF function during early B cell development *in vivo*, as described in Chapter 7.

Transcriptional regulation: relevance for human phenotypic variation and disease

Studying gene regulatory mechanisms is crucial if we are to understand how an organism develops and functions. However, studying transcriptional control is not just an academic exercise, as it has important implications for human health. Many diseases and syndromes, such as cancer, autoimmunity, diabetes and developmental disorders are associated with misregulation of gene expression (reviewed by Lee and Young[73]). Also among healthy individuals, phenotypic variation (e.g. differences in blood hemoglobin levels) and disease susceptibility (the risk of acquiring a particular disease) have been linked to changes in gene regulatory mechanisms[160,161]. In this section, I will illustrate several ways in which altered transcriptional regulation can cause disease and how such alterations underlie phenotypic variation, including susceptibility to common diseases.

Dysfunctional regulatory proteins and elements as a cause of disease

It is not difficult to imagine the potentially catastrophic consequences of a dysfunctional TF or chromatin-modifying enzyme at work. Genes under the control of the affected regulatory protein(s) are at high risk of misregulation, which can result in a multitude of detrimental effects. In the case of cancer for example, genes that promote cell proliferation (called 'oncogenes') can be aberrantly activated, leading to

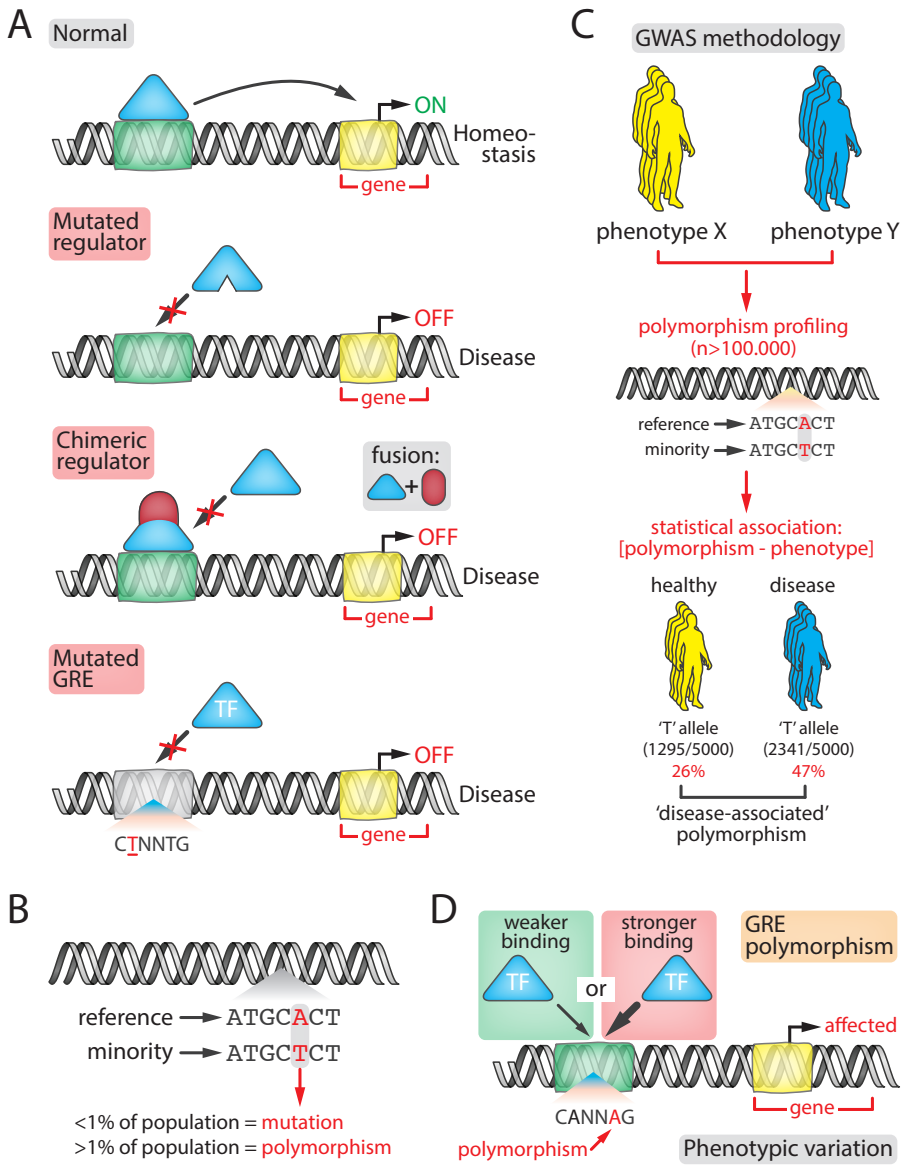


Figure 12: Transcriptional regulation and its relevance to human disease and phenotypic variation. (**A**) Genetic mutations can affect gene regulatory mechanisms and result in disease. Mutations in coding sequences of a regulator gene (e.g. a TF) can for example alter the DNA binding capacity of this regulatory protein, resulting in target gene misregulation and the possible development of disease. Chimeric regulators created by chromosomal abnormalities combine characteristics of 2 regulatory proteins in a single fusion protein, which can also have detrimental effects on gene regulation. Apart from interfering with protein function, mutations can also cripple GREs through a loss (or gain) of TF binding motifs, again resulting in aberrant gene expression. (**B**) The difference between a mutation and polymorphism is based solely on the frequency with which they occur in a population: mutation are rare (<1%), while polymorphisms are more common (>1%). (**C**) Overview of the methodology used in a genome-wide association study (GWAS). By genotyping thousands of polymorphisms in large groups of individuals with distinct phenotypes (e.g. healthy vs. suffering from a disease or blue eyes vs. brown eyes), statistical associations between the presence of a specific polymorphism and phenotype can be determined. (**D**) Many of the GWAS-identified polymorphisms localize to gene regulatory elements (GREs). Here they can affect TF binding affinity and subsequent gene regulation, resulting in (subtle) phenotypic variation.

uncontrolled cell growth and tumor formation[73]. Many of the hematopoietic TFs studied in this thesis (e.g. TAL1, LMO2) actually function as a double-edged sword: while they are indispensable for normal blood cell development, they can also act as powerful drivers of leukemia[162].

Three general mechanisms can cause transcriptional regulation to go astray (Figure 12A)[73,163,164]. First, mutations in the coding sequence of the gene encoding a regulatory factor can interfere with protein production (by introducing a premature stop codon), stability (amino acid changes result in an unstable protein) or function (mutant proteins can interact with an altered DNA motif or different cofactors). For example, mutations in the *MECP2* gene, encoding the MeCP2 protein, cause a severe neurodevelopmental disorder called Rett syndrome. The *MECP2* mutations affect either the DNA binding domain (MeCP2 binds methylated DNA) or an important corepressor-interaction domain, in both cases impairing the protein's normal gene regulatory function[165].

Second, chromosomal abnormalities (e.g. deletions, inversions, translocations) can result in the loss, duplication or even fusion of critical regulator genes[163]. The protein products of the latter events, so-called fusion proteins, are notorious drivers of cancer. For example, a translocation involving the *RUNX1* gene (on chromosome 21, encoding the RUNX1 TF) and the *ETO* gene (on chromosome 8, encoding the ETO corepressor) creates a fusion gene encoding the 'chimeric' RUNX1-ETO protein[162]. RUNX1-ETO possesses the same DNA-binding capacity as the normal RUNX1 protein, but now combines this with the repressor function of ETO. RUNX1-ETO will compete for DNA binding with the normal RUNX1 protein, resulting in the misregulation of RUNX1 target genes and ultimately in leukemia development[166].

Third, interference with the regulatory landscape that controls the expression level of the regulator gene. This can occur through mutations and chromosomal abnormalities that involve GREs[164]. Historically, scientists have focused their efforts mostly on protein-coding regions when trying to discover disease-associated mutations. Now that genome-wide identification of GREs has become feasible, the number of diseases and syndromes identified as having a genetic cause involving GREs has rapidly increased[164]. Nevertheless, scientists already observed a causal relationship between GREs and human disease several decades ago when studying gene regulation at model loci. A classic example is Burkitt's lymphoma (a highly aggressive B cell cancer), which is caused by chromosomal translocations that place the *MYC* oncogene in close proximity to a powerful *IgH* or *IgL* enhancer[167]. This results in a dramatic and B cell-specific increase of MYC production, with lymphoma development as the consequence. Another early example is a rare deletion of the β-globin LCR[168], resulting in a loss of β-globin expression and severe anemia (a syndrome referred to as β-thalassemia[169]).

Genetic variation, gene regulatory mechanisms and human phenotypes

The above-mentioned mutations and chromosomal anomalies are rare and tend to have profound effects on gene regulatory mechanisms, therefore resulting in dramatic phenotypes (like the development of cancer). However, the overwhelming majority of mutations and small insertions/deletions have no or very small effects on cellular phenotype, and are therefore well tolerated[170]. Some of these genetic variants can be quite common (present in >1% of the population, Figure 12B) and are then referred to as 'polymorphisms'. Based on our current knowledge of human genetic variation, a pair of random humans is expected to differ at 1 position every 1000 nucleotides of DNA sequence[171].

Single nucleotide polymorphisms (SNPs) are the most abundant type of polymorphism in the human genome (>38 million SNPs have been identified in the human genome[172]). Although rarely causing disease, common genetic polymorphisms can have a significant impact on several aspects of human health and phenotypic variation, ranging from eye color[173] to average red blood cell size[174]. Most of the correlations between the presence of certain variants and a particular human phenotype or trait have been obtained from genome-wide association studies (GWAS). A GWAS usually involves thousands of individuals for whom both genotype (in most cases a large number of SNPs) and phenotype (e.g. eye color or suffering from a specific disease) are determined. If the presence of a specific variant is more frequent in individuals with the particular phenotype or disease investigated, the variant is said to be 'associated' with that phenotype or disease (Figure 12C)[175]. Perhaps not too surprising, the vast majority of GWAS-identified variants (93%) reside in non-coding regions of our genome[160], complicating their functional evaluation. An obvious explanation for this phenomenon is that the associated variants affect GREs, a concept that has received support from several recent studies[160,176].

Although the GWAS approach has been a controversial one[177], it has yielded several success stories that illustrate how common genetic variation can significantly influence human traits and disease susceptibility. One such story is the identification of a cancer-associated SNP 500 kb upstream of the *MYC* oncogene. Individuals with a T nucleotide at this position (a T-allele, instead of the G-allele) have a reduced risk of developing colorectal cancer[178]. It turned out that the T-allele disrupts a TF binding motif in a long-range *MYC* enhancer[179,180], resulting in a modest reduction of *MYC* expression[181]. Mice lacking this conserved enhancer were resistant to intestinal tumor formation[182], confirming the protective effect of the T-allele.

Another successful 'translation' of a GWAS-identified association into a potentially clinically relevant insight is the case of persistent fetal hemoglobin (HbF) expression in adults. Around birth, humans progressively silence expression of the fetal γ globin genes and activate adult β globin gene expression (as previously discussed in this chapter). As a result, most adults produce only minor amounts of HbF (<1% of total Hb levels)[42]. However, already in the early 1960s a screening of 3000 Swiss army recruits showed that HbF levels vary considerably among humans, and individuals with as much as 30% HbF have been described[183]. The persistence of high HbF levels has no negative effects on adult human health. However, the clinical observation that HbF can completely compensate for the absence of functional adult hemoglobin (HbA) in individuals suffering from β-hemoglobinopathies (e.g. β-thalassemia) sparked great interest into adult HbF persistence[184]. Currently, therapeutic 'reactivation' of γ globin gene expression is the holy grail of β-hemoglobinopathy research.

Twin studies already showed that 89% of the variation in HbF levels could be explained by genetic factors[185]. Subsequent GWASs in both healthy populations and β-hemoglobinopathy patients identified three loci that explain ±50% of the variation in HbF levels: the β-globin locus itself, the second intron of the *BCL11A* gene and an intergenic region between the *HBS1L* and *MYB* genes (the *HBS1L-MYB* intergenic region)[183,186]. In addition to these common genetic variants, rare loss-of-function mutations in the KLF1 erythroid TF were also shown to result in elevated HbF levels[187,188]. Follow-up studies of the common HbF-associated variants by the Orkin laboratory demonstrated that the BCL11A TF is a potent repressor of γ globin expression[189] and that the associated intronic *BCL11A* variants affect the long-range regulation of *BCL11A* expression[190]. Work described in Chapter 6, performed in collaboration with the laboratory of Swee Lay Thein, addresses the association between *HBS1L-MYB* intergenic variation and HbF levels. In a joint effort, we could show that these variants appear to disrupt the function of erythroid-specific intergenic enhancers that regulate *MYB*, which encodes a TF that regulates HbF levels in adults. Together, these findings provide a mechanistic explanation for the original GWAS results. Additionally, in combination with the rapid advances in genome editing technology[191], they suggest that the GWAS-marked *BCL11A* and *MYB* enhancers are potential therapeutic targets. Excitingly, recent work from the Orkin laboratory provided a first *in vitro* proof-of-principle that supports 'enhancer editing' as a viable strategy for γ globin reactivation[190].

Understanding our genome: a perspective

More than ever it truly seems that the human genome can be considered, as recently coined by John Mattick, a 'ZIP file extraordinaire'[192]. The combinatorial actions of the numerous different TFs, cofactors, chromatin modifiers and GREs it encodes allows for a myriad of gene regulatory possibilities, allowing it to store staggering amounts of information. Although we have gained significant insight into the molecular mechanisms controlling gene expression, much still remains unclear. Truly understanding all the information encoded in our genome, how it is put to use to orchestrate embryonic development and to sustain adult life, and how it is 'misused' under pathological conditions, remain some of the greatest challenges facing modern biology. The studies described in Chapters 2-8 of this thesis aim to contribute to addressing these challenges.

References

1 Aristotle & Peck, A. L. *Generation of animals*. (W. Heinemann ; Harvard University Press, 1943).
2 Magner, L. N. *A history of the life sciences*. 3rd edn, (M. Dekker, 2002).
3 Correia, C. P. *The ovary of Eve : egg and sperm and preformation*. (University of Chicago Press, 1997).
4 Mazzarello, P. A unifying concept: the history of cell theory. *Nature cell biology* **1**, E13-15, doi:10.1038/8964 (1999).
5 Harris, H. *The birth of the cell*. (Yale University Press, 1999).

6 Alberts, B. *Molecular biology of the cell*. 5th edn, (Garland Science, 2008).

7 Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 7634-7638 (1981).

8 Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154-156 (1981).

9 Keller, G. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes & development* **19**, 1129-1155, doi:10.1101/gad.1303605 (2005).

10 Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

11 Graf, T. Historical origins of transdifferentiation and reprogramming. *Cell stem cell* **9**, 504-516, doi:10.1016/j.stem.2011.11.012 (2011).

12 Briggs, R. & King, T. J. Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proceedings of the National Academy of Sciences of the United States of America* **38**, 455-463 (1952).

13 Gurdon, J. B. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Journal of embryology and experimental morphology* **10**, 622-640 (1962).

14 Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810-813, doi:10.1038/385810a0 (1997).

15 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676, doi:10.1016/j.cell.2006.07.024 (2006).

16 Hanahan, D., Wagner, E. F. & Palmiter, R. D. The origins of oncomice: a history of the first transgenic mice genetically engineered to develop cancer. *Genes & development* **21**, 2258-2270, doi:10.1101/gad.1583307 (2007).

17 Reeves, R. H. *et al.* A mouse model for Down syndrome exhibits learning and behaviour deficits. *Nature genetics* **11**, 177-184, doi:10.1038/ng1095-177 (1995).

18 Nguyen, D. & Xu, T. The expanding role of mouse genetics for understanding human biology and disease. *Disease models & mechanisms* **1**, 56-66, doi:10.1242/dmm.000232 (2008).

19 Robinton, D. A. & Daley, G. Q. The promise of induced pluripotent stem cells in research and therapy. *Nature* **481**, 295-305, doi:10.1038/nature10761 (2012).

20 Till, J. E. & Mc, C. E. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation research* **14**, 213-222 (1961).

21 Till, J. E., McCulloch, E. A. & Siminovitch, L. A Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells. *Proceedings of the National Academy of Sciences of the United States of America* **51**, 29-36 (1964).

22 Becker, A. J., Mc, C. E. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**, 452-454 (1963).

23 Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631-644, doi:10.1016/j.cell.2008.01.025 (2008).

24 Bryder, D., Rossi, D. J. & Weissman, I. L. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *The American journal of pathology* **169**, 338-346, doi:10.2353/ajpath.2006.060312 (2006).

25 Bunting, K. D. & Hawley, R. G. Integrative molecular and developmental biology of adult stem cells. *Biology of the cell / under the auspices of the European Cell Biology Organization* **95**, 563-578 (2003).

26 Wagers, A. J. & Weissman, I. L. Plasticity of adult stem cells. *Cell* **116**, 639-648 (2004).

27 Barker, N., Bartfeld, S. & Clevers, H. Tissue-resident adult stem cell populations of rapidly self-renewing organs. *Cell stem cell* **7**, 656-670, doi:10.1016/j.stem.2010.11.016 (2010).

28 Catacchio, I. *et al.* Evidence for bone marrow adult stem cell plasticity: properties, molecular mechanisms, negative aspects, and clinical applications of hematopoietic and mesenchymal stem cells transdifferentiation. *Stem cells international* **2013**, 589139, doi:10.1155/2013/589139 (2013).

29 Li, Y. & Wingert, R. A. Regenerative medicine for the kidney: stem cell prospects & challenges. *Clinical and translational medicine* **2**, 11, doi:10.1186/2001-1326-2-11 (2013).

30 Duncan, A. W., Dorrell, C. & Grompe, M. Stem cells and liver regeneration. *Gastroenterology* **137**, 466-481, doi:10.1053/j.gastro.2009.05.044 (2009).

31 Huch, M. *et al.* In vitro expansion of single Lgr5+ liver stem cells induced by Wnt-driven regeneration. *Nature* **494**, 247-250, doi:10.1038/nature11826 (2013).

32 He, S., Nakada, D. & Morrison, S. J. Mechanisms of stem cell self-renewal. *Annual review of cell and developmental biology* **25**, 377-406, doi:10.1146/annurev.cellbio.042308.113248 (2009).

33 Li, L. & Clevers, H. Coexistence of quiescent and active adult stem cells in mammals. *Science* **327**, 542-545, doi:10.1126/science.1180794 (2010).

34 Shi, X. & Garry, D. J. Muscle stem cells in development, regeneration, and disease. *Genes & development* **20**, 1692-1708, doi:10.1101/gad.1419406 (2006).

35 Prendergast, A. M. & Essers, M. A. Hematopoietic stem cells, infection, and the niche. *Annals of the New York Academy of Sciences* **1310**, 51-57, doi:10.1111/nyas.12400 (2014).

36 Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242-245 (1996).

37 Muller, A. M., Medvinsky, A., Strouboulis, J., Grosveld, F. & Dzierzak, E. Development of hematopoietic stem cell activity in the mouse embryo. *Immunity* **1**, 291-301 (1994).

38 Boisset, J. C. *et al.* In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464**, 116-120, doi:10.1038/nature08764 (2010).

39 Gekas, C., Dieterlen-Lievre, F., Orkin, S. H. & Mikkola, H. K. The placenta is a niche for hematopoietic stem cells. *Developmental cell* **8**, 365-375, doi:10.1016/j.devcel.2004.12.016 (2005).

40 Kumaravelu, P. *et al.* Quantitative developmental anatomy of definitive haematopoietic stem cells/long-term repopulating units (HSC/RUs): role of the aorta-gonad-mesonephros (AGM) region and the yolk sac in colonisation of the mouse embryonic liver. *Development* **129**, 4891-4899 (2002).

41 Boisset, J. C. & Robin, C. On the origin of hematopoietic stem cells: progress and controversy. *Stem cell research* **8**, 1-13, doi:10.1016/j.scr.2011.07.002 (2012).

42 Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harbor perspectives in medicine* **3**, a011601, doi:10.1101/cshperspect.a011601 (2013).

43 Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295-306, doi:10.1016/j.cell.2005.02.013 (2005).

44 Liu, K. & Nussenzweig, M. C. Origin and development of dendritic cells. *Immunological reviews* **234**, 45-54, doi:10.1111/j.0105-2896.2009.00879.x (2010).

45 Moore, A. J. & Anderson, M. K. Dendritic cell development: a choose-your-own-adventure story. *Advances in hematology* **2013**, 949513, doi:10.1155/2013/949513 (2013).

46 Hajdu, S. I. A note from history: The discovery of blood cells. *Annals of clinical and laboratory science* **33**, 237-238 (2003).

47 Schierbeek, A. *Measuring the invisible world; the life and works of Antoni van Leeuwenhoek*. (Abelard-Schuman, 1959).

48 Harris, J. W. & Kellermeyer, R. W. *The red cell; production, metabolism, destruction: normal and abnormal*. Rev. edn, (Published for the Commonwealth Fund by Harvard University Press, 1970).

49 Palis, J. Primitive and definitive erythropoiesis in mammals. *Frontiers in physiology* **5**, 3, doi:10.3389/fphys.2014.00003 (2014).

50 Keerthivasan, G., Wickrema, A. & Crispino, J. D. Erythroblast enucleation. *Stem cells international* **2011**, 139851, doi:10.4061/2011/139851 (2011).

51 Chasis, J. A. & Mohandas, N. Erythroblastic islands: niches for erythropoiesis. *Blood* **112**, 470-478, doi:10.1182/blood-2008-03-077883 (2008).

52 Spits, H. *et al.* Innate lymphoid cells--a proposal for uniform nomenclature. *Nature reviews. Immunology* **13**, 145-149, doi:10.1038/nri3365 (2013).

53 Constantinides, M. G., McDonald, B. D., Verhoef, P. A. & Bendelac, A. A committed precursor to innate lymphoid cells. *Nature* **508**, 397-401, doi:10.1038/nature13047 (2014).

54 Klose, C. S. *et al.* Differentiation of type 1 ILCs from a common progenitor to all helper-like innate lymphoid cell lineages. *Cell* **157**, 340-356, doi:10.1016/j.cell.2014.03.030 (2014).

55 Walker, J. A., Barlow, J. L. & McKenzie, A. N. Innate lymphoid cells--how did we miss them? *Nature reviews. Immunology* **13**, 75-87, doi:10.1038/nri3349 (2013).

56 Murphy, K., Travers, P., Walport, M. & Janeway, C. *Janeway's immunobiology*. 8th edn, (Garland Science, 2012).

57 Morbach, H., Eichhorn, E. M., Liese, J. G. & Girschick, H. J. Reference values for B cell subpopulations from infancy to adulthood. *Clinical and experimental immunology* **162**, 271-279, doi:10.1111/j.1365-2249.2010.04206.x (2010).

58 LeBien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112**, 1570-1580, doi:10.1182/blood-2008-02-078071 (2008).

59 Nossal, G. J. One cell-one antibody: prelude and aftermath. *Nature immunology* **8**, 1015-1017, doi:10.1038/ni1007-1015 (2007).

60 Hardy, R. R. & Hayakawa, K. B cell development pathways. *Annual review of immunology* **19**, 595-621, doi:10.1146/annurev.immunol.19.1.595 (2001).

61 Nutt, S. L. & Kee, B. L. The transcriptional regulation of B cell lineage commitment. *Immunity* **26**, 715-725, doi:10.1016/j.immuni.2007.05.010 (2007).

62 Jung, D., Giallourakis, C., Mostoslavsky, R. & Alt, F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology* **24**, 541-570, doi:10.1146/annurev.immunol.23.021704.115830 (2006).

63 Bossen, C., Mansson, R. & Murre, C. Chromatin topology and the regulation of antigen receptor assembly. *Annual review of immunology* **30**, 337-356, doi:10.1146/annurev-immunol-020711-075003 (2012).

64 Hendriks, R. W. & Middendorp, S. The pre-BCR checkpoint as a cell-autonomous proliferation switch. *Trends in immunology* **25**, 249-256, doi:10.1016/j.it.2004.02.011 (2004).

65 Herzog, S., Reth, M. & Jumaa, H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. *Nature reviews. Immunology* **9**, 195-205, doi:10.1038/nri2491 (2009).

66 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

67 Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human genetics* **122**, 565-581, doi:10.1007/s00439-007-0433-0 (2008).

68 Watson, J. D. *The double helix: a personal account of the discovery of the structure of DNA*. (Weidenfeld & Nicolson, 1968).

69 Woodcock, C. L. & Ghosh, R. P. Chromatin higher-order structure and dynamics. *Cold Spring Harbor perspectives in biology* **2**, a000596, doi:10.1101/cshperspect.a000596 (2010).

70 Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-294 (1999).

71 Hillier, L. W. *et al.* Genomics in C. elegans: so many genes, such a little worm. *Genome research* **15**, 1651-1660, doi:10.1101/gr.3729105 (2005).

72 Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94, doi:10.1016/j.cell.2014.03.008 (2014).

73 Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251, doi:10.1016/j.cell.2013.02.014 (2013).

74 Ramskold, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5**, e1000598, doi:10.1371/journal.pcbi.1000598 (2009).

75 Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226, doi:10.1016/j.cell.2008.09.050 (2008).

76 Bunn, H. F. Erythropoietin. *Cold Spring Harbor perspectives in medicine* **3**, a011619, doi:10.1101/cshperspect.a011619 (2013).

77 Kornberg, R. D. An autobiographic conversation with Roger D. Kornberg on his work on transcription regulation. *Cell death and differentiation* **14**, 1977-1980, doi:10.1038/sj.cdd.4402250 (2007).

78 Lee, T. I. & Young, R. A. Transcription of eukaryotic protein-coding genes. *Annual review of genetics* **34**, 77-137, doi:10.1146/annurev.genet.34.1.77 (2000).

79 Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature reviews. Genetics* **13**, 233-245, doi:10.1038/nrg3163 (2012).

80 Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and*

molecular biology **41**, 105-178, doi:10.1080/10409230600648736 (2006).
81 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).
82 Carlsten, J. O., Zhu, X. & Gustafsson, C. M. The multitalented Mediator complex. *Trends in biochemical sciences* **38**, 531-537, doi:10.1016/j.tibs.2013.08.007 (2013).
83 Malik, S. & Roeder, R. G. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature reviews. Genetics* **11**, 761-772, doi:10.1038/nrg2901 (2010).
84 Cheung, A. C. & Cramer, P. A movie of RNA polymerase II transcription. *Cell* **149**, 1431-1437, doi:10.1016/j.cell.2012.06.006 (2012).
85 Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Molecular cell* **36**, 541-546, doi:10.1016/j.molcel.2009.10.019 (2009).
86 Phatnani, H. P. & Greenleaf, A. L. Phosphorylation and functions of the RNA polymerase II CTD. *Genes & development* **20**, 2922-2936, doi:10.1101/gad.1477006 (2006).
87 Allison, L. A. *Fundamental molecular biology*. (Blackwell Pub., 2007).
88 Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics* **13**, 720-731, doi:10.1038/nrg3293 (2012).
89 Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & development* **26**, 2119-2137, doi:10.1101/gad.200303.112 (2012).
90 Ben-Ari, Y. *et al.* The life of an mRNA in space and time. *Journal of cell science* **123**, 1761-1774, doi:10.1242/jcs.062638 (2010).
91 Kuehner, J. N., Pearson, E. L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature reviews. Molecular cell biology* **12**, 283-294, doi:10.1038/nrm3098 (2011).
92 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705, doi:10.1016/j.cell.2007.02.005 (2007).
93 Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707-719, doi:10.1016/j.cell.2007.01.015 (2007).
94 Bell, O., Tiwari, V. K., Thoma, N. H. & Schubeler, D. Determinants and dynamics of genome accessibility. *Nature reviews. Genetics* **12**, 554-564, doi:10.1038/nrg3017 (2011).
95 Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: the industrial revolution of DNA around histones. *Nature reviews. Molecular cell biology* **7**, 437-447, doi:10.1038/nrm1945 (2006).
96 Hargreaves, D. C. & Crabtree, G. R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell research* **21**, 396-420, doi:10.1038/cr.2011.32 (2011).
97 Khare, S. P. *et al.* HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic acids research* **40**, D337-342, doi:10.1093/nar/gkr1125 (2012).
98 Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell research* **21**, 381-395, doi:10.1038/cr.2011.22 (2011).
99 Talbert, P. B. & Henikoff, S. Histone variants--ancient wrap artists of the epigenome. *Nature reviews. Molecular cell biology* **11**, 264-275, doi:10.1038/nrm2861 (2010).
100 Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41-45, doi:10.1038/47412 (2000).
101 Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics* **12**, 7-18, doi:10.1038/nrg2905 (2011).
102 Chi, P., Allis, C. D. & Wang, G. G. Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nature reviews. Cancer* **10**, 457-469, doi:10.1038/nrc2876 (2010).
103 Yun, M., Wu, J., Workman, J. L. & Li, B. Readers of histone modifications. *Cell research* **21**, 564-578, doi:10.1038/cr.2011.42 (2011).
104 Guibert, S. & Weber, M. Functions of DNA methylation and hydroxymethylation in mammalian development. *Current topics in developmental biology* **104**, 47-83, doi:10.1016/B978-0-12-416027-9.00002-4 (2013).
105 Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-485, doi:10.1038/35013100 (2000).
106 Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes & development* **23**, 781-783, doi:10.1101/gad.1787609 (2009).
107 Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* **13**, 613-626, doi:10.1038/nrg3207 (2012).
108 Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology* **3**, 318-356 (1961).
109 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics* **10**, 252-263, doi:10.1038/nrg2538 (2009).
110 Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147-151, doi:10.1038/nature01763 (2003).
111 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes & development* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).
112 Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes & development* **9**, 1250-1262 (1995).
113 Ferreira, R., Ohneda, K., Yamamoto, M. & Philipsen, S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Molecular and cellular biology* **25**, 1215-1227, doi:10.1128/MCB.25.4.1215-1227.2005 (2005).
114 Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes & development* **12**, 2403-2412 (1998).
115 Hodges, V. M., Rainey, S., Lappin, T. R. & Maxwell, A. P. Pathophysiology of anemia and erythrocytosis. *Critical reviews in oncology/hematology* **64**, 139-158, doi:10.1016/j.critrevonc.2007.06.006 (2007).
116 Aranda, A. & Pascual, A. Nuclear hormone receptors and gene expression. *Physiological reviews* **81**, 1269-1304 (2001).
117 Latchman, D. S. Transcription factors: an overview. *The international journal of biochemistry & cell biology* **29**, 1305-1312 (1997).
118 Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* **12**, 283-293, doi:10.1038/nrg2957 (2011).

119 de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506, doi:10.1038/nature12753 (2013).

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

121 Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell stem cell* **14**, 762-775, doi:10.1016/j.stem.2014.05.017 (2014).

122 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).

123 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 996-1001, doi:10.1073/pnas.1317788111 (2014).

124 Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).

125 Orom, U. A. & Shiekhattar, R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* **154**, 1190-1193, doi:10.1016/j.cell.2013.08.028 (2013).

126 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

127 Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729-740 (1983).

128 Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**, 717-728 (1983).

129 Ebert, A., Medvedovic, J., Tagoh, H., Schwickert, T. A. & Busslinger, M. Control of Antigen Receptor Diversity through Spatial Regulation of V(D)J Recombination. *Cold Spring Harbor symposia on quantitative biology*, doi:10.1101/sqb.2013.78.019943 (2014).

130 Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature reviews. Genetics* **7**, 703-713, doi:10.1038/nrg1925 (2006).

131 Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51**, 975-985 (1987).

132 Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077-3086, doi:10.1182/blood-2002-04-1104 (2002).

133 Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* **10**, 1453-1465 (2002).

134 Palstra, R. J. *et al.* The beta-globin nuclear compartment in development and erythroid differentiation. *Nature genetics* **35**, 190-194, doi:10.1038/ng1244 (2003).

135 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

136 Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17921-17926, doi:10.1073/pnas.1317023110 (2013).

137 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

138 Wang, H. *et al.* NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 705-710, doi:10.1073/pnas.1315023111 (2014).

139 Eddy, S. R. The C-value paradox, junk DNA and ENCODE. *Current biology : CB* **22**, R898-899, doi:10.1016/j.cub.2012.10.002 (2012).

140 Prasanth, K. V. & Spector, D. L. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes & development* **21**, 11-42, doi:10.1101/gad.1484207 (2007).

141 Chen, L., Tovar-Corona, J. M. & Urrutia, A. O. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *International journal of evolutionary biology* **2012**, 596274, doi:10.1155/2012/596274 (2012).

142 Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays : news and reviews in molecular, cellular and developmental biology* **29**, 288-299, doi:10.1002/bies.20544 (2007).

143 Messina, D. N., Glasscock, J., Gish, W. & Lovett, M. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome research* **14**, 2041-2047, doi:10.1101/gr.2584104 (2004).

144 Tsiftsoglou, A. S., Vizirianakis, I. S. & Strouboulis, J. Erythropoiesis: model systems, molecular regulators, and developmental programs. *IUBMB life* **61**, 800-830, doi:10.1002/iub.226 (2009).

145 deBoer, E., Antoniou, M., Mignotte, V., Wall, L. & Grosveld, F. The human beta-globin promoter; nuclear protein factors and erythroid specific induction of transcription. *The EMBO journal* **7**, 4203-4212 (1988).

146 Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Molecular cell* **36**, 667-681, doi:10.1016/j.molcel.2009.11.001 (2009).

147 Yu, M. *et al.* Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Molecular cell* **36**, 682-695, doi:10.1016/j.molcel.2009.11.002 (2009).

148 Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes & development* **24**, 277-289, doi:10.1101/gad.551810 (2010).

149 Cantor, A. B. & Orkin, S. H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**, 3368-3376, doi:10.1038/sj.onc.1205326 (2002).

150 Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *The EMBO journal* **16**, 3145-3157, doi:10.1093/emboj/16.11.3145 (1997).

151 Matthews, J. M. & Visvader, J. E. LIM-domain-binding protein 1: a multifunctional cofactor that interacts with diverse proteins. *EMBO reports* **4**, 1132-1137, doi:10.1038/sj.embor.7400030 (2003).

152 Love, P. E., Warzecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends in genetics*

: TIG **30**, 1-9, doi:10.1016/j.tig.2013.10.001 (2014).
153 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).
154 Laslo, P., Pongubala, J. M., Lancki, D. W. & Singh, H. Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Seminars in immunology* **20**, 228-235, doi:10.1016/j.smim.2008.08.003 (2008).
155 Murre, C. Developmental trajectories in early hematopoiesis. *Genes & development* **23**, 2366-2370, doi:10.1101/gad.1861709 (2009).
156 Nutt, S. L., Heavey, B., Rolink, A. G. & Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**, 556-562, doi:10.1038/44076 (1999).
157 Ma, S., Pathak, S., Trinh, L. & Lu, R. Interferon regulatory factors 4 and 8 induce the expression of Ikaros and Aiolos to down-regulate pre-B-cell receptor and promote cell-cycle withdrawal in pre-B-cell development. *Blood* **111**, 1396-1403, doi:10.1182/blood-2007-08-110106 (2008).
158 Herold, M., Bartkuhn, M. & Renkawitz, R. CTCF: insights into insulator function during development. *Development* **139**, 1045-1057, doi:10.1242/dev.065268 (2012).
159 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics* **15**, 234-246, doi:10.1038/nrg3663 (2014).
160 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).
161 Visser, M., Kayser, M. & Palstra, R. J. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome research* **22**, 446-455, doi:10.1101/gr.128652.111 (2012).
162 Look, A. T. Oncogenic transcription factors in the human acute leukemias. *Science* **278**, 1059-1064 (1997).
163 Rowley, J. D. Chromosome translocations: dangerous liaisons revisited. *Nature reviews. Cancer* **1**, 245-250, doi:10.1038/35106108 (2001).
164 Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. *Nature structural & molecular biology* **21**, 210-219, doi:10.1038/nsmb.2784 (2014).
165 Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature genetics* **23**, 185-188, doi:10.1038/13810 (1999).
166 Gardini, A. *et al.* AML1/ETO oncoprotein is directed to AML1 binding regions and co-localizes with AML1 and HEB on its targets. *PLoS genetics* **4**, e1000275, doi:10.1371/journal.pgen.1000275 (2008).
167 Hecht, J. L. & Aster, J. C. Molecular biology of Burkitt's lymphoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **18**, 3707-3721 (2000).
168 Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* **306**, 662-666 (1983).
169 Cao, A. & Galanello, R. Beta-thalassemia. *Genetics in medicine : official journal of the American College of Medical Genetics* **12**, 61-76, doi:10.1097/GIM.0b013e3181cd68ed (2010).
170 Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature reviews. Genetics* **10**, 241-251, doi:10.1038/nrg2554 (2009).
171 Barbujani, G., Ghirotto, S. & Tassi, F. Nine things to remember about human genome diversity. *Tissue antigens* **82**, 155-164, doi:10.1111/tan.12165 (2013).
172 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
173 Sturm, R. A. *et al.* A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American journal of human genetics* **82**, 424-431, doi:10.1016/j.ajhg.2007.11.005 (2008).
174 Andrews, N. C. Genes determining blood cell traits. *Nature genetics* **41**, 1161-1162, doi:10.1038/ng1109-1161 (2009).
175 Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS computational biology* **8**, e1002822, doi:10.1371/journal.pcbi.1002822 (2012).
176 Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer research* **73**, 4180-4184, doi:10.1158/0008-5472.CAN-13-0789 (2013).
177 Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7-24, doi:10.1016/j.ajhg.2011.11.029 (2012).
178 Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988, doi:10.1038/ng2085 (2007).
179 Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature genetics* **41**, 885-890, doi:10.1038/ng.406 (2009).
180 Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics* **41**, 882-884, doi:10.1038/ng.403 (2009).
181 Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Molecular and cellular biology* **30**, 1411-1420, doi:10.1128/MCB.01384-09 (2010).
182 Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360-1363, doi:10.1126/science.1228606 (2012).
183 Thein, S. L., Menzel, S., Lathrop, M. & Garner, C. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. *Human molecular genetics* **18**, R216-223, doi:10.1093/hmg/ddp401 (2009).
184 Olivieri, N. F. & Weatherall, D. J. The therapeutic reactivation of fetal haemoglobin. *Human molecular genetics* **7**, 1655-1658 (1998).
185 Garner, C. *et al.* Genetic influences on F cells and other hematologic variables: a twin heritability study. *Blood* **95**, 342-346 (2000).
186 Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature genetics* **42**, 1049-1051, doi:10.1038/ng.707 (2010).
187 Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin.

Nature genetics **42**, 801-805, doi:10.1038/ng.630 (2010).
188 Liu, D. *et al.* Erythroid Kruppel-like factor mutations are relatively more common in a thalassemia endemic region and ameliorate the clinical and hematological severity of beta-thalassemia. *Blood*, doi:10.1182/blood-2014-03-561779 (2014).
189 Sankaran, V. G. *et al.* Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science* **322**, 1839-1842, doi:10.1126/science.1165409 (2008).
190 Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-257, doi:10.1126/science.1242088 (2013).
191 Gaj, T., Gersbach, C. A. & Barbas, C. F., 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology* **31**, 397-405, doi:10.1016/j.tibtech.2013.04.004 (2013).
192 Marx, V. A blooming genomic desert. *Nature methods* **11**, 135-138, doi:10.1038/nmeth.2817 (2014).
193 Noordermeer, D. & de Laat, W. Joining the loops: beta-globin gene regulation. *IUBMB life* **60**, 824-833, doi:10.1002/iub.129 (2008).

Chapter 2

Control of developmentally poised erythroid genes by combinatorial corepressor actions

Ralph Stadhouders[1]†, Supat Thongjuea[2,3], Petros Kolovos[1], H. Irem Baymaz[1], Xiao Yu[1], Jeroen Demmers[4], Karel Bezstarosti[4], Alex Maas[1], Christel Kockx[5], Zeliha Ozgur[5], Wilfred van IJcken[5], Marie-Laure Arcangeli[6], Charlotte Andrieu-Soler[1,7], Boris Lenhard[2,8], Frank Grosveld[1,9,11] & Eric Soler[1,9,10,11]†

[1]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands.
[2]Computational Biology Unit, Bergen Center for Computational Science, Bergen, Norway.
[3]MRC Molecular Haematology Unit, WIMM, Oxford, United Kingdom .
[4]Department of Proteomics, Erasmus Medical Center, Rotterdam, The Netherlands.
[5]Center for Biomics, Erasmus Medical Center, Rotterdam, The Netherlands.
[6]Laboratory of Hematopoietic and Leukemic Stem cells, CEA/INSERM U967, France.
[7]Institut National pour la Santé Et la Recherche Médicale (INSERM) U872, Paris, France.
[8]Department of Molecular Sciences, Imperial College London, and MRC Clinical Sciences Centre, London, United Kingdom.
[9]Cancer Genomics Center, Erasmus Medical Center, Rotterdam, The Netherlands.
[10]Laboratory of Molecular Hematopoiesis, CEA/INSERM U967, Fontenay-aux-Roses, France.

[11]These authors contributed equally.
†Corresponding authors.

Submitted

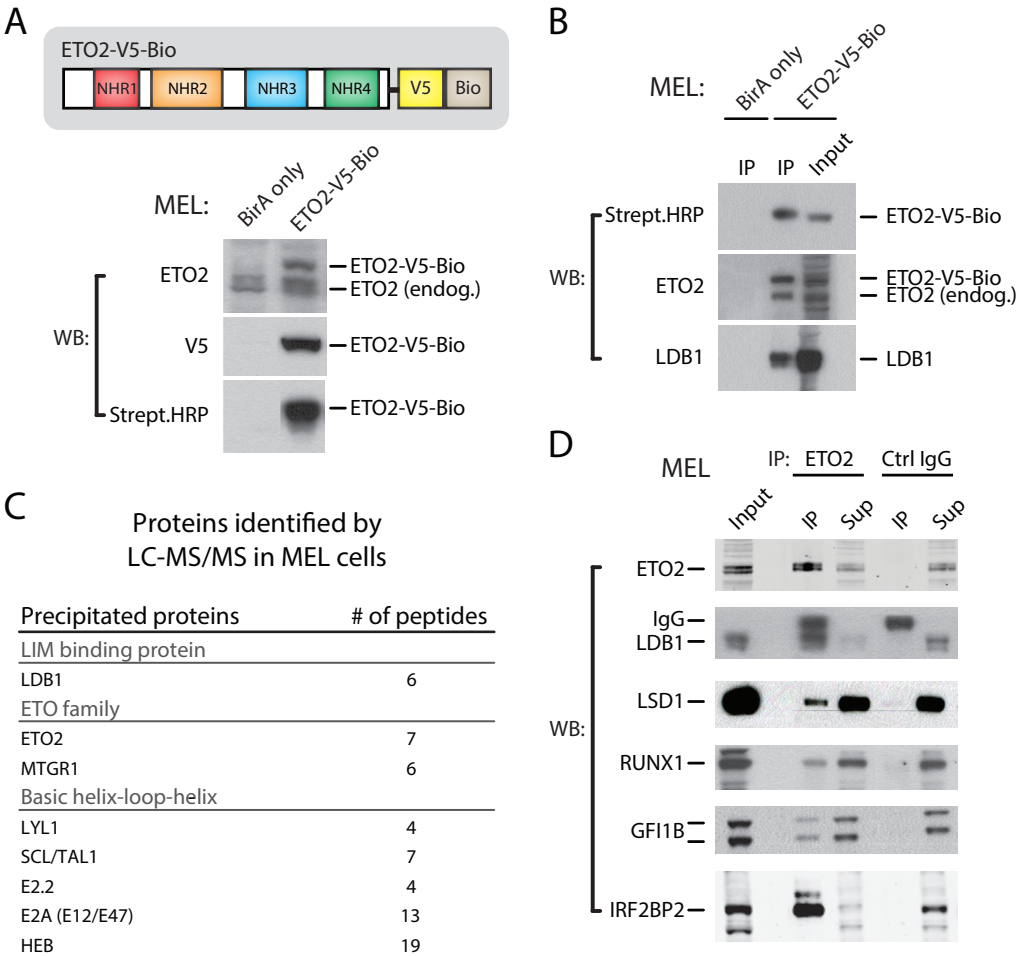
Abstract

How transcription factors (TFs) cooperate within large multimeric complexes to allow rapid modulation of gene expression during differentiation is still largely unknown. Here we investigate the actions of a large activating TF complex nucleated by the hematopoietic master regulators LDB1, GATA1, TAL1, LMO2 and ETO2 during erythroid differentiation. This 'LDB1-complex' already binds to the regulatory elements of its target genes prior to their activation; ETO2 is thought to prevent premature activation in progenitors. How ETO2 establishes this poised state is unclear. Using a combination of proteomics and functional genomics we identified corepressor proteins that cooperate with ETO2 to repress LDB1-complex target genes in erythroid progenitors. The IRF2BP2 corepressor strongly enhances ETO2-mediated gene repression, likely through recruitment of the NCOR1/SMRT corepressor complex. The ETO2-IRF2BP2 axis suppresses the expression of the vast majority of archetypical erythroid genes and pathways until its decommissioning at the onset of terminal differentiation. A novel mouse model confirmed the importance of IRF2BP2 for erythropoiesis *in vivo*. Thus, a collaborative action of multiple corepressor proteins within the LDB1-complex maintains the late erythroid transcriptome poised for rapid activation by its activating members. Our experiments show that multimeric regulatory complexes feature a dynamic interplay between activating and repressing components that determines lineage-specific gene expression.

Introduction

Hematopoietic development relies on the stepwise activation and repression of lineage-specific gene expression programs. This process is regulated by sets of conserved transcription factors (TFs) acting in a combinatorial and/or antagonistic fashion to establish cellular identity through tight control of gene regulatory networks (Orkin and Zon 2008). Exactly how TFs and the cofactors they recruit cooperate within large protein complexes to rapidly modulate gene expression during differentiation is still not completely understood. We set out to address this issue using a well-characterized erythroid differentiation system driven by a multimeric TF complex nucleated by the hematopoietic master regulators LDB1, GATA1, TAL1, LMO2 and ETO2 (hereafter referred to as the LDB1-complex). The LDB1-complex plays a pivotal role in promoting differentiation of the erythroid and megakaryocytic lineages (Szalai et al. 2006; Love et al. 2014). It was previously shown to bind the regulatory regions of developmentally poised erythroid genes, which are rapidly induced upon terminal erythroid differentiation (Schuh et al. 2005; Goardon et al. 2006; Meier et al. 2006; Soler et al. 2010; Li et al. 2013). Early activation of these poised erythroid genes in immature progenitors is prevented by the LDB1-complex member ETO2 (also referred to as MTG16), a transcriptional corepressor (Schuh et al. 2005; Goardon et al. 2006; Meier et al. 2006; Soler et al. 2010; Kiefer et al. 2011). ETO2 belongs to a family of transcriptional repressors known as the ETO family, which further consists of the founder member ETO (or MTG8) and the MTGR1 proteins (Davis et al. 2003). ETO2 plays key roles in the maintenance of hematopoietic stem cells (Fischer et al. 2012), the development of the lymphoid system (Hunt et al. 2011) and regulating effective (stress) erythropoiesis (Chyla et al. 2008). The importance of a functional ETO2 protein in maintaining hematopoietic homeostasis is further underlined by its causal involvement in acute myeloid leukemia (AML) (Gamou et al. 1998). Whereas ETO2 is well known for its repressor function in several cell types (Schuh et al. 2005; Barrett et al. 2012; Kumar et al. 2013), the molecular mechanisms of erythroid gene suppression in the context of the LDB1-complex remain largely unknown. Unraveling these mechanisms is important to provide novel insight into how TFs and cofactors within a multimeric complex impose a 'poised' status onto their target genes.

To begin addressing these questions, we performed a proteomics screen for novel ETO2 binding partners. This screen identified the IRF2BP2, GFI1B and LSD1 transcriptional repressors as ETO2 interacting proteins. We show here that IRF2BP2 is a novel component of the LDB1 complex able to strongly enhance ETO2-mediated transcriptional repression. ChIP-Sequencing (ChIP-Seq) analysis and loss-of-function studies revealed that ETO2 and IRF2BP2 chromatin occupancy significantly overlap at a genome-wide scale, and that both factors regulate a common set of key erythroid target genes and regulatory pathways. Subsequent analysis of IRF2BP2 protein partners showed that IRF2BP2 is able to recruit the well-known NCOR1 corepressor, which is shown here to bind ETO2/IRF2BP2 erythroid target genes to potentially mediate their repression. We finally confirmed the *in vivo* relevance of the newly identified IRF2BP2 corepressor by using an IRF2BP2-deficient mouse model. Animals homozygous for the genetrap *Irf2bp2* allele display an ineffective fetal liver erythropoiesis during gestation and die around birth. Thus, our data reveal a complex collaborative action of multiple corepressor proteins within the LDB1-complex at the erythroid progenitor stage. As a result,



C

Proteins identified by LC-MS/MS in MEL cells

Precipitated proteins	# of peptides
LIM binding protein	
LDB1	6
ETO family	
ETO2	7
MTGR1	6
Basic helix-loop-helix	
LYL1	4
SCL/TAL1	7
E2.2	4
E2A (E12/E47)	13
HEB	19
SSBP family	
SSBP2	5
SSBP3	4
SSBP4	3
Transcriptional activator	
FUS/TLS	2
Repressor / co-repressor	
IRF2BP2	6
GFI1B	2
Runt related	
RUNX1	4
CBF-beta	1
BHC repressor complex	
LSD1	13
GSE1	9
REST co-repressor 3	4
HMG20a	2
HDAC2	3
Jumonji-domain demethylase 1C/2C	1

Figure 1. Identification of ETO2 binding partners in erythroid progenitor cells. (A) Schematic of the ETO2 protein, its 4 Nervy homology regions (NHR1-4) and the C-terminal V5-Bio tag (Top). Fusion protein expression and proper tag function in MEL cells were validated by Western Blot (WB) analysis. MEL cells expressing only the BirA enzyme were used as a control. (B) Efficient streptavidin immunoprecipitation (IP) of ETO2-V5-Bio in MEL cells. Interaction of ETO2-V5-Bio with LDB1 (a known binding partner) was used for validation. (C) ETO2-V5-Bio interacting proteins identified by LC-MS/MS in MEL cells. Only proteins pulled down in 2 independent experiments and with low background scores are shown. (D) Co-IP validations of selected ETO2-V5-Bio interacting proteins in MEL cells using an endogenous ETO2 antibody. Species-matched IgG was used to control for aspecific binding. Strept-HRP, streptavidin-HRP; Sup, supernatant; endog., endogenous

late erythroid-specific genes are maintained in a poised state prior to their rapid activation upon terminal differentiation.

Results

Identification of ETO2 protein partners in erythroid cells

We first employed a proteomics approach to characterize the molecular determinants of ETO2's repressive activity. An epitope-tagged form of ETO2 (ETO2-V5-Bio) was expressed in the MEL erythroid progenitor cell line (Soler et al. 2010) and used in single-step protein complex capture experiments (de Boer et al. 2003; Soler et al. 2011). The affinity tag contains a Bio peptide sequence that is efficiently biotinylated by the bacterial BirA enzyme, resulting in the biotinylation of ETO2-V5-Bio (Fig.1A). The C-terminal tag fused to ETO2 did not interfere with its functions since ETO2-V5-Bio shows (i) proper intracellular localization (Supplemental Fig.1A), (ii) the ability to interact with endogenous ETO2 (El Omari et al. 2013) (Fig.1B), (iii) interaction with its known binding partner LDB1 (Meier et al. 2006) (Fig.1B), and (iv) binding to known genomic target sites (Soler et al. 2010) (Supplemental Fig.1C). These results demonstrate that tag addition does not affect ETO2 in its ability to form complexes in erythroid cells. A streptavidin pull-down was carried out and co-purified proteins were identified by mass spectrometry (LC-MS/MS) (Fig.1C). In addition to known components of the LDB1 complex (e.g. TAL1, E2A, HEB, SSBP2/3/4) (Meier et al. 2006), additional interactions with the LSD1/Co-REST repressor complex, the hematopoietic transcription factor GFI1B, and the transcriptional repressor IRF2BP2 (Interferon regulatory factor 2-binding protein 2) were also detected. Endogenous interaction of ETO2 with these factors was confirmed by co-immunoprecipitation experiments in MEL cells (Fig.1D and Supplemental Fig.1B). Whereas the ability of ETO2 to interact with GFI1B was reported previously (Schuh et al. 2005), and the LSD1 complex was found to be associated with the LDB1 complex (including ETO2) in erythroid cells (Hu et al. 2009), the involvement of IRF2BP2 in these complexes has not been reported yet. We therefore set out to investigate this interaction in more detail.

ETO2 interacts with IRF2BP2 via a unique N-terminal domain

IRF2BP2 is a highly conserved Zinc-finger/RING-finger protein belonging to a family of three evolutionary conserved factors (IRF2BP1, IRF2BP2 and IRF2BPL) sharing high sequence homology. IRF2BP1 and IRF2BP2 were originally identified as interacting partners of IRF2, mediating its ability to repress *in vitro* reporter expression (Childs and Goodbourn 2003). Recently, other studies reported a repressive role for IRF2BP2 in complex with NFAT1 (Carneiro et al. 2011), p53 (Koeppel et al. 2009) or EAP1 (Yeung et al. 2011). In order to map the domains mediating the interaction between ETO2 and IRF2BP2, a series of deletion mutants was generated and used in co-immunoprecipitations experiments. ETO2 contains four highly conserved domains (NHR1-4) shared with the other members of the ETO family (ETO and MTGR1) (Davis et al. 2003), and two unique sequences at its N-terminus that are not shared with ETO and MTGR1, which we termed US1 and US2 (for Unique Sequence 1 and 2) (Fig.2A). As shown in Fig.2B, ETO2 interacts with IRF2BP2 via its US2 domain, suggesting that ETO2 is the only protein from the ETO family able to bind IRF2BP2. Using a similar strategy, we found that the IRF2BP2 RING finger domain mediates the interaction with ETO2 (Fig.2C). RING finger domains are characteristic of E3 ubiquitin ligases catalyzing the ubiquitination of target proteins, which often leads to protein degradation (Lipkowitz and Weissman 2011). Since ETO2 interacts with the RING finger domain of IRF2BP2, we tested whether ETO2 stability could be affected by this interaction. Increasing amounts of IRF2BP2 were co-expressed together with ETO2 in HEK 293T cells and ETO2 protein levels were monitored by Western blot analysis. As shown in Supplemental Figure 2, even when expressed in large excess, IRF2BP2 does not significantly affect ETO2 protein levels under these conditions.

IRF2BP2 enhances ETO2-mediated transcriptional repression

The functional role of the ETO2-IRF2BP2 interaction was first investigated *in vitro* using luciferase reporter assays. ETO2 was fused to a Gal4 DNA binding domain and co-expressed in HEK 293T cells together with a luciferase reporter plasmid containing Gal4 responsive elements. As previously reported, ETO2 induces a 20 to 30 fold repression of luciferase activity (Fig.2D) (Amann et al. 2001). Co-expression of IRF2BP2 further increased ETO2-mediated transcriptional repression, in a dose-dependent manner. This effect was not seen when using a RING finger deletion mutant of IRF2BP2 (IRF2BP2deltaRING), which is unable to interact with ETO2. Importantly, the ETO2 interacting partner LSD1 (Fig.1C and D), a known transcriptional repressor, did not significantly enhance ETO2-mediated repression (Fig.2D).

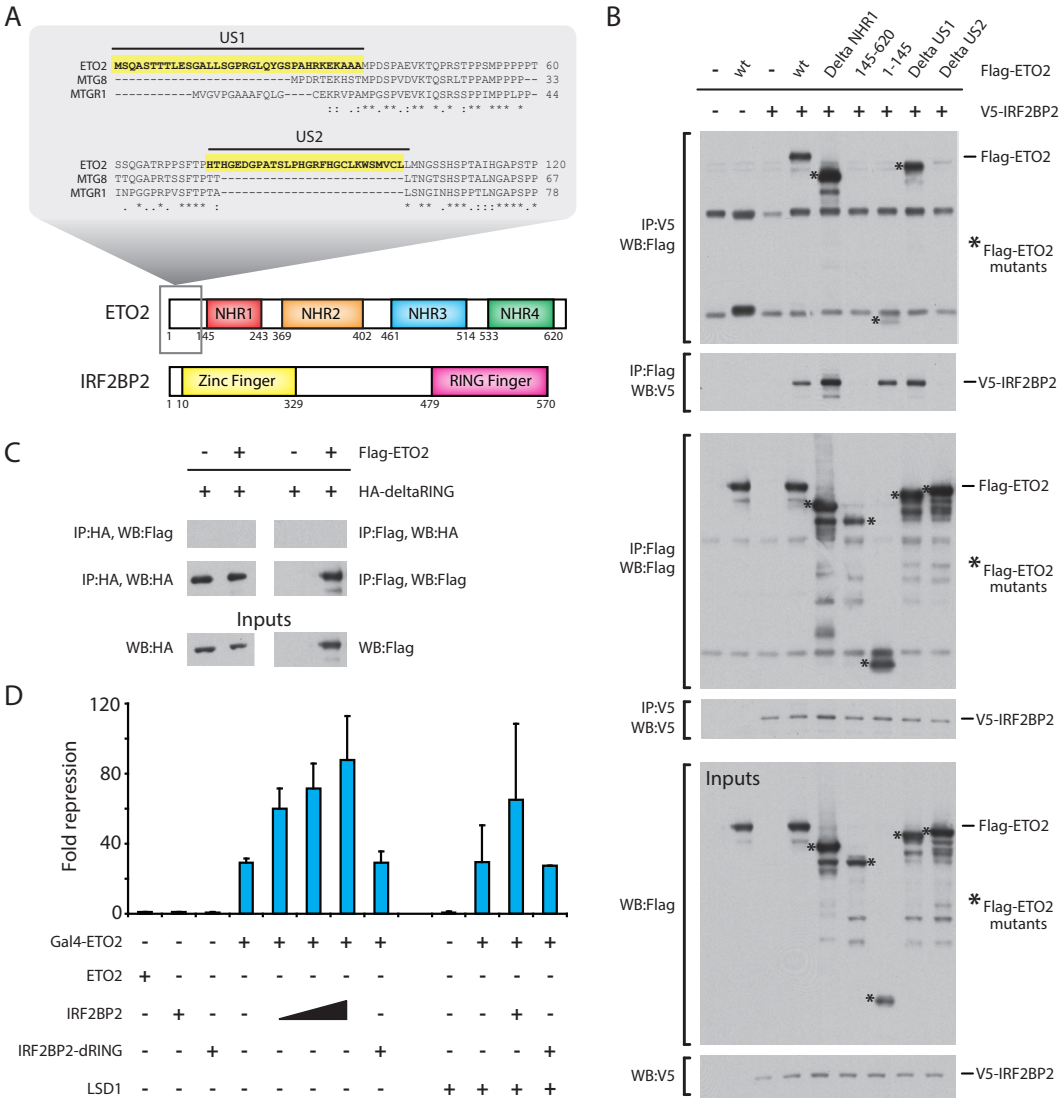


Figure 2. ETO2 and IRF2BP2 interact via their US2 and RING domains respectively to cooperatively repress reporter gene activity. (A) Schematic of the ETO2 and IRF2BP2 proteins and known functional domains. First and last amino acid positions of known functional domains are indicated by numbers. Highlighted are two unique N-terminal amino acid sequences (US1 and US2) only present in ETO2. (B) ETO2 interaction domain mapping using a collection of Flag-tagged deletion mutants that were overexpressed in HEK 293T cells together with V5-IRF2BP2. Bands representing the ETO2 mutant proteins are marked by an asterisk. (C) An HA-tagged IRF2BP2 lacking the C-terminal RING finger domain (HA-deltaRING) was used in co-IP experiments with Flag-ETO2. (D) Luciferase reporter assay to test repression of a Gal4-responsive promoter (coupled to a firefly luciferase gene) by ETO2 and its interacting partners IRF2BP2 and LSD1. Fusion to a Gal4 DNA binding domain (Gal4-ETO2) was used to target ETO2 to the promoter. Different combinations of Gal-ETO2 and IRF2BP2, deltaRING and LSD1 were co-transfected and firefly luciferase expression was measured after 48h. Co-transfection with equal amounts of a Renilla luciferase expression plasmid was used for normalization. Bars represent average values of at least three independent transfection experiments; error bars denote s.d. WB, Western Blot; IP, immunoprecipitation

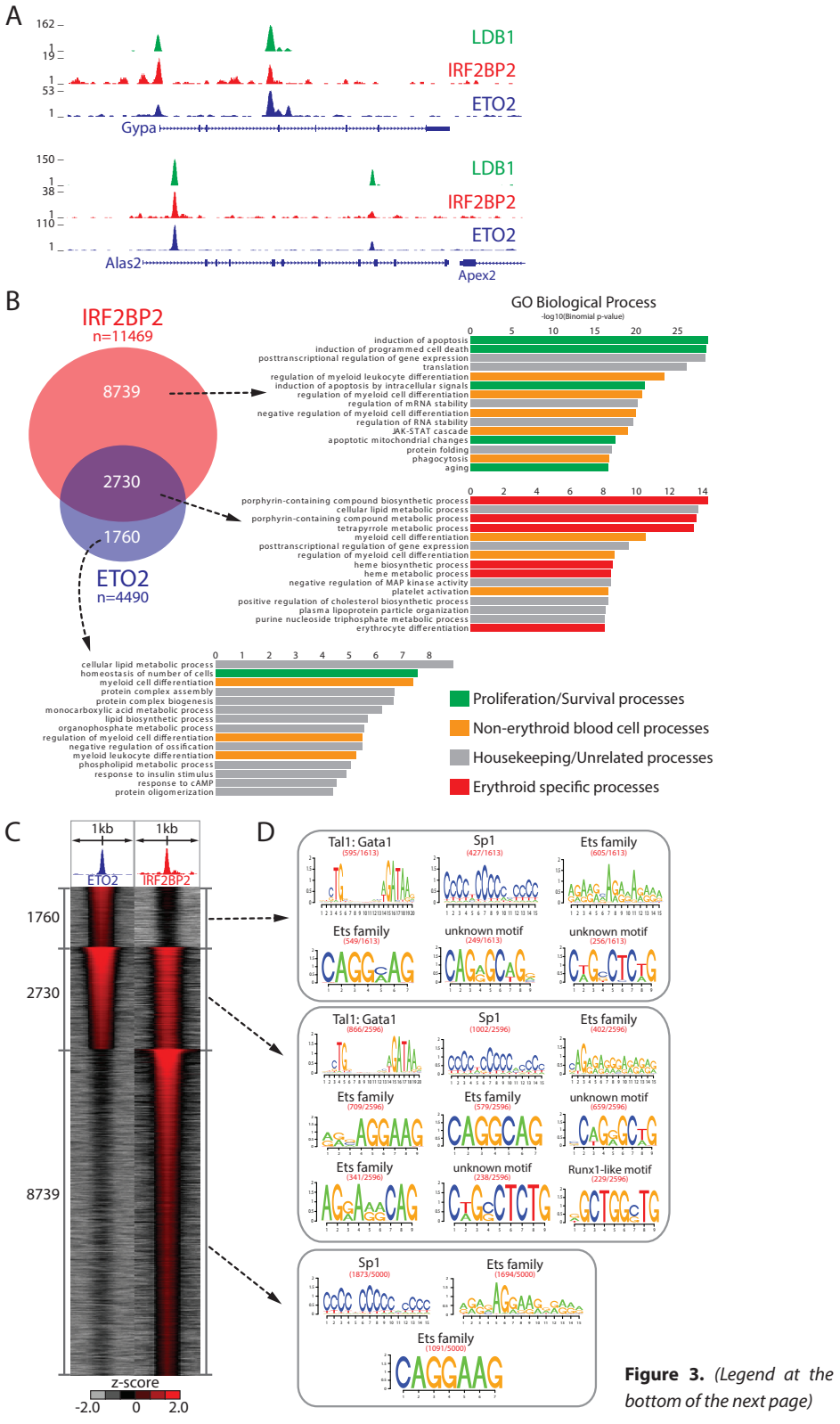


Figure 3. *(Legend at the bottom of the next page)*

Genome-wide analysis of ETO2 and IRF2BP2 chromatin occupancy reveals overlapping binding patterns on erythroid genes

We next performed ChIP-Sequencing (ChIP-Seq) experiments to determine whether IRF2BP2 is enriched at critical regulatory sites occupied by ETO2. IRF2BP2 binding sites were found at numerous *cis*-regulatory regions of erythroid genes controlled by ETO2 and LDB1. In particular, IRF2BP2 and ETO2 show co-occupancy on the *Gypa*, *Slc22a4*, *Epb4.2*, *Alas2* and *Slc4a1* genes, as well as the α- and β-globin clusters (see Fig.3A for examples). These genes are critical markers of mature erythroid cells, but are not yet expressed (or expressed at low levels) in erythroid progenitors such as MEL cells (Soler et al. 2010; Li et al. 2013). This suggests that ETO2 and IRF2BP2 might be involved in maintaining these erythroid genes in a 'poised' state, as the LDB1-complex is required for their rapid activation upon terminal differentiation (Li et al. 2013; Love et al. 2014). A genome-wide comparison of ETO2 and IRF2BP2 binding patterns revealed that 61% of ETO2 binding sites are also occupied by IRF2BP2 (Fig.3B). However many genomic locations are bound by IRF2BP2 in the absence of ETO2 (e.g. *Jund*, *Tnf*) and *vice versa* (e.g. *Gpr64*, *Rbm51*) (Fig.3C) indicating that both proteins are also involved in different regulatory complexes. Interestingly, the LSD1 and GFI1B repressors were also found enriched at ETO2/IRF2BP2 binding sites, overlapping with the positioning of the LDB1 complex (Supplemental Fig.3).

We tried to substantiate these observations on ETO2 and IRF2BP2 chromatin (co-)occupancy by performing a GO term analysis on putative target genes assigned to the different binding site subsets (using GREAT (McLean et al. 2010), see Methods for a detailed description). This confirmed a strong enrichment for erythroid functions among the common target genes (Fig.3B). ETO2-specific target genes showed some enrichment for common blood cell related functions, as well as for several housekeeping processes. Intriguingly, IRF2BP2-specific target genes showed strong associations with biological processes and functions involved in survival, apoptosis and cancer (Fig.3B).

A *de novo* DNA motif search performed on ETO2 and IRF2BP2 occupied genomic binding sites revealed enrichment of several different TF binding motifs. Both ETO2-only and ETO2-IRF2BP2 shared sites are enriched for SP1 and ETS motifs, as well as two unknown motifs (Fig.3D). In addition, the typical LDB1 complex signature represented by a composite E-box/GATA motif (CTGN(6-8)WGATAR) (Kassouf et al. 2010; Soler et al. 2010) was also found. Interestingly, this motif is completely absent from the IRF2BP2-only binding sites, and no enrichment for GATA motifs was observed. This suggests that in MEL cells, the IRF2BP2-only binding sites are GATA1-independent. Instead, mainly SP1 and ETS motifs are associated with IRF2BP2-only binding sites (Fig.3D).

Expression pattern and transcriptional regulation of Irf2bp2 during erythroid differentiation

It is well established that ETO2 expression levels diminish as erythroid progenitors undergo terminal differentiation (Schuh et al. 2005; Goardon et al. 2006; Meier et al. 2006). Furthermore, in a G1E-ER model system of erythroid differentiation, expression of *Cbfa2t3* (encoding ETO2) was repressed upon GATA1-driven erythroid maturation (Welch et al. 2004; Fujiwara et al. 2009). These and other observations (Fujiwara et al. 2009) suggest that *Cbfa2t3* expression is regulated by the ETO2-containing LDB1-complex, which involves an ETO2-negative auto-regulatory loop. To gain more insight into the regulation of *Irf2bp2* during erythropoiesis, we examined its expression levels during mouse fetal liver (FL) erythropoiesis. RNA-

Figure 3. ETO2-IRF2BP2 genomic co-occupancy is associated with genes involved in key erythroid processes. (A) Selected examples of overlapping ChIP-Seq profiles for LDB1, IRF2BP2 and ETO2 in MEL cells on key erythroid gene loci. (B) Venn diagram showing the genome-wide overlap between ETO2 and IRF2BP2 binding sites in MEL cells. GREAT analysis (McLean et al. 2010) (see Methods for more details) was performed for each group of binding sites (ETO2 only, co-occupied and IRF2BP2 only) to identify their putative target genes and possible significantly associated Gene Ontology (GO) terms. The top 15 GO terms is shown for each group of binding sites, and individual GO terms were categorized into four classes (erythroid-related, non-erythroid blood-related, proliferation/survival-related and housekeeping/unrelated). (C) Heatmap visualization of ETO2 and IRF2BP2 ChIP-Seq data, depicting all significant binding events centered on the peak region within a 1 kb window around the peak (binding sites were ranked by intensity). (D) A motif analysis (see Methods for more details) on the three groups of binding sites (ETO2 only, co-occupied and IRF2BP2 only) was performed to identify possible overrepresented transcription factor binding motifs within the peak sequences. Red numbers denote [number of motifs]/[total number of binding sites].

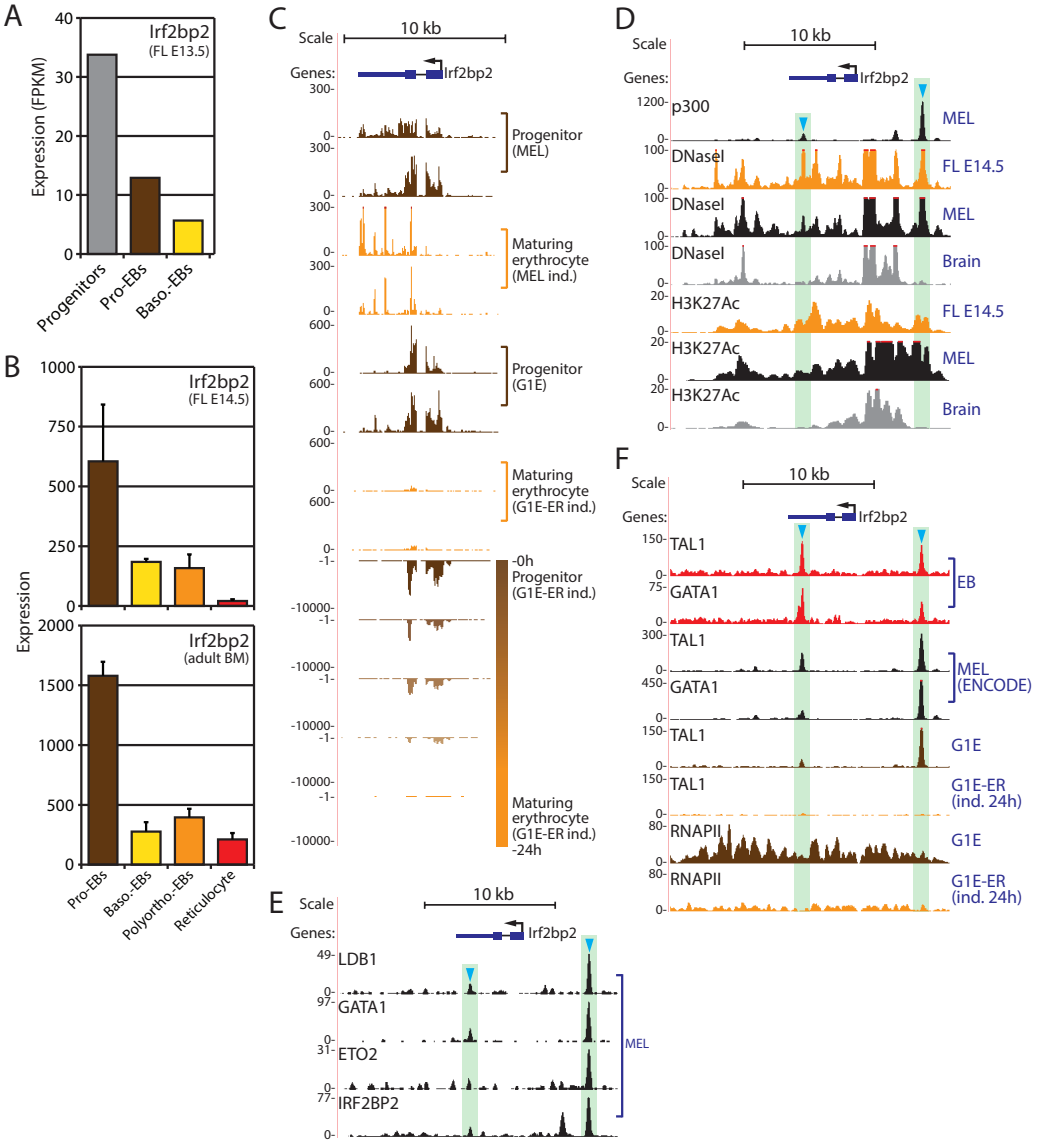


Figure 4. *Irf2bp2* gene expression and transcriptional regulation during erythroid development. (A) *Irf2bp2* expression levels at different stages of erythroid development ('Progenitors', CD71^{-}/Ter119^{-}; 'Pro-EBs', CD71^{+}/Ter119^{-}; 'Baso.-EBs, CD71^{+}/Ter119^{+}) as determined by RNA-Seq analysis of sorted E13.5 fetal liver (FL) cells. (B) *Irf2bp2* gene expression values at different stages of erythroid development in E14.5 FL and adult bone marrow (BM). Data were obtained from the online ErythronDB database (Kingsley et al. 2013). (C-F) Genome-wide datasets centered on the *Irf2bp2* locus from MEL, G1E(-ER), E14.5 FL, erythroblast (EB) and whole brain cells. Panel C shows RNA-Seq data from (differentiating) erythroid progenitors. Panel D depicts ChIP-Seq (p300 and H3K27Ac, both associated with enhancer activity) and DNase I-Seq (denotes regions of open chromatin) tracks; note the presence of two erythroid-specific putative enhancer elements (blue arrowheads). Panel E shows LDB1-complex (including IRF2BP2) occupancy of these putative enhancer elements in MEL cells. Panel F depicts GATA1, TAL1 and RNAPII binding to the *Irf2bp2* locus in erythroid cells. Note the loss of TAL1 binding to the putative enhancer elements in differentiating G1E-ER cells, accompanied by a loss of RNAPII enrichments. Data shown in panels C, D and F were obtained from the ENCODE consortium (Consortium et al. 2012) (see Methods for details on data access). Pro-EBs, pro-erythroblasts; Baso.-EBs, basophilic erythroblasts; Polyortho.-EBs, polyorthochromatic erythroblasts; RNAPII, RNA polymerase II

Sequencing (RNA-Seq) analysis of FACS-sorted populations of developing erythroid cells indicated that *Irf2bp2* expression is reduced upon differentiation (Fig.4A). A similar trend was observed by others using various *in vivo* and *in vitro* model systems for erythroid development (Consortium et al. 2012; Kingsley et al. 2013) (Fig.4B-C). As was reported for *Cbfa2t3*, *Irf2bp2* expression was lost upon GATA1-driven erythroid maturation in a G1E-ER model system (Fig.4C). Additionally, using genome-wide datasets previously generated by our laboratory (Soler et al. 2010) and the ENCODE consortium (Consortium et al. 2012), we identified two putative enhancer elements within the Irf2bp2 locus bound by the ETO2/IRF2BP2-containing

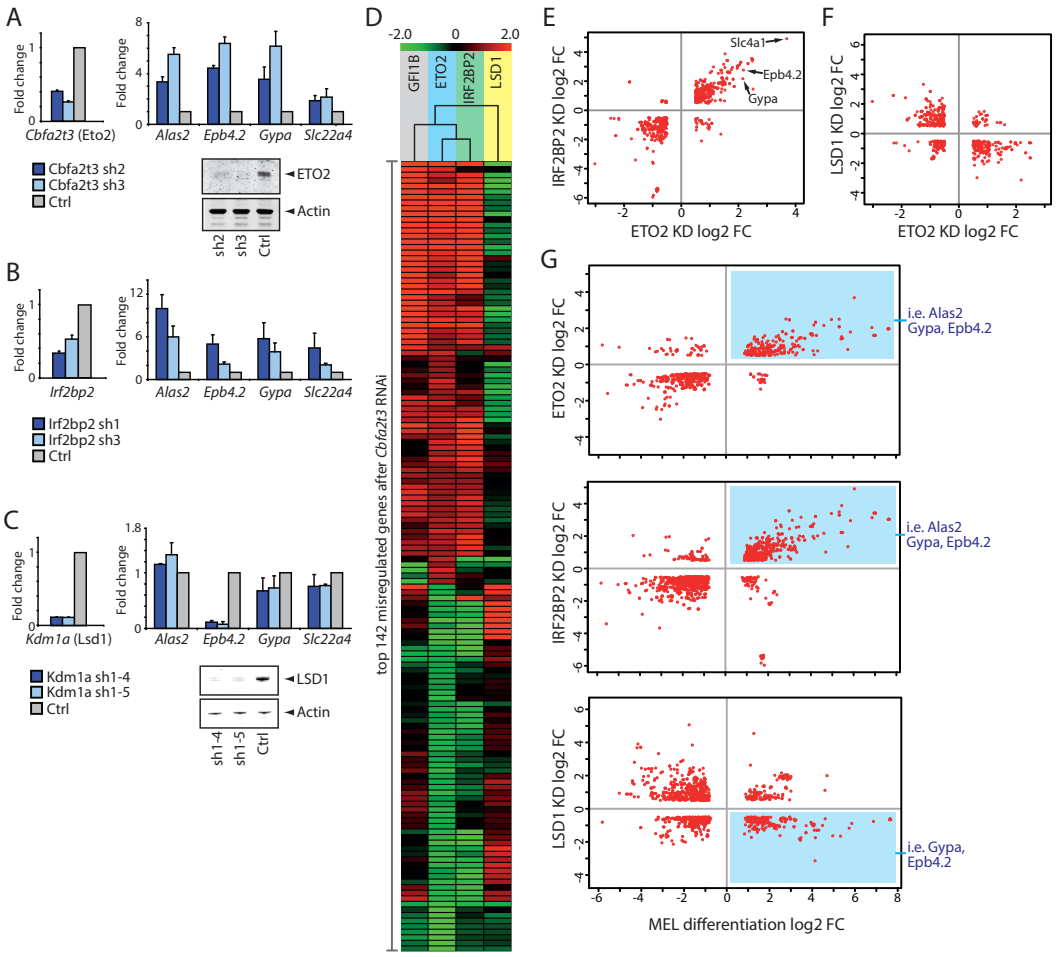


Figure 5. Genome-wide analysis of gene expression changes shows that ETO2 and IRF2BP2, but not LSD1, repress the late erythroid transcriptome. Lentiviral delivery of shRNAs against *Cbfa2t3* (A), *Irf2bp2* (B) and *Kdm1a* (C) mRNA to deplete MEL cells of the ETO2, LSD1 and IRF2BP2 proteins respectively. A non-targeting shRNA ('Ctrl') was used as a control. After 72 hours, mRNA levels were measured by qPCR (normalized versus *Rnh1* levels); protein levels (for ETO2 and LSD1) by Western Blot analysis (actin was used as a loading control). Expression levels of four archetypical late erythroid genes (*Alas2*, *Epb4.2*, *Gypa* and *Slc22a4*) were quantified by qPCR. (D) Unsupervised clustering of the top 142 misregulated genes after *Cbfa2t3* (ETO2) knockdown and the expression changes of the same set of genes induced after GFI1B, IRF2BP2 and LSD1 depletion. Gene expression changes are shown as log2 fold change (FC). (E-G) Correlations between gene expression changes (log2 FC) after ETO2/IRF2BP2/LSD1 knockdown (KD; 72h post-transduction) or MEL cell differentiation (96 hours). Red dots represent individual genes (see Methods for more information on thresholds used). Locations of archetypical late erythroid genes (e.g. *Alas2*, *Epb4.2*, *Gypa*) within the graphs is indicated. Bars represent averages of at least three independent experiments; error bars denote s.d.

LDB1-complex (Fig.4D-F). When G1E-ER cells were differentiated by translocation of GATA1 into the nucleus, the TAL1 activator was displaced from these putative regulatory elements (Fig.4F), along with a loss of *Ir2bp2* expression (Fig.4C) and RNA polymerase II (RNAPII) occupancy of the locus (Fig.4F). Collectively, these data show that during erythroid differentiation, as was reported for *Cbfa2t3*, *Irf2bp2* expression is repressed in a GATA1-dependent manner, very likely involving negative auto-regulation by ETO2/IRF2BP2.

IRF2BP2 cooperates with ETO2 to impose transcriptional repression on erythroid genes

We next tried to address the functional roles played by ETO2 and IRF2BP2 in erythroid cells. ShRNA-mediated knockdowns (KD) of *Cbfa2t3* and *Irf2bp2* were performed in MEL cells, after which the expression of several ETO2-LDB1 target genes was measured. As shown in Figure 5, depleting ETO2 (Fig.5A) or IRF2BP2 (Fig.5B) results in increased *Alas2*, *Epb4.2*, *Gypa* and *Slc22a4* expression levels, establishing the repressive roles of ETO2 and IRF2BP2 in regulating archetypical erythroid target genes. This result also corroborates that ETO2 and IRF2BP2 form a functional erythroid co-repressor complex. In marked contrast, when performing the same experiments for LSD1 (encoded by the *Kdm1a* gene, Fig.5C), which co-occupies the same genes (Supplemental Fig. 3), either no significant change (*Alas2*, *Gypa*, *Slc22a4*) or decreased expression (*Epb4.2*) was observed. This result, together with the data derived from the reporter assays (Fig.2D) suggests that LSD1 does not mediate transcriptional repression by ETO2, and might even play an opposite role. In order to more comprehensively identify genes controlled by ETO2 and IRF2BP2, transcriptome analyses were carried-out by RNA-Sequencing (RNA-Seq) after ETO2 and IRF2BP2 depletion in MEL cells. Differentially expressed genes were also compared to the ones obtained after LSD1 depletion. Strikingly, we observed a high degree of correlation when comparing genes significantly misregulated after ETO2 or IRF2BP2 KD (Fig.5D-E), showing that genes controlled by ETO2 are also regulated by IRF2BP2, both in a positive and negative manner. Conversely, comparison of genes misregulated in both the *Cbfa2t3* (ETO2) and *Kdm1a* (LSD1) KD showed an inverse trend, as genes repressed by ETO2 were activated by LSD1 and vice versa (Fig.5D and F). In addition, the knockdown of another ETO2 interacting repressor *Gfi1b* (encoding GFI1B), which is known to interact with both ETO2 and LSD1, results in a very similar profile of differentially expressed genes when compared to the *Cbfa2t3* and *Irf2bp2* KD results (Fig.5D). This suggests that ETO2, IRF2BP2 and GFI1B negatively regulate a set of common genes and form a repressive complex in erythroid cells. Finally, we compared misregulated genes from the ETO2, IRF2BP2 and LSD1depletion experiments to the gene expression changes obtained after MEL cell differentiation (Fig.5G). The emerging correlations confirm the results presented in Fig.5A-C: genes de-repressed upon ETO2/IRF2BP2 depletion are upregulated during erythroid differentiation (including many archetypical late erythroid genes, Fig.5G), while the opposite trend emerged for LSD1.

IRF2BP2 and ETO2 repress essential erythroid pathways

To obtain functional insight into the genes affected in the *Cbfa2t3* and *Irf2bp2* KD experiments, we applied Ingenuity Pathway Analysis (IPA) on the misregulated genes to link the transcriptional regulatory activities of ETO2 and IRF2BP2 to biological functions. In MEL cells, 2625 genes were found differentially expressed upon IRF2BP2 depletion, and 724 upon ETO2 depletion. Combining these datasets, 58% of the ETO2 misregulated genes (420) were also found affected in the IRF2BP2 dataset (Fig.6A). Approximately 55% of the commonly misregulated genes were found upregulated and therefore appear to be repressed by ETO2/IRF2BP2. These 234 genes were highly enriched for erythroid functions (Fig.6B).

In fact, 71% of the genes coding for the major components of the heme biosynthesis pathway were bound by ETO2 and IRF2BP2 (Fig.6C; left graph). Furthermore, over 77% of the erythrocyte-specific membrane structural components and ion transporters are also targeted by the ETO2/IRF2BP2 complex (Fig.6C; right graph). In correspondence with this binding pattern, almost all of the above mentioned erythroid genes are misregulated upon ETO2 and/or IRF2BP2 depletion (100% of the heme biosynthesis genes and 78% of the erythrocyte membrane proteins are affected in at least one KD, see Fig.6C), with a strong preference for de-repression. In agreement with their co-occupancy by both proteins, many genes were upregulated after either *Cbfa2t3* or *Irf2bp2* KD (71% of heme biosynthesis genes and 36% of erythrocyte membrane proteins, see Fig.6C). Additionally, α- and β-globin gene activation was observed in both KD experiments (data not shown). In agreement with these observations, 51% of the ETO2/IRF2BP2 co-repressed genes were also found co-bound, again exhibiting a significant enrichment for erythroid functions (Fig.6D). Together, these observations strongly suggest that the ETO2/IRF2BP2 complex controls

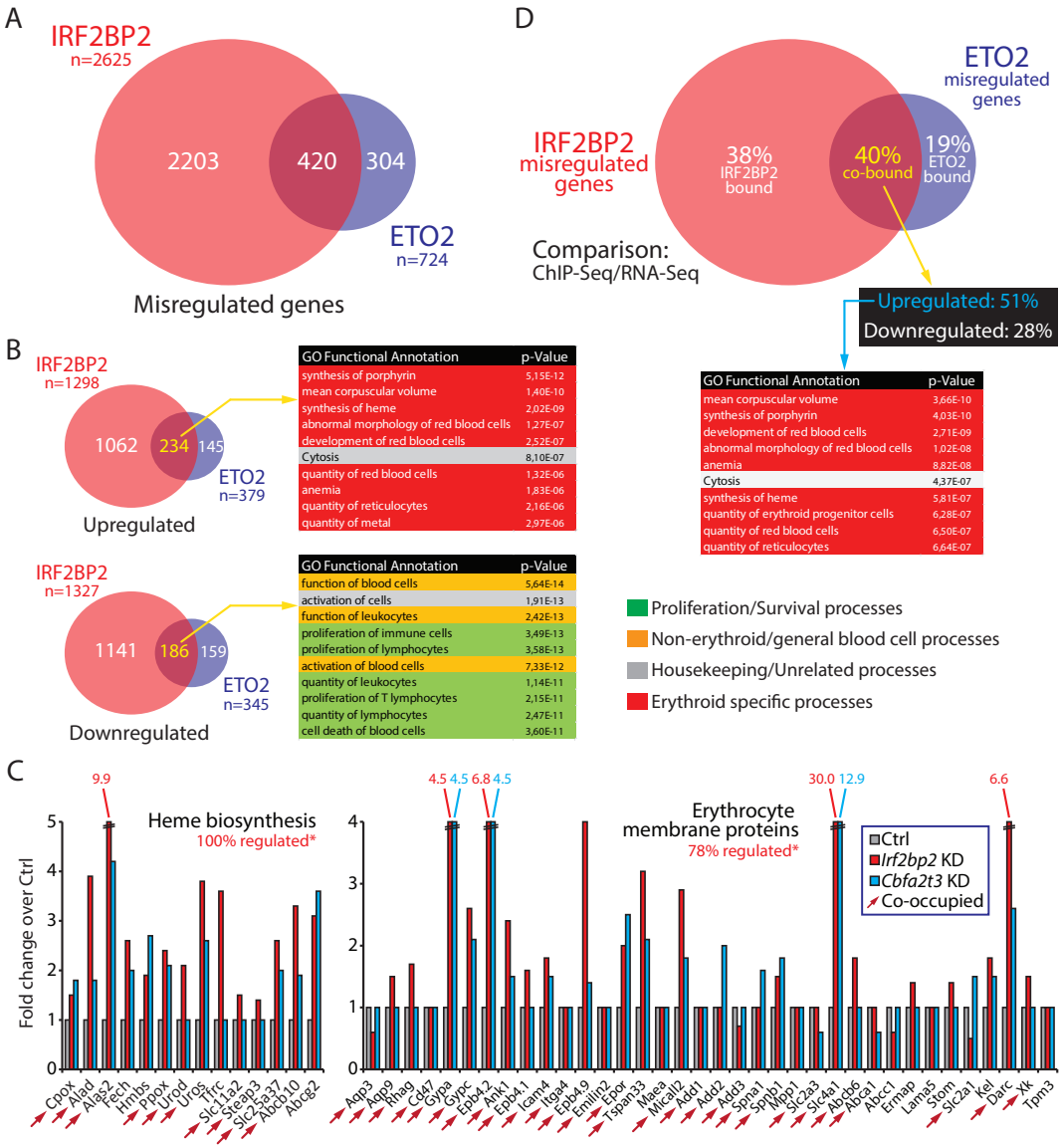


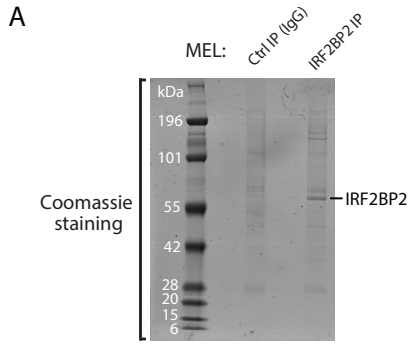
Figure 6. The ETO2-IRF2BP2 axis directly controls the expression of key heme biosynthesis and erythrocyte membrane proteins. (A) Venn diagram of differentially expressed genes (log2 FC >0.5/-0.5, P<0.05) after ETO2 or IRF2BP2 depletion in MEL cells. (B) Venn diagrams of upregulated genes (top, log2 FC >0.5, P<0.05) and downregulated genes (bottom, log2 FC >-0.5, P<0.05) after ETO2 or IRF2BP2 depletion. Genes found commonly up- or downregulated were subjected to Gene Ontology (GO) analysis using Ingenuity Pathway Analysis (IPA); the top 10 significantly associated GO functional annotations are shown. GO terms were categorized as in Fig.3B. (C) Fold changes in gene expression (log2 FC >0.5/-0.5, P<0.05; genes not significantly affected were given a fold change of 1) of heme biosynthesis and erythrocyte membrane protein genes upon *Cbfa2t3* (encoding ETO2*)* and *Irf2bp2* knockdown. Expression levels obtained from MEL cells transduced with a non-targeting shRNA ('ctrl') were set to 1. Genes bound by both ETO2 and IRF2BP2 in MEL cells are marked by a red arrow. *'% regulated' refers to the % of total genes in the group misregulated upon *Irf2bp2* and/or *Cbfa2t3* KD. (D) Combinatorial analysis of ETO2/IRF2BP2 ChIP-Seq data and differentially expressed genes after *Cbfa2t3*/*Irf2bp2* RNAi: 38% of the IRF2BP2 misregulated genes are also bound by IRF2BP2; 19% of ETO2 misregulated genes are bound by ETO2. Of the commonly misregulated genes 40% was bound by both factors, which increased to 51% when only upregulated genes were considered (as determined by GREAT analysis, see Methods of more details). GO analysis using IPA was also performed on this gene set.

the expression of key genes critical for erythroid cell identity and function.

ETO2 and IRF2BP2 also modulate a set of 186 genes that are downregulated upon factor depletion (Fig.6B), of which 28% was also co-occupied (Fig.6C). This suggests that ETO2/IRF2BP2 containing complexes can also function in gene activation, perhaps due to a recruitment of activating factors or a loss of key co-repressor molecules. Overrepresented among these are genes known to play a role in blood cell activation, proliferation and cell death (Fig.6B). Such pathways are known to be suppressed upon erythroid differentiation and might (in part) be activated by ETO2/IRF2BP2 in progenitor cells (Testa 2004). Surprisingly, a large fraction of the overrepresented processes were related to leukocyte and lymphocyte biology (Fig.6B).

IRF2BP2 interacts with NCOR corepressor proteins in erythroid cells

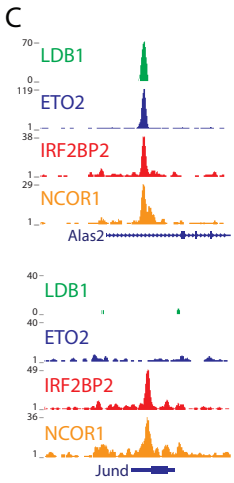
Although our data strongly suggest a repressor function for IRF2BP2 in erythroid gene regulation, how IRF2BP2 achieves gene repression is still unclear. We therefore purified endogenous IRF2BP2-containing protein complexes from MEL cells and identified the interacting proteins by mass spectrometry. As shown in Figure 7, we could retrieve known interacting proteins such as the other IRF2BP family members and several LDB1-complex members. Additionally, IRF2BP2 was also found to interact with proteins involved in the cell cycle and transcriptional regulation (Fig.7B). Among the latter group were several protein complexes known to mediate transcriptional repression. Prominent among these was the NCOR/SMRT corepressor complex. Key components of this complex are the nuclear receptor corepressor protein 1 (NCOR1) and 2 (NCOR2, also known as SMRT), and their repressive actions have been well documented (Mottis et al. 2013). Intriguingly, *Ncor1*$^{-/-}$ mice die *in utero* due to abnormal erythropoiesis (Jepsen et al. 2000). To test whether NCOR proteins are indeed recruited to the regulatory elements of ETO2/IRF2BP2 target genes, we performed NCOR1 ChIP-Seq in MEL cells. This revealed a significant overlap between NCOR1 and ETO2/IRF2BP2 binding sites (1164 sites, Fig.7C-D). In accordance with a possible cooperative relationship between these proteins, we found that these co-occupied sites include >64% of the erythroid-specific genes involved



B

Proteins identified by LC-MS/MS in MEL cells

Precipitated proteins	# of peptides
IRF2BP family	
IRF2BP2	43
IRF2BP1	30
IRF2BPL	34
LIM binding protein	
LDB1	7
ETO family	
ETO2	16
MTGR1	4
Basic helix-loop-helix	
SCL/TAL1	9
E2.2	3
E2A (E12/E47)	3
HEB	14
SSBP family	
SSBP3	3
Histone acetyl transferase complexes	
TRRAP	27
Cyclin-dependent kinases	
CDK11B	7
BHC repressor complex	
HMG20a	3
Ncor/SMRT corepressor complex	
NCOR1	10
NCOR2/SMRT	11
SHARP	25
SIN3A	4
HDAC3	2*



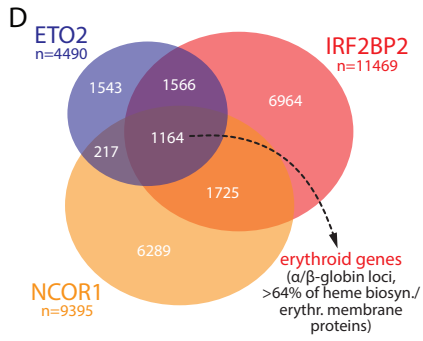


Figure 7. *(continued on next page)*

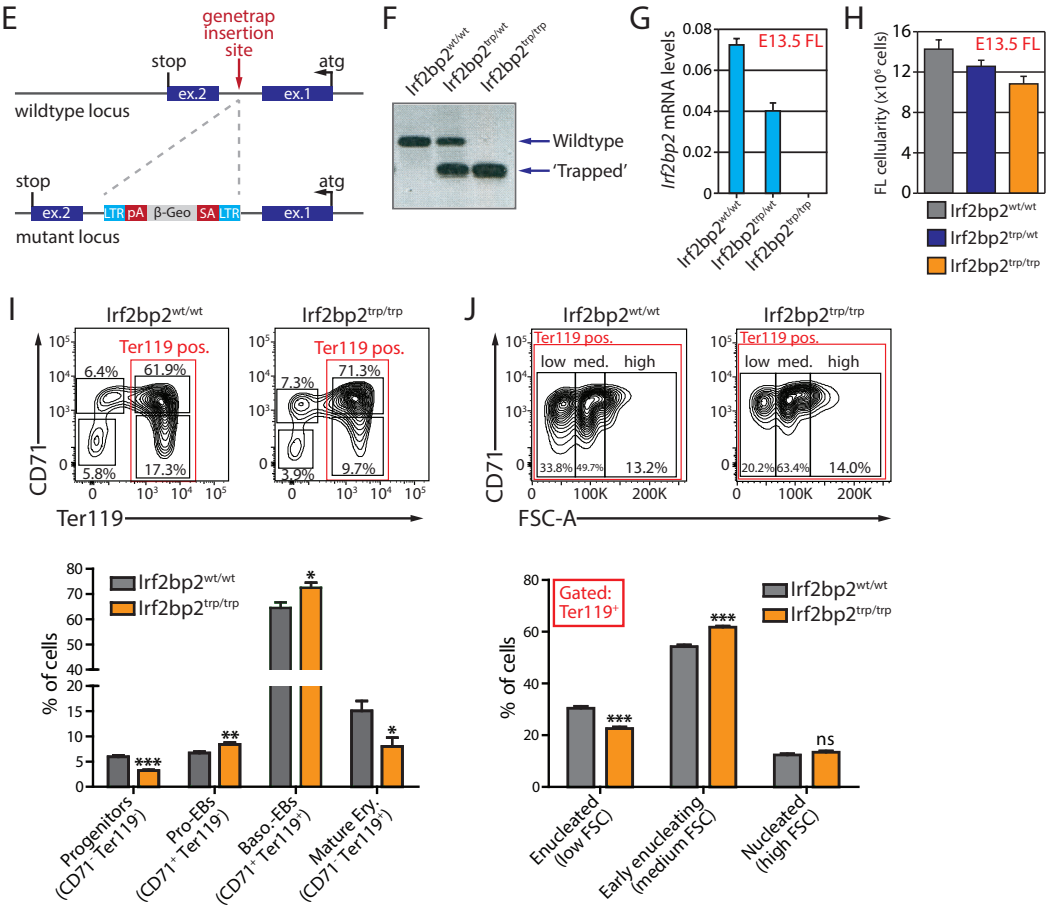


Figure 7. Characterization of IRF2BP2 protein partners reveals a possible NCOR1-mediated mechanism of repression and IRF2BP2-deficient mice show defects in fetal liver erythropoiesis. (A) Coomassie staining of IRF2BP2 and Control IgG immunoprecipitated proteins separated by SDS-PAGE. (B) IRF2BP2 interacting proteins identified by LC-MS/MS in MEL cells. Proteins pulled down in 2 independent experiments and with low background scores are shown, except for HDAC3 (*only detected in one pull down). (C) Examples of NCOR1 recruitment to IRF2BP2 binding sites. (D) Venn diagram showing the genome-wide overlap between ETO2, IRF2BP2 and NCOR1 binding sites in MEL cells. Note the significant co-localization of all three factors on the chromatin (1164 sites), which included the α- and β-globin loci and >64% of heme biosynthesis and erythrocyte membrane protein genes shown in Fig.6C. (E) A genetrap vector (containing a strong splice-acceptor [SA] and a polyadenylation sequence [pA]) was retrovirally inserted in the *Irf2bp2* intron to disrupt full-length mRNA production (genetrap allele is referred to as 'Irf2bp2^{trp}'). (F) Typical genotyping results obtained from a standard 3-primer PCR strategy. (G) *Irf2bp2* mRNA levels in whole fetal livers (FL) from E13.5 mouse embryos with the indicated genotypes (n=4-6 embryos per genotype, normalized to *Rnh1* levels). (H) Total FL cellularity in E13.5 embryos with the indicated genotypes (n=5-21 embryos per genotype). (I-J) Flowcytometry analysis (CD71-Ter119 double-staining) of FLs from E13.5 embryos with the indicated genotypes (n=9-11 embryos per genotype). Representative flowcytometry plots are shown on top; average values are plotted as bar graphs underneath. Panel I shows a quadrant analysis of CD71-Ter119 staining on all live (Hoechst negative) single cells to visualize erythroid differentiation. Panel J shows Ter119^{+} FL cells separated into three populations based on FSC profile (Sui et al. 2014). Differences between wildtype and Irf2bp2$^{trp/trp}$ embryos were tested for statistical significance (Mann Whitney U test; *P<0.05, **P<0.01, ***P<0.001) Error bars denote s.d.

in heme biosynthesis and red cell membrane function (Fig.7C-D). Furthermore, the ETO2-IRF2BP2-NCOR1 triad occupies key regulatory elements within the α and β globin loci (data not shown). These data indeed suggest that IRF2BP2-mediated gene repression involves the NCOR1 corepressor complex.

IRF2BP2 deficient mice are not viable and show abnormal fetal liver erythropoiesis

Next, we interrogated IRF2BP2 function *in vivo*. For this purpose, we used an IRF2BP2-deficient mouse model generated by a genetrap strategy (Fig.7E). The genetrap vector (containing a strong splice-acceptor) was retrovirally inserted in the *Irf2bp2* intron, resulting in a complete disruption of full-length mRNA production (Fig.7E-G). Mice homozygous for the *Irf2bp2* genetrap allele (hereafter referred to as *Irf2bp2*$^{trp/trp}$ mice) were rarely obtained and did not survive past 4 weeks of age, displaying severe growth retardation (data not shown). In fact, although *Irf2bp2*$^{trp/trp}$ embryos appeared to develop normally up to E18.5, live births were very rare (<5% of the expected number). This indicates that *Irf2bp2*$^{trp/trp}$ mice die either late during gestation or immediately after birth, for yet unknown reasons. To determine whether definitive erythropoiesis was affected in these mice, we collected E13.5 FL tissue from litters obtained after crossing *Irf2bp2*$^{trp/wt}$ mice. At this stage of murine embryonic development, the FL is the main site of definitive hematopoiesis and consists mainly of developing erythrocytes (Orkin and Zon 2008). *Irf2bp2*$^{trp/trp}$ FLs showed reduced total cellularity (Fig.7H). When stained with antibodies against the developmental CD71 and Ter119 surface markers, erythroid development in *Irf2bp2*$^{trp/trp}$ FLs showed several defects (Fig.7I). We observed a marked reduction in the double-negative immature progenitor compartment, while cells belonging to the more mature erythroblast stages (the CD71^{+}Ter119^{-} and double-positive stages) were more abundant in *Irf2bp2*$^{trp/trp}$ FLs. Furthermore, the relative number of mature Ter119^{+}CD71^{-} erythrocytes was significantly reduced in the absence of IRF2BP2. These data indicate that IRF2BP2 is important for effective FL erythropoiesis, as the output of mature erythrocytes is impaired in the absence of a functional *Irf2bp2* allele.

Finally, we further characterized terminal differentiation in *Irf2bp2*$^{trp/trp}$ FLs by separating the Ter119^{+} population based on its forward scatter (FSC) profile (Sui et al. 2014) (Fig.7J). As erythroid differentiation is paralleled by a reduction in cell size, this analysis visualizes a terminal differentiation gradient ranging from large and nucleated cells (high FSC) to small, enucleated cells (low FSC). Early enucleating cells (medium FSC) were more abundantly present in IRF2BP2 deficient FLs, while the percentage of small and enucleated erythrocytes was reduced (Fig.7J). These observations point at a block in terminal erythroid maturation in the absence of IRF2BP2, confirming the notion that IRF2BP2 is required for effective erythropoiesis *in vivo*.

Discussion

Developmental processes are coordinated by spatio-temporal changes in gene expression laid down by the combinatorial actions of TFs and the cofactors they recruit. Exactly how TFs in large multimeric complexes cooperate to create a regulatory environment that allows for rapid modulation of gene expression programs is under intense investigation. Here we address the observation of a master hematopoietic TF complex, containing key factors required for the activation of a tissue-specific gene expression program, that binds its target genes but maintains them in a developmental stage-specific poised state. Previous studies have shown that the activating LDB1 TF complex is already recruited to genes of the late erythroid-specific transcriptome in erythroid progenitors, before their full activation (Schuh et al. 2005; Goardon et al. 2006; Meier et al. 2006; Soler et al. 2010; Li et al. 2013). One particular complex member, the ETO2 corepressor, was found to mediate this 'poised' state by repressing LDB1-complex target gene expression (Schuh et al. 2005; Goardon et al. 2006; Meier et al. 2006; Soler et al. 2010). ETO2-mediated repression remains poorly understood, although the GFI1B TF, HDACs and the Sin3A repressor protein have been implicated (either directly or via their interaction with TAL1) (Amann et al. 2001; Schuh et al. 2005; Fujiwara et al. 2010). We set out to further investigate the molecular mechanisms used by ETO2 to suppress terminal erythroid gene expression in progenitor cells.

A proteomics approach was first used to catalogue ETO2 interacting proteins in MEL erythroid progenitors (Fig.1) and identified several repressor candidates known to bind ETO2 or other LDB1-complex members (e.g. GFI1B (Schuh et al. 2005) and LSD1 (Hu et al. 2009)). Interestingly, we also detected the IRF2BP2 corepressor in our interaction screen. Follow-up experiments firmly establish a cooperative role for ETO2 and IRF2BP2 in maintaining the late erythroid transcriptional program in a 'poised' state: (1) IRF2BP2 strongly enhances ETO2-mediated repression *in vitro*, which is fully dependent on the ETO2-BP2 interaction (Fig.2); (2) ETO2 and IRF2BP2 chromatin occupancy shows extensive genome-wide colocalization at genes

involved in red blood cell development and function (Fig.3); (3) Like *Cbfa2t3* (ETO2), *Irf2bp2* expression is reduced upon erythroid differentiation, concomitant with the upregulation of its erythroid target genes (Fig.4); (4) Depletion of ETO2 or IRF2BP2 leads to very similar effects on gene expression, in particular the strong derepression of the late erythroid-specific transcriptome (Fig.5); (5) ETO2 and IRF2BP2 bind the regulatory regions of >70% of the critical heme biosynthesis and erythrocyte membrane genes, the majority of which are repressed by both factors (Fig.6).

Our biochemical analyses of IRF2BP2 protein complexes in MEL cells revealed the presence of NCOR/SMRT corepressor complex members (Fig.7B). In accordance, a key component of this complex, NCOR1, showed extensive genomic co-occupancy with IRF2BP2 and the ETO2/LDB1-complex (Fig.7D). Among these co-occupied sites we found the vast majority of ETO2/IRF2BP2-repressed erythroid genes. Based on these data, we propose that IRF2BP2 confers repression upon ETO2/LDB1-complex target genes via its interaction with the NCOR/SMRT corepressor complex. In accordance with our hypothesis, NCOR1-deficient mice showed abnormal FL erythropoiesis and developed severe anemia during mid-gestation (Jepsen et al. 2000).

We have also investigated the role of other ETO2-interacting putative repressor proteins. Although we could not detect mSin3A in our ETO2 IPs, we did find the GFI1B TF and the LSD1 lysine demethylase, both of which have been implicated in the repression of LDB1-complex target genes (Schuh et al. 2005; Hu et al. 2009; Foudi et al. 2014). Both proteins colocalize with the ETO2-containing LDB1-complex on the erythroid genome (Supplemental Fig.3). In discordance with the findings of Hu et al. (Hu et al. 2009), we found no evidence for LSD1-mediated repression of the erythroid-specific *epb4.2* gene (Fig.5). In fact, we observed the opposite effect of LSD1 depletion on the late red cell transcriptome when compared to the *Cbfa2t3*/*Irf2bp2* KD (Fig.5), similar to the loss of erythroid marker expression and differentiation observed upon LSD1 KD by Saleque et al. (Saleque et al. 2007). We conclude that LSD1, as part of the LDB1-complex, in general fulfills an activating role in erythroid differentiation (i.e. possibly through the control of H3K4 methylation status (Wang et al. 2007)). In contrast, GFI1B, a DNA-binding repressor previously found to be required for terminal erythroid differentiation (Saleque et al. 2002), appeared to repress LDB1-complex target genes in a similar manner as ETO2/IRF2BP2 (Fig.5D). As was reported for ETO2, interactions between GFI1B and the activating LDB1-complex member TAL1 were strongly diminished upon terminal erythroid differentiation (Schuh et al. 2005). Cooperation of GFI1B with ETO2 and IRF2BP2 seems a plausible scenario warranting further investigation.

Intriguingly, IRF2BP2 binds many genomic regions independent of ETO2 and the LDB1-complex (Fig.3 and data not shown). Furthermore, IRF2BP2 depletion affected the expression of numerous genes in an ETO2-independent fashion (Fig.6). These observations suggest that IRF2BP2 plays additional roles in erythroid progenitors, independent of ETO2 and the LDB1-complex. In such cases, targeting of IRF2BP2 to the DNA could be mediated by ETS TFs or SP1, as binding motifs for these factors were enriched at sites only bound by IRF2BP2 (Fig.3C-D). Surprisingly, we did not detect a significant enrichment of IRF binding motifs at these regions, nor did we find IRF TFs interacting with IRF2BP2 in our mass spectrometry experiments. IRF2BP2 was originally identified as an IRF2 interacting factor in a yeast two-hybrid screen (Childs and Goodbourn 2003). An IRF2-IRF2BP2 complex was recently detected in the K562 human erythroleukemia cell line (Xu et al. 2012) and IRF2 is expressed in MEL and primary murine erythroid cells (data not shown). Whether this discrepancy reflects a species-specific difference or differences in experimental systems is unclear. Nevertheless, our combined analysis of IRF2BP2 binding sites and protein partners does provide preliminary insight into the ETO2/LDB1-independent functions of IRF2BP2. Genes bound only by IRF2BP2 were significantly enriched for functions related to proliferation and apoptosis (Fig.3B), and the cell-cycle regulator CDK11B (Li et al. 2004) interacts with IRF2BP2 (Fig.7B). Interestingly, several studies have implicated IRF2BP2 in the regulation of cell survival (Koeppel et al. 2009; Tinnikov et al. 2009; Yeung et al. 2011).

In agreement with our experiments in MEL cells, IRF2BP2 also appears to be important for erythropoiesis *in vivo*. Perinatal lethality of IRF2BP2-deficient mice precluded the analysis of adult erythropoiesis in our *Irf2bp2* genetrap model. However, analysis of mid-gestation definitive FL erythropoiesis in these mice showed that IRF2BP2 is required for an effective output of terminal erythroid differentiation (Fig.7I-J). The exact nature of this defect remains to be determined, but our experiments indicate the presence of a partial differentiation block at the erythroblast stage, just prior to enucleation (Fig.7I-J). Alternatively, the observed erythroblast expansion could be a consequence of accelerated progenitor differentiation or represent a compensatory mechanism, which could also explain the partially exhausted

progenitor compartment.

In summary, we show that the control of developmentally poised erythroid genes depends on the cooperative actions of ETO2 and its novel binding partner IRF2BP2. Repression by the ETO2-IRF2BP2 axis is lost during erythroid differentiation, resulting in the full activation of the late erythroid-specific transcriptome by the LDB1-complex. These results provide new insight into the control of lineage-specific transcriptional programs, as they suggest that an intricate balance between the activating and repressive components of a TF complex underlies the implementation of lineage-specific gene expression. Furthermore, using an IRF2BP2-deficient mouse model, we confirmed the relevance of a functional *Irf2bp2* allele for effective erythropoiesis *in vivo*.

Methods

Cell culture and Irf2bp2 genetrap animals

Mouse erythroleukemia (MEL) and HEK 293T cells were maintained in DMEM containing 10% FCS and penicillin/streptomycin. ETO2-V5-Bio MEL cells expressing BirA were generated and maintained as described previously (Soler et al. 2010; Soler et al. 2011). *Irf2bp2*$^{trp/wt}$ C57BL/6 ES cells were produced by the Texas A&M Institute for Genomic Medicine (College Station, TX) through the insertion of a genetrap construct in the first intron of the *Irf2bp2* gene (clone IST11591C1). Gene trap location was verified using standard PCR and sequencing methods. Mouse ES cells were injected into blastocysts and implanted into pseudopregnant albino fosters according to standard methods. Chimeric animals were further crossed to obtain heterozygous founders and mice were further crossed on a mixed FVB/N-C57BL/6 background. Mice were genotyped using a standard 3-primer PCR method. All animal experiments were carried out according to institutional and national guidelines.

(Co-)Immunoprecipitations and mass spectrometry analysis in MEL cells

Protocols for the preparation of nuclear extracts, streptavidin-mediated protein capture and LC-MS/MS in MEL cells have been described previously in detail (Soler et al. 2011). For endogenous co-immunoprecipitations, MEL nuclear extracts were diluted to reach 100mM KCl salt concentration using Heng 0 buffer (20mM HEPES KOH pH7.9, 20% glycerol, 0.25mM EDTA, 0.05% Np40). For co-IP experiments in MEL cells, 0.5 mg nuclear extract was used per IP. Extracts were treated with 1U Benzonase nuclease (Millipore). Protein extracts were incubated with the specific antibody overnight at 4°C, followed by addition of protein A or G Sepharose bead slurry (50µl slurry per IP; Millipore) and incubation at 4°C for 1 hour. Beads were pelleted, washed 3 times in Heng 100 buffer (Heng buffer containing 100mM KCl) and boiled for 5 min. at 95°C in Laemmli buffer before being subjected to Western Blot analysis. Proteomics analysis of IRF2BP2 interacting proteins was carried-out by direct immune-capture as described previously (van den Berg et al. 2010). Briefly, purification of endogenous IRF2BP2 protein complexes was performed by crosslinking 10µg of anti-IRF2BP2 monoclonal antibody (see below), or control Ig to 50µl protein G Sepharose beads (Amersham). Antibody-bead complexes were blocked with 0.1 mg/ml insulin (Sigma), 0.2 mg/ml chicken egg albumin (Sigma) and 1% fish skin gelatin (Sigma) for 1h at RT and directly added to 1.5ml of MEL nuclear extracts containing benzonase. After 3h incubation at 4°C, antibody-bead complexes were washed 5 times in C-100 buffer (20 mM Hepes pH 7.6, 20% glycerol, 100 mM KCl, 1.5 mM MgCl2, 0.2 mM EDTA, 0.02% NP40), and boiled in Laemmli buffer. Proteins were loaded on a 4-12% acrylamide gel and lanes were cut for LC-MS/MS analysis. For mass spectrometry analysis of ETO2 and IRF2BP2 interacting proteins, two independent biological replicates (for both experimental and control samples) were analyzed to ensure reproducible and specific binding partner identification. The following antibodies were used: ETO2 G-20 (Santa Cruz, sc9741), an IRF2BP2 rat monoclonal clone 10G3 (produced by Absea Antibodies, Beijing), GFI1B B-7 (Santa Cruz, sc8559), LDB1 N-18 (Santa Cruz, sc-11198), LSD1 (Abcam, ab17721), RUNX1 H-65 (Santa Cruz, sc28679), E2A V-18 (sc-349), HEB A-20 (sc-357), SSBP3 (Abcam, ab83815), V5 (Invitrogen, R960-25), Flag M2 (Sigma, F1804) and HA (Sigma, H6908).

Transfections, co-immunoprecipitations and luciferase assays in HEK 293T cells

HEK 293T cells were transfected with Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. For ETO2-IRF2BP2 interaction domain mapping we constructed a series of Flag-tagged ETO2 deletion mutants, V5-IRF2BP2 and the HA-IRF2BP2deltaRING deletion mutant (see Fig.2) in the pcDNA3.1 expression vector (Invitrogen). HEK 293T cells were lysed 48h post-transfection in whole cell lysis buffer

(20mM HEPES KOH pH7.5, 150mM KCl, 10% glycerol, 2.5mM EDTA, 5mM DTT, 0.1% Triton X-100 (Sigma) and protease inhibitor cocktail (Roche)). Extracts were treated with 1U Benzonase nuclease (Millipore). Protein extracts were incubated with the anti-Flag, anti-V5 or anti-HA antibodies overnight at 4°C, followed by addition of protein A or G Sepharose bead slurry (50µl slurry per IP; Millipore) and incubation at 4°C for 1 hour. Beads were pelleted, washed 3 times in lysis buffer and boiled for 5 min. at 95°C in Laemmli buffer before being subjected to Western Blot analysis. Full length *Kdm1a* (LSD1) cDNA was cloned in pcDNA3.1 for reporter assay experiments. The Gal4-ETO2 fusion protein was generated by fusing full length *Cbfa2t3* cDNA sequence to a Gal4 DNA binding domain in pcDNA3.1. The Gal4-responsive firefly luciferase plasmid was a kind gift from Dr. Jan van der Knaap (Erasmus MC). A Renilla luciferase expressing vector (pRL-TK, Promega) was co-transfected and used for normalization. Luciferase assays were performed using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions (Stadhouders et al. 2012).

ChIP and ChIP-Seq experiments

Protocols for the preparation of chromatin from MEL cells, immunoprecipitation and sample preparation for Illumina sequencing have been previously described in great detail (Soler et al. 2010; Soler et al. 2011). Antibodies used for ChIP are identical to those used for immunoprecipitation (detailed above), except for GFI1B (D-19 Santa Cruz, sc8559). Reads were mapped against NCBI build 37.1 of the mouse genome using Bowtie (Langmead et al. 2009). Uniquely mapped reads were extended to 200 bp in the 3' direction and were transformed into a genome-wide read density (coverage) using custom R scripts. MACS (Zhang et al. 2008), CCAT (Xu et al. 2010), and in-house peak calling software with default parameters were used to comprehensively identify binding sites. We combined binding sites identified by all three methods to define consensus-binding regions using GenomicRanges (Lawrence et al. 2013). Consensus-binding regions were given p-values based on a negative binomial distribution (Rozowsky et al. 2009) and assigned p-values were adjusted using the Benjamini-Hochberg (BH) method. Candidate binding sites were then selected for the downstream analysis based on the following criteria: read counts ≥10 reads, fold changes ≥2 compared to IgG control and adjusted p-values ≤0.01. To classify co-binding patterns, ETO2 and IRF2BP2 binding sites were combined using GenomicRanges. Binding signal coverage for each site was then normalized to obtain equal levels of background signal in both Antibody and IgG control experiments (normalization method was modified from Peakseq (Rozowsky et al. 2009)). Normalized coverage for the Antibody experiment was subtracted from the normalized coverage for the IgG control. We next retrieved the subtracted coverage within ± 0.5kb relative to the center of each binding site and calculated the standard z-scores in each sub-window (25bp). The matrix of standard z-scores per individual binding site was then subjected to K-means (K=3) clustering. Clustering analysis results were visualized with Java Treeview (Saldanha 2004). After K-means clustering, we selected representative binding sites for each co-binding pattern. We retrieved repeat-masked 200bp DNA sequences centered on each binding site and performed *de novo* motif discovery using MEME (Bailey and Elkan 1994). Results from MEME were subjected to an in-house ChIP-Seq analysis pipeline to generate motif logos and to calculate the proportion of motif containing sites (repeat motifs were discarded). Derived motifs were then compared to known motifs in the JASPAR database (Portales-Casamar et al. 2010) using Tomtom (Gupta et al. 2007). The online Genomic Regions Enrichment of Annotations Tool (GREAT (McLean et al. 2010)) was used to assign TF binding sites to genes and identify associated biological processes. Different GREAT analysis parameters were tested and yielded highly comparable results. Results using the 'single nearest gene method (within 1 Mb)' parameter are shown.

Real-time quantitative PCR (qPCR)

For gene expression analysis, RNA extractions were performed using TRIPure (Sigma) and cDNA synthesized using SuperScript II reverse transcriptase and Oligo(dT) primers (Invitrogen). ChIP DNA or cDNA were used as template in triplicate qPCR reactions (Platinum Taq DNA polymerase, Invitrogen) and analyzed on a CFX96 system (Bio-Rad). SYBR Green (Invitrogen) was used for quantification. Gene expression values were normalized to *Rnh1* mRNA levels (Stadhouders et al. 2012).

Lentivirus production and RNAi

Lentivirus particles were produced as described (Stadhouders et al. 2012). *Kdm1a* shRNA sequences were obtained from the MISSION TRC shRNA library (Sigma), designed manually and cloned into pLKO.1 (*Irf2bp2*; sh1: CTCCAGACAAAGCATTAAA and sh3: CAACGGGTCTAAAGCAGTT) or described before (*Cbfa2t3*

(Soler et al. 2010)). For *Gfi1b* knockdowns MEL cells were transfected with FlexiTube *Gfi1b* siRNA #1 and #7 (SI01011227 and SI05169871, Qiagen) using HiPerfect transfection reagent (Qiagen) according to the manufacturer's instructions. Non-targeting shRNAs/siRNAs were used as controls. Cells were harvested 48 or 72 hours after transduction/transfection and processed for RNA/protein extraction as described above.

RNA Sequencing
Total RNA was extracted from MEL or E13.5 sorted fetal liver populations using the RNeasy mini kit (Qiagen). After qPCR validation, RNA was used for mRNA-Sequencing on an Illumina HiSeq 2000 (standard TruSeq RNA sequencing protocol). At least two independent biological replicate samples were sequenced and used for downstream analysis. Raw reads were mapped with Bowtie (Langmead et al. 2009) against the murine transcriptome (NCBI build 37.1 Ensembl transcripts); non-uniquely mapped reads were discarded. Count number of reads per individual transcript and reads per kilobase per million mapped reads (RPKMs) were calculated and assigned to each transcript. Overlapping Ensembl transcripts were collapsed and the single highest expression value per gene locus was used. The non-adjusted read counts for each gene were used for statistical calculation of global differential expression using DESeq (Anders and Huber 2010). Differentially expressed genes were selected at an adjusted p-value of ≤0.05 (BH corrected). We selected differentially expressed genes with log2 fold changes ≥0.5 and log2 fold change ≤-0.5 in each knockdown data set to generate the correlation plots shown in Figure 5. We selected the top 142 differentially expressed genes (log2 fold change ≥1 and log2 fold change ≤-1) from the *Cbfa2t3* knockdown data set for the clustering analysis. We then retrieved all log2 fold change values for the 142 genes in the other knockdown data sets. Hierarchical clustering and visualization were performed using MeV (Saeed et al. 2006). For Ingenuity Pathway Analysis (IPA, Qiagen) only genes with a log2 FC ≥0.5 or ≤-0.5 and a P≤0.05 were considered. Core Analysis (standard settings) was used to extract GO terms that were associated with a gene-set in a statistically significant fashion.

Flow cytometry
E13.5 embryos were harvested and dissected to collect the fetal liver (FL). Single cell suspensions of whole E13.5 FLs were stained with CD71-FITC (553266) and Ter119-PE (553673) antibodies (BD Pharmingen). Flowcytometric analysis was performed using a BD LSRFortessa flowcytometer (BD Biosciences). FACS sorting was performed using a BD FACSAria III (BD Biosciences).

Immunofluorescence
MEL cells were fixed on poly-prep glass slides (Sigma) and fixed in 4% paraformaldehyde for 15 min. at room temperature. Cells were permeabilized with 0.1% Triton X-100, blocked with 0.5% BSA/0.15% Glycin (in PBS) and incubated overnight with ETO2 or V5 antibodies at 4°C. After a 2h incubation with appropriate secondary antibodies at room temperature, coverslips were mounted on glass slides with Vectashield (+DAPI, Vector Laboratories).

Published genome-wide datasets used
The following publicly available datasets were used: LDB1, GATA1 and ETO2 ChIP-Seq data (MEL, SRA ERA000161 (Soler et al. 2010)); RNA-Seq data (MEL/G1E/G1E-ER, ENCODE Penn State University; available at the UCSC Genome Browser [mouse genome, mm9]); p300 ChIP-Seq data (MEL, ENCODE Stanford/Yale; available at the UCSC Genome Browser [mouse genome, mm9]); H3K27Ac ChIP-Seq data (MEL/FL E14.5/Brain, ENCODE Ludwig Institute for Cancer Research; available at the UCSC Genome Browser [mouse genome, mm9]); DNAse I-Seq data (MEL/FL E14.5/Brain, ENCODE University of Washington; available at the UCSC Genome Browser [mouse genome, mm9]); GATA1, TAL1 and RNAPII ChIP-Seq data (MEL/G1E/G1E-ER, ENCODE Penn State University; available at the UCSC Genome Browser [mouse genome, mm9]); microarray gene expression data (Fetal and adult erythroid populations, ErythronDB database online (Kingsley et al. 2013)).

Data access
Newly generated datasets were submitted to the Gene Expression Omnibus (GEO), accession number GSE59859.

Acknowledgements

The authors would like to thank members of the Soler and Grosveld laboratories for helpful discussions. We thank Anouk van Oosten for constructing the V5-IRF2BP2 expression plasmid and Dr. Stephen Goodbourn (St George's Hospital Medical School, London, UK) for providing human *IRF2BP2* cDNA, Dr. Jan van der Knaap (Erasmus MC, Rotterdam, the Netherlands) for the Gal4-responsive luciferase reporter construct and Erasmus MC Experimental Animal Facility personnel for animal care and handling. We would also like to thank Dr. Paul-Henri Romeo (INSERM, Paris, France) for providing murine *Eto2* cDNA, and Erasmus MC Biomics personnel for excellent technical assistance with Illumina sequencing and data analysis. R.S. was supported by the the Royal Netherlands Academy of Arts and Sciences (KNAW; 'Academy Assistant' fellowship). C.A.S. is supported by a Marie Curie European Reintegration Grant (FP7-PEOPLE-2010-RG). P.K. is supported by grants from EpiGenSys/ERASysBio + /FP7 (NL: NWO, UK: BSRC, D: BMBF) the Bluescript EU Integrated Project and the Netherlands Genomics Initiative (MEC Booster grant). F.G. is supported by a KNAW Academy Professorship, the Cancer Genomics Center (NGI, NL), the NIRM (NL) consortium and the SyBoSS EU Consortium. E.S. is supported by grants from the ARC – 'projet ARC' and the Atip-Avenir program.

Author contributions:

R.S., F.G. and E.S. conceived and designed the experiments. R.S., P.K., H.I.B., X.Y., C.A.S. and E.S. performed the experiments. R.S., S.T., M.A., B.L. and E.S. analyzed the data. J.D. and K.B. performed proteomics experiments. C.K., Z.O. and W.v.IJ performed high-throughput sequencing experiments. A.M. performed ES cell injections into blastocysts to generate *Irf2bp2* genetrap chimeric mice. R.S., S.T., F.G. and E.S. wrote the paper.

References

Amann JM, Nip J, Strom DK, Lutterbach B, Harada H, Lenny N, Downing JR, Meyers S, Hiebert SW. 2001. ETO, a target of t(8;21) in acute leukemia, makes distinct contacts with multiple histone deacetylases and binds mSin3A through its oligomerization domain. *Molecular and cellular biology* **21**(19): 6470-6483.

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

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* **2**: 28-36.

Barrett CW, Smith JJ, Lu LC, Markham N, Stengel KR, Short SP, Zhang B, Hunt AA, Fingleton BM, Carnahan RH et al. 2012. Kaiso directs the transcriptional corepressor MTG16 to the Kaiso binding site in target promoters. *PLoS One* **7**(12): e51205.

Carneiro FR, Ramalho-Oliveira R, Mognol GP, Viola JP. 2011. Interferon regulatory factor 2 binding protein 2 is a new NFAT1 partner and represses its transcriptional activity. *Molecular and cellular biology* **31**(14): 2889-2901.

Childs KS, Goodbourn S. 2003. Identification of novel co-repressor molecules for Interferon Regulatory Factor-2. *Nucleic acids research* **31**(12): 3016-3026.

Chyla BJ, Moreno-Miralles I, Steapleton MA, Thompson MA, Bhaskara S, Engel M, Hiebert SW. 2008. Deletion of Mtg16, a target of t(16;21), alters hematopoietic progenitor cell proliferation and lineage allocation. *Molecular and cellular biology* **28**(20): 6234-6247.

Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Davis JN, McGhee L, Meyers S. 2003. The ETO (MTG8) gene family. *Gene* **303**: 1-10.

de Boer E, Rodriguez P, Bonte E, Krijgsveld J, Katsantoni E, Heck A, Grosveld F, Strouboulis J. 2003. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc Natl Acad Sci U S A* **100**(13): 7480-7485.

El Omari K, Hoosdally SJ, Tuladhar K, Karia D, Hall-Ponsele E, Platonova O, Vyas P, Patient R, Porcher C, Mancini EJ. 2013. Structural basis for LMO2-driven recruitment of the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. *Cell Rep* **4**(1): 135-147.

Fischer MA, Moreno-Miralles I, Hunt A, Chyla BJ, Hiebert SW. 2012. Myeloid translocation gene 16 is required for maintenance of haematopoietic stem cell quiescence. *EMBO J* **31**(6): 1494-1505.

Foudi A, Kramer DJ, Qin J, Ye D, Behlich AS, Mordecai S, Preffer FI, Amzallag A, Ramaswamy S, Hochedlinger K et al. 2014. Distinct, strict requirements for Gfi-1b in adult bone marrow red cell and platelet generation. *The Journal of experimental medicine* **211**(5): 909-927.

Fujiwara T, Lee HY, Sanalkumar R, Bresnick EH. 2010. Building multifunctionality into a complex containing master regulators of hematopoiesis. *Proceedings of the National Academy of Sciences of the United States of America* **107**(47): 20429-20434.

Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ, Bresnick EH. 2009. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol Cell* **36**(4): 667-681.

Gamou T, Kitamura E, Hosoda F, Shimizu K, Shinohara K, Hayashi Y, Nagase T, Yokoyama Y, Ohki M. 1998. The partner gene of AML1 in t(16;21) myeloid malignancies is a novel member of the MTG8(ETO) family. *Blood* **91**(11): 4028-4037.

Goardon N, Lambert JA, Rodriguez P, Nissaire P, Herblot S, Thibault P, Dumenil D, Strouboulis J, Romeo PH, Hoang T. 2006. ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *EMBO J* **25**(2): 357-366.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome biology* **8**(2): R24.

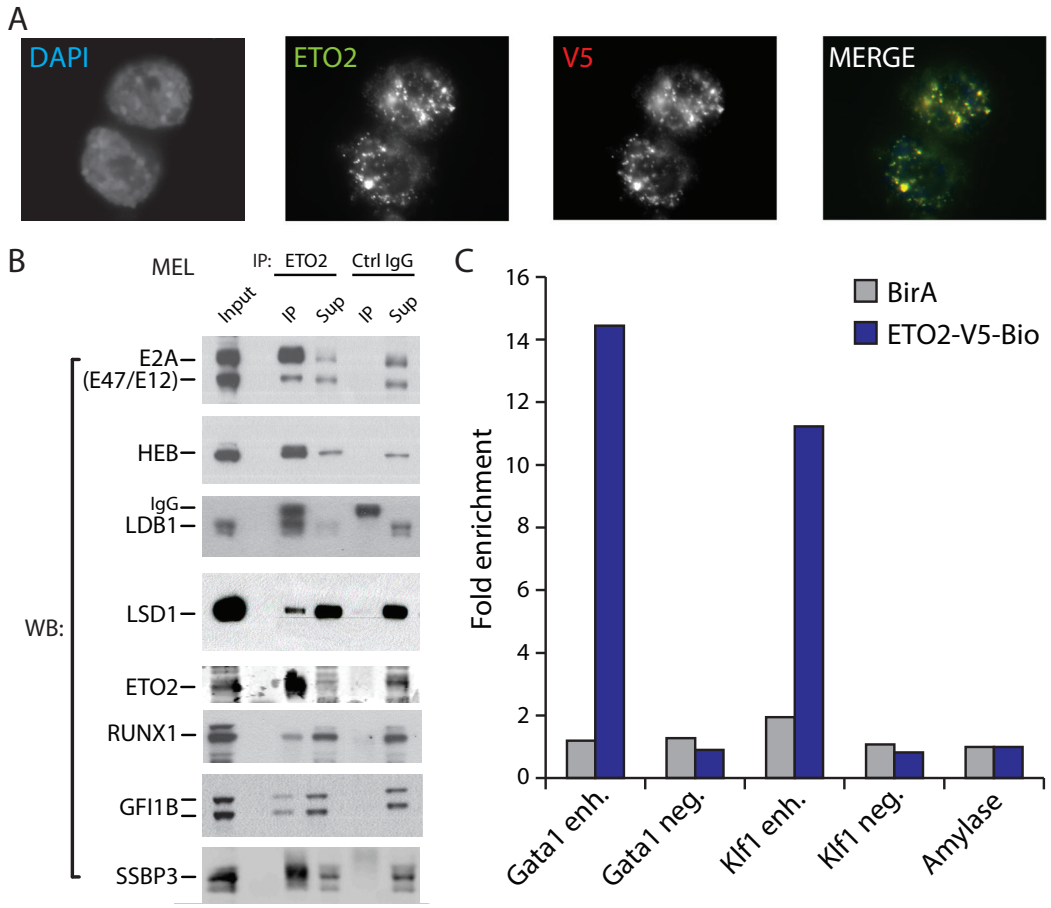
Hu X, Li X, Valverde K, Fu X, Noguchi C, Qiu Y, Huang S. 2009. LSD1-mediated epigenetic modification is required for TAL1 function and hematopoiesis. *Proceedings of the National Academy of Sciences of the United States of America* **106**(25): 10141-10146.

Hunt A, Fischer M, Engel ME, Hiebert SW. 2011. Mtg16/Eto2 contributes to murine T-cell development. *Molecular and cellular biology* **31**(13): 2544-2551.

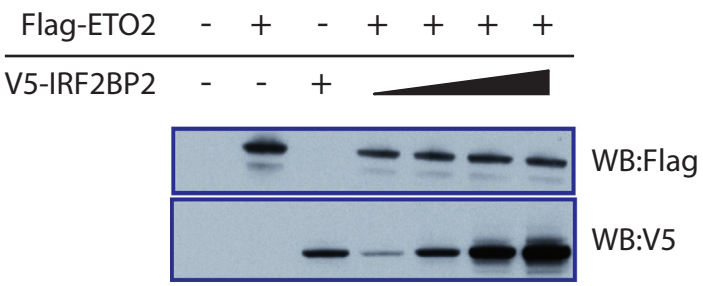
Jepsen K, Hermanson O, Onami TM, Gleiberman AS, Lunyak V, McEvilly RJ, Kurokawa R, Kumar V, Liu F, Seto E et al. 2000. Combinatorial roles of the nuclear receptor corepressor in transcription and development. *Cell* **102**(6): 753-763.
Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C. 2010. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res* **20**(8): 1064-1083.
Kiefer CM, Lee J, Hou C, Dale RK, Lee YT, Meier ER, Miller JL, Dean A. 2011. Distinct Ldb1/NLI complexes orchestrate gamma-globin repression and reactivation through ETO2 in human adult erythroid cells. *Blood* **118**(23): 6200-6208.
Kingsley PD, Greenfest-Allen E, Frame JM, Bushnell TP, Malik J, McGrath KE, Stoeckert CJ, Palis J. 2013. Ontogeny of erythroid gene expression. *Blood* **121**(6): e5-e13.
Koeppel M, van Heeringen SJ, Smeenk L, Navis AC, Janssen-Megens EM, Lohrum M. 2009. The novel p53 target gene IRF2BP2 participates in cell survival during the p53 stress response. *Nucleic acids research* **37**(2): 322-335.
Kumar P, Sharoyko VV, Spegel P, Gullberg U, Mulder H, Olsson I, Ajore R. 2013. The transcriptional co-repressor myeloid translocation gene 16 inhibits glycolysis and stimulates mitochondrial respiration. *PLoS One* **8**(7): e68502.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.
Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS computational biology* **9**(8): e1003118.
Li L, Freudenberg J, Cui K, Dale R, Song SH, Dean A, Zhao K, Jothi R, Love PE. 2013. Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. *Blood* **121**(22): 4575-4585.
Li T, Inoue A, Lahti JM, Kidd VJ. 2004. Failure to proliferate and mitotic arrest of CDK11(p110/p58)-null mutant mice at the blastocyst stage of embryonic cell development. *Molecular and cellular biology* **24**(8): 3188-3197.
Lipkowitz S, Weissman AM. 2011. RINGs of good and evil: RING finger ubiquitin ligases at the crossroads of tumour suppression and oncogenesis. *Nat Rev Cancer* **11**(9): 629-643.
Love PE, Warzecha C, Li L. 2014. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends Genet* **30**(1): 1-9.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**(5): 495-501.
Meier N, Krpic S, Rodriguez P, Strouboulis J, Monti M, Krijgsveld J, Gering M, Patient R, Hostert A, Grosveld F. 2006. Novel binding partners of Ldb1 are required for haematopoietic development. *Development* **133**(24): 4913-4923.
Mottis A, Mouchiroud L, Auwerx J. 2013. Emerging roles of the corepressors NCoR1 and SMRT in homeostasis. *Genes Dev* **27**(8): 819-835.
Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**(4): 631-644.
Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research* **38**(Database issue): D105-110.
Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* **27**(1): 66-75.
Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 microarray software suite. *Methods in enzymology* **411**: 134-193.
Saldanha AJ. 2004. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**(17): 3246-3248.
Saleque S, Cameron S, Orkin SH. 2002. The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. *Genes Dev* **16**(3): 301-306.
Saleque S, Kim J, Rooke HM, Orkin SH. 2007. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Mol Cell* **27**(4): 562-572.
Schuh AH, Tipping AJ, Clark AJ, Hamlett I, Guyot B, Iborra FJ, Rodriguez P, Strouboulis J, Enver T, Vyas P et al. 2005. ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Molecular and cellular biology* **25**(23): 10235-10250.
Soler E, Andrieu-Soler C, Boer E, Bryne JC, Thongjuea S, Rijkers E, Demmers J, Ijcken W, Grosveld F. 2011. A systems approach to analyze transcription factors in mammalian cells. *Methods* **53**(2): 151-162.
Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Stevens M, Kockx C, van Ijcken W et al. 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24**(3): 277-289.
Stadhouders R, Thongjuea S, Andrieu-Soler C, Palstra RJ, Bryne JC, van den Heuvel A, Stevens M, de Boer E, Kockx C, van der Sloot A et al. 2012. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J* **31**(4): 986-999.
Sui Z, Nowak RB, Bacconi A, Kim NE, Liu H, Li J, Wickrema A, An XL, Fowler VM. 2014. Tropomodulin3-null mice are embryonic lethal with anemia due to impaired erythroid terminal differentiation in the fetal liver. *Blood* **123**(5): 758-767.
Szalai G, LaRue AC, Watson DK. 2006. Molecular mechanisms of megakaryopoiesis. *Cell Mol Life Sci* **63**(21): 2460-2476.
Testa U. 2004. Apoptotic mechanisms in the control of erythropoiesis. *Leukemia* **18**(7): 1176-1199.
Tinnikov AA, Yeung KT, Das S, Samuels HH. 2009. Identification of a novel pathway that selectively modulates apoptosis of breast cancer cells. *Cancer Res* **69**(4): 1375-1382.
van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA. 2010. An Oct4-centered protein interaction network in embryonic stem cells. *Cell stem cell* **6**(4): 369-381.
Wang J, Scully K, Zhu X, Cai L, Zhang J, Prefontaine GG, Krones A, Ohgi KA, Zhu P, Garcia-Bassets I et al. 2007. Opposing LSD1 complexes function in developmental gene activation and repression programmes. *Nature* **446**(7138): 882-887.
Welch JJ, Watts JA, Vakoc CR, Yao Y, Wang H, Hardison RC, Blobel GA, Chodosh LA, Weiss MJ. 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**(10): 3136-3147.
Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK. 2010. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **26**(9): 1199-1204.
Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC et al. 2012. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev Cell* **23**(4): 796-811.
Yeung KT, Das S, Zhang J, Lomniczi A, Ojeda SR, Xu CF, Neubert TA, Samuels HH. 2011. A novel transcription complex that selectively modulates apoptosis of breast cancer cells through regulation of FASTKD2. *Molecular and cellular biology* **31**(11): 2287-2298.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**(9): R137.

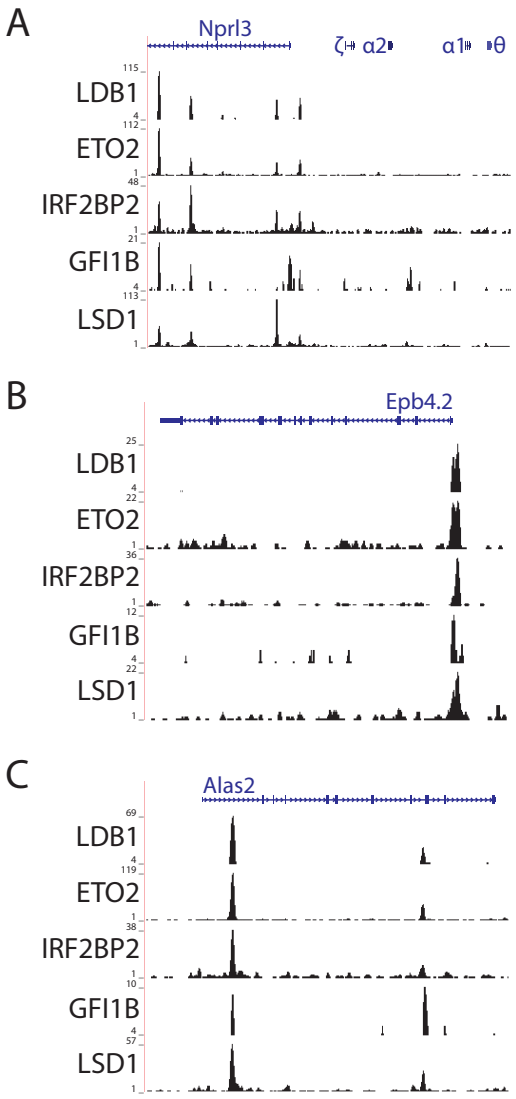
Supplementary Figures



Supplemental Figure 1. Validation of ETO2-V5-Bio functionality in MEL cells. (A) MEL cells stably expressing BirA and ETO2-V5-Bio cells were fixed and stained for endogenous ETO2 (in green) or for ETO2-V5-Bio (using a V5 antibody, in red). Note the nuclear localization (as compared to the DAPI nuclear stain) of ETO2-V5-Bio and co-localization with endogenous ETO2. (B) Endogenous co-immunoprecipitation validations of ETO2-V5-Bio interacting proteins in MEL cells identified by LC-MS/MS. Species-matched IgG was used to control for non-specific binding. (C) Bio-ChIP qPCR experiments showing recruitment of ETO2-V5-Bio to known endogenous ETO2 genomic binding sites (Gata1 -3.5 HS enhancer and the Klf1 upstream enhancer (Meier et al. 2006)). Regions immediately up- or downstream (+ or – 1kb) of the enhancer (enh.) sites were used as negative (neg.) controls. Enrichments were normalized to Amylase promoter values. A representative of two independent experiments is shown. WB, Western Blot; IP, immunoprecipitation



Supplemental Figure 2. Increasing IRF2BP2 levels do not have an impact on ETO2 protein stability when co-transfected in HEK 293T cells. Equal amounts of Flag-ETO2 and variable amounts of V5-IRF2BP2 expression constructs were co-transfected into HEK 293T cells. Protein extracts were prepared 48h post-transfection and Flag-ETO2 and V5-IRF2BP2 protein levels were visualized using Western Blot (WB) analysis.



Supplemental Figure 3. Co-occupancy of LDB1-complex target genes by IRF2BP2, GF1B and LSD1 in erythroid progenitors. ChIP-Seq data for LDB1, ETO2, IRF2BP2, GFI1B and LSD1 (from MEL cells) is shown for the α-globin (A), Epb4.2 (B) and Alas2 (C) loci. Note the high degree of co-occupancy of the ETO2-interacting corepressor proteins (IRF2BP2, GFI1B and LSD1) on known LDB1/ETO2-complex target genes.

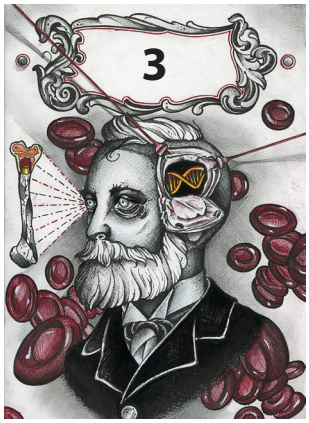
Chapter 3

Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions

Ralph Stadhouders[1]*, Petros Kolovos[1]*, Rutger Brouwer[2,3]*, Jessica Zuin[1], Anita van den Heuvel[1], Christel Kockx[2], Robert-Jan Palstra[1], Kerstin S Wendt[1], Frank Grosveld[1,4], Wilfred van IJcken[2]† & Eric Soler[1,4,5]†

[1]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands.
[2]Center for Biomics, Erasmus Medical Center, Rotterdam, The Netherlands.
[3]Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands.
[4]Cancer Genomics Center, Erasmus Medical Center, Rotterdam, The Netherlands.
[5]Laboratory of Hematopoiesis and Leukemic Stem Cells (LSHL), French Alternative Energies and Atomic Energy Commission (CEA)/Institut National de la Santé et de la Recherche Médicale (INSERM) U967, Fontenay-aux-Roses, France.

***These authors contributed equally.**
†Corresponding authors.



Published in:
Nature Protocols
2013; 8:509-24

Abstract

Chromosome conformation capture (3C) technology is a powerful and increasingly popular tool for analyzing the spatial organization of genomes. Several 3C variants have been developed (e.g., 4C, 5C, ChIA-PET, Hi-C), allowing large-scale mapping of long-range genomic interactions. Here we describe multiplexed 3C sequencing (3C-seq), a 4C variant coupled to next-generation sequencing, allowing genome-scale detection of long-range interactions with candidate regions. Compared with several other available techniques, 3C-seq offers a superior resolution (typically single restriction fragment resolution; approximately 1–8 kb on average) and can be applied in a semi-high-throughput fashion. It allows the assessment of long-range interactions of up to 192 genes or regions of interest in parallel by multiplexing library sequencing. This renders multiplexed 3C-seq an inexpensive, quick (total hands-on time of 2 weeks) and efficient method that is ideal for the in-depth analysis of complex genetic loci. The preparation of multiplexed 3C-seq libraries can be performed by any investigator with basic skills in molecular biology techniques. Data analysis requires basic expertise in bioinformatics and in Linux and Python environments. The protocol describes all materials, critical steps and bioinformatics tools required for successful application of 3C-seq technology.

Introduction

In recent years, it has become evident that the 3D organization of genomes is not random. Numerous studies have implicated long-range chromosomal interactions in several crucial cellular processes, including the regulation of gene expression[1,2,3,4]. Indeed, chromatin coassociations mediated by chromatin looping provide a means by which distal enhancers communicate with their target genes and stimulate transcription[5,6,7]. Accordingly, methods providing efficient and sensitive detection of chromatin looping events with high resolution are becoming increasingly popular. The development of 3C technology has revolutionized the analysis of spatial genomic organization by allowing the detection of chromatin coassociations with a resolution far beyond that provided by light microscopy–based studies[8]. 3C relies on the ability of distal DNA fragments to be ligated together when positioned in close proximity in the nuclear space. Over the past decade, several 3C variants have been developed, offering the possibility of analyzing chromatin looping events on a genome-wide scale (e.g., 4C[9,10,11,12], 5C[13], ChIA-PET[14], Hi-C[15]). We describe here in detail multiplexed 3C-seq, a 3C variant coupled to high-throughput sequencing that we recently developed[16,17]. Multiplexed 3C-seq allows genome-scale simultaneous detection of long-range chromatin interactions of numerous genomic elements in parallel and can be applied to low numbers of cells (from 1×10^6 cells[18] to as low as 300,000 cells (P.K. and E.S., unpublished data)). We recently used this technique to analyze the spatial organization of several loci, including the mouse β-globin (*Hbb*), myeloblastosis oncogene (*Myb*) and Ig kappa loci (*Igκ*), revealing crucial enhancer-gene communications[16,17,18].

Overview of the procedure

All 3C-based procedures use formaldehyde fixation of living cells or fresh tissues to preserve genomic architecture in its native state before fragmentation by restriction enzyme digestion. The digested cross-linked chromatin is subjected to a ligation reaction under dilute conditions, favoring intramolecular ligation events over intermolecular ligation events (proximity ligation). This step yields a 3C library composed of chimeric DNA molecules resulting from the ligation of (distal) chromatin fragments that were in physical proximity in the nuclear space (Fig. 1). The subsequent steps differ depending on the type of assay used. The 3C library can be directly analyzed by probing for specific interactions by PCR[19,20] or further processed for more global analyses using bait-specific primers (e.g., promoter-specific primer pair[9,10,11,12,16,17,18]) or whole-genome looping assays as in Hi-C[15]. In the 3C-seq procedure, the 3C library is subjected to a second restriction enzyme digestion using a frequent cutter, and fragments are circularized before an inverse PCR step using bait-specific primers (Fig. 1), similar to the original microarray-based 4C protocol[11]. This second restriction digest is necessary to decrease the size of the DNA circles, resulting in fragments that can be PCR-amplified efficiently. The inverse PCR products contain the DNA elements that were captured (i.e., ligated) by the bait sequence and thereby represent its native chromatin environment in the nucleus. The 3C-seq library is then directly sequenced on an Illumina HiSeq2000 platform, with the possibility of multiplexing sample sequencing by pooling up to 12 different bait-specific 3C-seq libraries in a single lane of a HiSeq2000 flow cell, providing marked cost reduction and increased throughput. Other sequencing platforms are, in

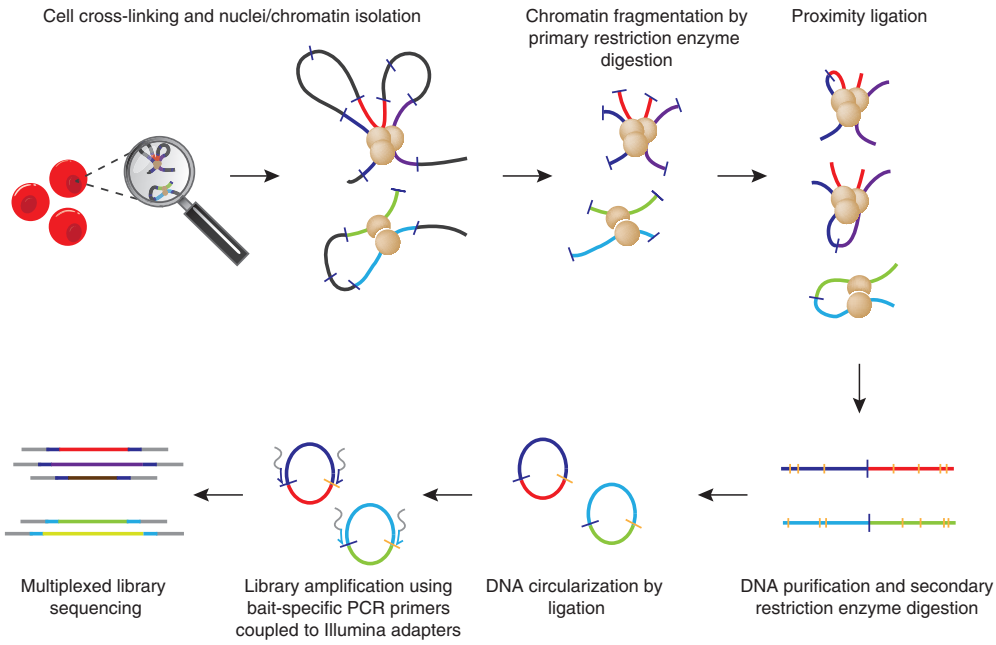


Figure 1: Overview of the multiplexed 3C-seq procedure. Nuclei from cross-linked cells are digested (primary restriction enzyme) and ligated under dilute conditions to physically link *in vivo* interacting DNA fragments. After a secondary digestion (secondary restriction enzyme) and ligation, inverse PCR is performed using bait-specific primers containing Illumina sequencing adapters to amplify unknown fragments interacting with the bait. PCR samples generated with different primer sets are then pooled and subjected to multiplexed library sequencing.

principle, compatible with multiplexed 3C-seq, but the multiplexing/de-multiplexing steps and associated informatics tools described here may need further optimization and adjustments.

Comparison of 3C-seq with other 3C-based methods

The choice between 3C and the different derivatives strongly depends on the biological question under consideration (Table 1). Although 3C-qPCR is particularly suited to quantitatively probe for specific interactions and interrogate a restricted number of chosen chromatin coassociations, it rapidly becomes technically demanding when large chromosomal domains are under investigation or when numerous interactions need to be analyzed in parallel for de novo detection of chromatin looping events. In the latter cases, high-throughput 3C derivatives such as 4C, 5C, 3C-seq or Hi-C technologies will be preferred. The 4C approach[10,11] consists of a large-scale analysis of chromatin interactions with a chosen bait sequence by probing the 4C library on DNA microarrays. It produces chromatin interaction maps of a single bait, with the coverage depending on the array used. 4C has the advantage of allowing unbiased detection of unknown bait-specific interactions, but is limited by the number of arrays needed to achieve genome-wide coverage and by the saturation of signals around the bait sequence, preventing the detection of medium- to close-range interactions (up to 200 kb away). The 5C variant[13] overcomes this limitation and offers the possibility of exploring every potential chromatin coassociation in large subchromosomal domains by using primer sets covering all possible interactions. It is, however, difficult to reach genome-wide coverage using 5C, as it requires extremely large numbers of primers for all possible intrachromosomal and interchromosomal interactions. HiC, in contrast, provides a global genome-wide analysis of all possible chromatin associations by coupling a modified 3C procedure to high-throughput sequencing[15]. Although it is extremely powerful, Hi-C requires substantial computational resources, and the number of sequence reads needed to obtain

TABLE 1 | Comparison between different 3C variants.

3C-based method	Applications	Advantages	Limitations
3C-(q)PCR[19,20]	One-to-one	Relatively simple analysis (no bioinformatics required)	Laborious, knowledge of locus required, proper controls are essential
3C-on-chip (4C)[9–11]	One-to-all	Relatively simple data analysis	Poor signal-to-noise ratio, difficult to obtain genome-wide coverage
3C sequencing (3C-seq or 4C-seq)[12,16]	One-to-all	Genome-wide coverage, high resolution, good signal-to-noise ratio, allows multiplexing for high-throughput	Restricted to a single view point per experiment (except when multiplexing), analysis requires some bioinformatics expertise
Multiplexed 3C-seq[17,18]	Many-to-all		
3C carbon copy (5C)[13]	Many-to-many	Explores interactions between many individual fragments simultaneously (instead of using a single viewpoint)	No genome-wide coverage, primer design can be challenging
Hi-C[15]	All-to-all	Explores the genome-wide interactions between all individual fragments simultaneously	Obtaining high resolution requires a massive sequencing effort; expensive, complicated analysis

high coverage of mammalian genomes renders it very expensive and, as a consequence, unaffordable for a large number of academic laboratories.

3C-seq provides a fast and affordable genome-scale 3C alternative (Fig. 2). The use of high-throughput sequencing eliminates the problems of limited coverage and saturating signals associated with microarray technology and markedly increases resolution and signal-to-noise ratios. A disadvantage of 3C-seq is that, as in 4C, the analysis is restricted to a single bait sequence and does not provide deep characterization of chromatin coassociations of several regulatory elements in parallel. The multiplexed 3C-seq protocol presented here (Figs. 1 and 2) addresses this limitation and shows that, by efficiently multiplexing bait-specific library sequencing, genome-scale interactions of up to 192 different genomic elements can be assessed in parallel on an Illumina HiSeq2000 platform, thereby markedly increasing the throughput of the technique and decreasing sequencing costs. Moreover, 3C-seq data analysis is facilitated by the availability of bioinformatics tools. We provide here a dedicated analysis pipeline facilitating the entire data handling process, including de-multiplexing, alignment and visualization. Together, this renders multiplexed 3C-seq an inexpensive and efficient method for in-depth analysis of complex genetic loci and genomic regulatory regions.

Applications of the method

3C-seq can be applied to any non-repetitive region of a genome. It is generally used to unravel medium- to long-range interactions (i.e., few kb to hundreds of kb) of a genomic element of interest. It is usually applied to detect interactions between promoter elements and the surrounding regions, or to connect distal enhancers to their target gene(s). With the recent developments in high-throughput chromatin occupancy profiling[21], large numbers of transcription factor binding and chromatin modification data sets are becoming available. Combined with this knowledge, 3C-seq can be used to analyze the functional relationships existing between regulatory elements, sites of active transcription, gene deserts or boundary elements where transitions in chromatin structure or transcription are observed (e.g., insulator elements or initiation sites for productive transcription elongation).

Limitations of 3C-seq

Similar to all 3C-based procedures, 3C-seq only provides topological information. The control experiments discussed in Experimental design will help validate and ensure the specificity of the observed interactions. Even so, it is recommended to combine 3C-seq data with results from complementary experiments (e.g., fluorescence in situ hybridization (FISH), gene expression analysis, chromatin immunoprecipitation (ChIP))[7,17,22] or, even better, with functional experiments, before drawing conclusions on the functional impact of chromatin coassociations.

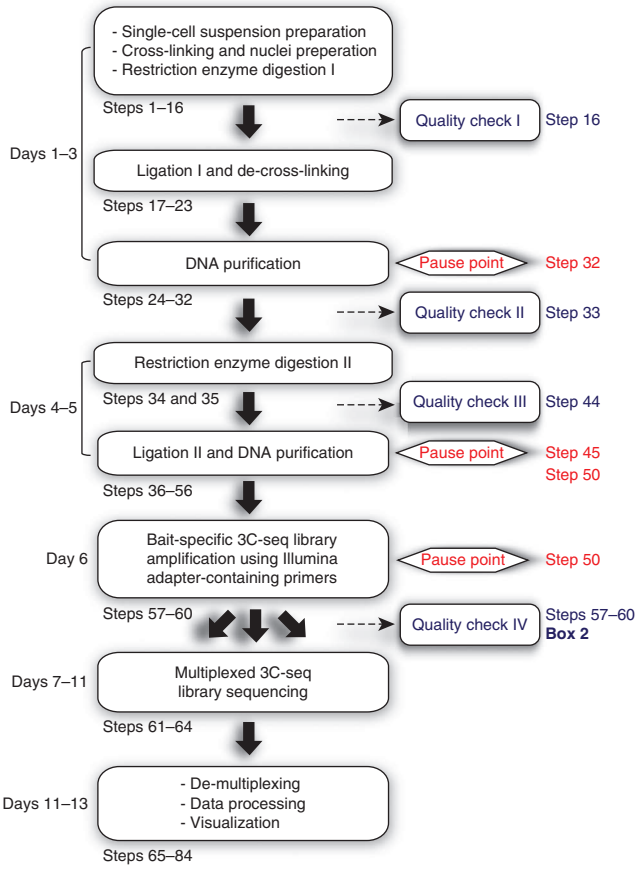


Figure 2: Flowchart of multiplexed 3C-seq data generation and processing. Steps involved in the multiplexed 3C-seq procedure are shown in blue rectangles. Time needed to complete these steps is depicted on the left. Pause points are indicated together with the timing of the different quality checkpoints: I, primary digestion efficiency (Step 16); II, ligation efficiency (Step 33); III, secondary digestion efficiency (Step 44); IV, 3C-seq PCR performance (Steps 57–60 and Box 2).

Experimental design

Fixing cells. Cell fixation, which represents the starting point of the procedure, provides the template for the essential proximity ligation step used to capture DNA-DNA interactions. Fixation conditions need to be standardized for increased reproducibility and efficient comparison between samples. In our hands, formaldehyde fixation conditions used in ChIP experiments (1–2% (vol/vol) formaldehyde, 10 min at room temperature (18–22 °C)) work well for 3C-seq[16,17,18]. More extensive fixation protocols have been reported to improve signal-to-noise ratios in the distance range of a few kb (ref. 23), although this protocol utilizes more frequently cutting restriction enzymes to obtain such resolution and might therefore be difficult to compare with our protocol.

Starting material. We have used many human and mouse cell or tissue types in 3C-seq experiments (Table 2), although certain cell or tissue types (e.g., fibroblasts) can be more difficult to handle. The use of single-cell suspensions is essential when performing 3C-seq (and other 3C-based protocols, for that matter). When working with tissues that are difficult to dissociate (e.g., brain, heart, lung), consider treating them with collagenase before formaldehyde fixation (see PROCEDURE Step 1 and TROUBLESHOOTING section). Previously published 3C (and derivate) protocols describe using 10^6 cells or more per experiment. We, however, have successfully applied 3C-seq on much smaller numbers of cells (i.e., FACS-sorted cell populations, using $<10^6$ cells), further extending its applicability (P.K. and E.S., unpublished data, and ref. 18).

Restriction enzyme choice. The resolution of a 3C-seq experiment depends on the first restriction enzyme used. Ideally, the restriction pattern given by the enzyme should provide evenly distributed fragments, separating the different regulatory elements of interest (e.g., promoter, enhancers). When possible, check for the presence of regulatory elements, transcription factor binding sites and histone modification patterns relevant for the tissue to be analyzed using publicly accessible databases such as ENCODE (http://genome.ucsc.edu/ENCODE/) in order to determine the most appropriate enzyme for the region of interest. We suggest using 6-base-recognizing enzymes (referred to as a 'six-cutter') such as EcoRI, HindIII, BglII, BamHI and XhoI, which perform well on cross-linked chromatin. The enzymes should be insensitive to mammalian DNA methylation in order to prevent introducing digestion biases. We observed that the use of a six-cutter yields better reproducibility at the single restriction fragment level than enzymes that cut more

TABLE 2 | Performance of different cell types and tissues successfully used for 3C-seq.

Cell or tissue type	Performance in 3C-seq	Special requirements
Hematopoietic cell types: mouse and human erythroid cells (FACS sorted and cultured), mouse B and T lymphocytes (FACS sorted and cultured), mouse erythroleukemia cell lines (MEL, I11) Hematopoietic tissue (mouse fetal liver E12.5-15.5, human fetal liver) Mouse ES cells (IB10), ES-derived Flk1$^+$ cells (magnetic-activated cell sorting (MACS)-sorted) HeLa cells	Excellent	None
Other mouse tissues (Mouse fetal brain E12.5-15.5) Rat tissues (liver, heart and lung)	Good	Use a collagenase treatment (PROCEDURE Step 1) to obtain a single-cell suspension for efficient cross-linking
Human primary melanocytes[33] Fibroblast cells: cell lines (NIH3T3) and primary cells (mouse dermal fibroblasts, mouse and human lung fibroblasts) HEK/293T cells K562 cells HUVEC cells Human ES cells (H9)	Poor: extensive nuclei aggregation resulting in poor digestion efficiencies	Ensure gentle handling of the cells and nuclei. Preferentially collect adherent cells with a scraper instead of trypsin. In case of aggregation, see **Table 3** for additional troubleshooting. Melanin produced by melanocytes is a potent PCR inhibitor and can be removed using a suitable column purification step[33]

frequently (e.g., 4-base-recognizing enzymes, referred to as a 'four-cutter'). The latter generate many more fragments per kb, which may lead to a poorer signal-to-noise ratio owing to more frequent intermolecular ligations. This could result in interaction signals being spread over several restriction fragments, thereby yielding interaction profiles that are sometimes more difficult to interpret. For instance, enhancer-promoter communication might be difficult to analyze using a small four-cutter bait fragment encompassing the transcription start site, as in some cases enhancers tend to associate with slightly more downstream or upstream sequences, which may not be encompassed by the four-cutter fragment used in the analysis[7,17,24]. We suggest using a four-cutter as the primary restriction enzyme only when you are refining interactions initially detected by a six-cutter or if interactions have to be investigated within a narrow genomic region. For the secondary restriction enzyme, any four-cutter insensitive to mammalian DNA methylation and with good re-ligation efficiencies can, in principle, be used. We have performed successful 3C-seq experiments using NlaIII, DpnII, HaeIII and MseI. The final combination of primary and secondary restriction enzymes will ultimately depend on their compatibility in terms of generating a suitable bait fragment for the inverse PCR primer design (see below and Box 1). To maximize efficient circularization in the second ligation step, the final bait fragment should be at least 250 bp (ref. 25), although we have succeeded in obtaining good interaction profiles with bait fragments

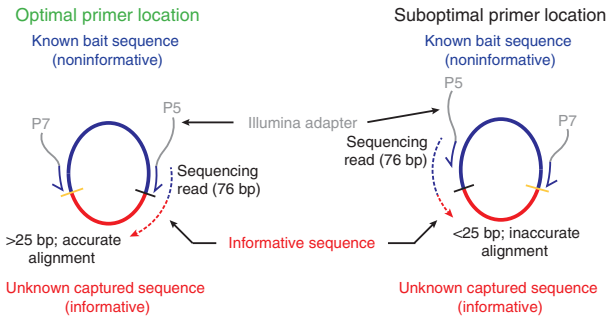


Figure 3: 3C-seq primer design and positioning. Schematic drawing of the location of the inverse PCR primers used to amplify a 3C-seq library. The ring represents a circular DNA molecule composed of the bait fragment (blue) ligated to an unknown captured fragment (red). The two PCR primers are located on the bait fragment next to the restriction sites, with adapters shown as gray overhangs. The P5 primer is located next to the primary restriction site (black dash), and the P7 primer is located next to the secondary restriction site (yellow dash). Illumina sequencing is initiated from the P5 primer and extends into the unknown fragment (dashed arrow). If the P5 primer is located right next to the primary restriction site (within 50 bp), sequence reads generated will be long enough for highly accurate alignment (>25 bp, left). If the distance between the P5 primer and the primary restriction site becomes too large (>50 bp, right), accurate alignment might be compromised.

as small as 120–180 bp (ref. 18; P.K. and E.S., unpublished data). Please note that for some potential interacting fragments both restriction enzyme sites will be very close (<50 bp). When such a fragment ligates to the bait, the resulting sequencing reads might be problematic to align (see TROUBLESHOOTING section). Such a read is not a combination of the bait sequence and a single interacting fragment, as it will also contain sequences from the other side of the bait fragment. By trimming the 3′ end of the reads (PROCEDURE Step 75), a large portion of these fragments can be rehabilitated.

Box 1 | 3C-seq primer design

Two primers, a P5 primer and a P7 primer, need to be designed for each bait fragment of interest:

The P5 primer must be located as close as possible to the primary restriction enzyme site (usually the six-cutter). As only the sequence located after the restriction site is informative for identifying interacting fragments, the distance between the primary restriction enzyme primer and the restriction site itself should be minimized to ensure unambiguous alignment and identification of the interacting fragments (**Fig. 3**). This primer contains the P5 Illumina adapter sequence (5′-AATGATACGGCGACCACCGAACACTCTTTCCCTACAC GACGCTCTTCCGATCT-3′ to be placed upstream of the annealing sequence; **Fig. 3**) from which library sequencing will be initiated. The sequencing reaction starts from the bait fragment, reads through the annealing primer sequence and extends into the unknown captured fragment. To allow more flexibility for primer design and to ensure optimal alignment of the sequences, we use a 76-bp sequencing read length (Step 64).

The second primer, located near the secondary restriction enzyme site (the four-cutter), contains the P7 Illumina adapter sequence (5′-CAAGCAGAAGACGGCATACGA-3′, **Fig. 3**), and although it is required for the inverse PCR and the Illumina sequencing chemistry it is not sequenced (in contrast to paired-end sequencing, for which a different adapter is required). Therefore, the location of the P7 primer with regard to the secondary restriction site is more flexible (within 100 bp of the restriction site).

Actual primer requirements are similar to those used in standard PCR reactions. Oligo length is kept between 17 and 24 nt to facilitate efficient amplification and annealing temperatures are generally chosen between 54 and 59 °C. We regularly use primer design software (DNAMAN 5.0) to check these parameters and to ensure that primers are not prone to form dimers.

Primer design. The 3C-seq library is amplified using primers annealing to the bait sequence, facing outward. Proper design of both primers for the inverse PCR is crucial in the 3C-seq procedure (Box 1 and Fig. 3). Efficiency and reproducibility of the PCR primers are first tested without the addition of the Illumina adapters (Box 2). If performing well, oligonucleotides containing appropriate Illumina adapters are then tested again before being used in the final library amplification PCR before sequencing. For multiplexing purposes, the bait-specific primer sequence itself is used as a bar code to identify reads originating from each individual 3C-seq library. If identical bait-specific libraries need to be sequenced in parallel (e.g., the same promoter for different biological conditions), small bar codes (2–6 nt) may be added to the primers (PROCEDURE Step 62; Box 3).

Controls. 3C-seq data need to be interpreted carefully, as high interaction signals are not necessarily indicators of functionally relevant chromatin coassociations (also see the 'Limitations' section). Furthermore, the PCR amplification step may introduce biases owing to differences in fragment length and GC content, which can affect amplification efficiencies. To ensure proper data interpretation, consider including several control experiments[26]. Whether an interaction is specific for a certain tissue/cell type or whether it correlates with the activity of a specific gene can be tested by analyzing different tissues/cell types or non-expressing cells, respectively. For example, we generally use embryonic stem (ES) cells, cell lines, tissues or FACS-sorted cells that do not express the gene under investigation as controls when investigating promoter-enhancer interactions of an active gene. In addition, using a captured interaction site of interest as bait in a 'reverse experiment' can provide excellent validation of the interaction.

Materials

- Freshly collected tissues, sorted populations of cells and/or cell lines

Caution: Approved governmental and institutional regulations must be followed and adhered to.

- FCS (Sigma-Aldrich, cat. no. A4781)
- DMEM (Gibco, cat. no. 41966)
- Glycine (1 M in PBS; Sigma-Aldrich, cat. no. G7126)

Critical: Glycine stocks should be stored at 4 °C and used cold. They can be stored for a maximum of 6 months.

- PBS (Sigma-Aldrich, cat. no. P4417)
- FCS/PBS (10% (vol/vol))
- Lysis buffer (see Reagent Setup)
- Sodium chloride (NaCl; Sigma-Aldrich, cat. no. S7653)
- Nonidet P-40 substitute (NP-40, Sigma-Aldrich, cat. no. 74385)
- Complete protease inhibitor, EDTA free (Roche, cat. no. 11873580001, see Reagent Setup)
- Milli-Q H_2O
- Collagenase, 2.5% (wt/vol) (Sigma-Aldrich, cat. no. C1639), in PBS
- Formaldehyde, 37% (vol/vol) (Merck, cat. no 1039992500)

Caution: Formaldehyde is toxic.

- Restriction enzymes with 6-bp and 4-bp recognition sites and their corresponding buffers (see INTRODUCTION; Roche or New England Biolabs)
- SDS (20% (wt/vol); Sigma-Aldrich, cat. no. 05030)
- Triton X-100 (20% (vol/vol); Sigma-Aldrich, cat. no. T8787)
- T4 DNA ligation buffer (Roche, cat. no. 10799009001)
- T4 DNA ligase, high concentration (Roche, cat. no. 10799009001)
- Proteinase K (10 mg ml^{-1}, Sigma-Aldrich, cat. no. P2308)
- RNase (10 mg ml^{-1}, Sigma-Aldrich, cat. no. R6513)
- Phenol/chloroform/isoamyl alcohol (25:24:1 (vol/vol/vol); pH 8; Sigma-Aldrich, cat. no. 77617)

Caution: Phenol/chloroform is toxic.

- Glycogen (20 mg ml^{-1}, Roche, cat. no. 10901393001)
- Ethanol (100% (vol/vol) or 70% (vol/vol); Sigma-Aldrich, cat. no. 459844)
- Sodium acetate (2 M, pH 5.6; Sigma-Aldrich, cat. no. S2889)
- Tris-HCl (10 mM, pH 7.5, or 1 M, pH 8.0)
- Liquid N_2
- Agarose electrophoresis gels (0.6% and 1.5% (wt/vol))
- Expand long template system 10× buffer 1 (Roche, cat. no. 11759060001)
- dNTPs (10 mM each)
- Expand long template system DNA polymerase (Roche, cat. no. 11759060001)
- PCR primers (see INTRODUCTION)
- QIAquick gel extraction kit (Qiagen, cat. no. 28706)
- TruSeq SR cluster kit v3-cBot-HS (Illumina, cat. no. GD-401-3001)
- TruSeq SBS kit v3-HS (50 cycles) (Illumina, cat. no. FC-401-3002)
- Python 2.6 (http://www.python.org/)
- Illumina offline base calling software (http://support.illumina.com/sequencing/sequencing_software/offline_basecaller_olb.ilmn)
- NARWHAL (https://trac.nbic.nl/narwhal/)
- Pysam (http://code.google.com/p/pysam/)
- Supplementary analysis scripts (see Supplementary Data; the scripts findSequence.py, regionsBetween.py, alignCounter.py and libutil.py should be extracted to the same directory)

EQUIPMENT

- Cell strainer, 40 µm (BD Falcon, cat. no. 352340)
- Polypropylene centrifugation tubes (Greiner bio-one, cat. no. 188271)
- Safe-Lock 1.5-ml centrifugation tubes (Eppendorf, cat. no. 0030120.086)
- Thermomixer (Eppendorf, cat. no. EF4283)
- Water bath
- Microcentrifuge (Eppendorf, cat. no. 5417R)

- PCR thermocycler (MJ Research, cat. no. PTC-200)
- Spectrophotometer (NanoDrop 2000c, Thermo Scientific)
- Agilent 2100 Bioanalyzer (Aligent Technologies, cat. no. G2938C) with the 7500 DNA chip (cat. no. 5067-1506)
- Illumina HiSeq2000 high-throughput sequencing machine (Illumina)
- Excel spreadsheet software (Microsoft)
- Computer with a minimum of 8 Gb RAM and 1.5 Tb attached storage running a Linux distribution and the software listed above

REAGENT SETUP

- Complete protease inhibitor, EDTA free

Dissolve one tablet in 1 ml of PBS to create a 50× working solution. Store the solution at −20 °C for up to 2–3 months; avoid repeated freeze-thaw cycles.

- Lysis buffer

Prepare the following solution in Milli-Q H_2O: 10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% (vol/vol) NP-40 and 1× protease inhibitor solution.
Critical: Because protease inhibitors degrade quickly in solution, use freshly prepared lysis buffer for each new experiment.

PROCEDURE

Steps 1 - 3: Single-cell preparation and cross-linking
Timing: 1–2 h

1. Obtain single-cell preparations from fresh tissue, FACS-sorted cells or cell lines in 10% (vol/vol) FCS/PBS (see Table 2 for cell types successfully used by us in 3C-seq experiments). Tissues rich in extracellular matrix (e.g., brain) can be treated with collagenase (0.125% (wt/vol) in PBS; incubate the tissues for 30–60 min at 37 °C) first. Filter tissue-harvested cell preparations through a 40-µM cell strainer to obtain single-cell suspensions (see ref. 19). Determine cell concentrations and dilute 0.3 × 106 to 10 × 106 cells (10 × 106 is preferred but substantially fewer starting cells can be used) in 12 ml of culture medium (e.g., DMEM) or 10% (vol/vol) FCS/PBS (15-ml polypropylene tube).

Critical step: Cell preparations need to be single-cell suspensions in order for proper formaldehyde cross-linking to be achieved.

2. Add 649 µl of 37% (vol/vol) formaldehyde to each 15-ml tube (2% (vol/vol) final formaldehyde concentration), and incubate it for 10 min at room temperature while tumbling.

Critical step: 1% (vol/vol) formaldehyde can also be used, especially if digestion efficiencies are suboptimal.

3. Transfer the tubes to ice and add 1.6 ml of cold 1 M glycine (0.125 M final concentration). Immediately proceed with Step 4.

Steps 4 - 16: Cell lysis, nuclei preparation and first restriction enzyme digestion
Timing: 18–20 h

4. Centrifuge the mixture for 8 min at 340g (4 °C) and remove all of the supernatant.

5. Carefully add ice-cold PBS to a volume of 14 ml and resuspend the pellet.

6. Pellet the cells again as in Step 4. Remove all of the supernatant.

7. Carefully resuspend the pellet in 1 ml of cold lysis buffer and add another 4 ml of lysis buffer to obtain a total volume of 5 ml for each tube. Incubate the mixture for 10 min on ice.

8. Centrifuge the mixture for 5 min at 650g (4 °C) to pellet the nuclei.

Pause point: The pelleted nuclei can be washed with PBS, snap-frozen in liquid N_2 and stored at −80 °C for several months.

9. Resuspend the nuclei in 0.5 ml of 1.2× restriction buffer and transfer them to a 1.5-ml Safe-Lock microcentrifuge tube.

10. Place the tubes at 37 °C in a thermomixer and add 7.5 µl of 20% (wt/vol) SDS (final: 0.3% SDS).
    - *Troubleshooting*

11. Incubate the mixture at 37 °C for 1 h while shaking (900 r.p.m.).

12. Add 50 µl of 20% (vol/vol) Triton X-100 (final: 2% Triton X-100).

13. Incubate the mixture at 37 °C for 1 h while shaking (900 r.p.m.).

14. Take a 5-µl aliquot (undigested control sample) of each sample and store it at −20 °C until analysis of digestion efficiency is required (see Step 16).

15. Add 400 U of the selected six-cutter restriction enzyme to the remaining samples and incubate them overnight at 37 °C while shaking (900 r.p.m.).

Critical step: More unconventional primary restriction enzymes with optimal temperatures of 38–50 °C (e.g., ApoI) are also used at 37 °C to avoid partial de-cross-linking of the sample. Prolonged incubation times and/or addition of more enzyme might be required in these cases.

16. Take a 5-µl aliquot (digested control sample) of each sample. At this point, digestion efficiencies can be analyzed by purifying the genomic DNA from the control samples using a standard phenol/chloroform extraction and running it on a 0.6% (wt/vol) agarose gel (see ref. 19). A successful six-cutter restriction enzyme digestion results in a DNA smear with the majority of fragments located between 5 and 10 kb (Fig. 4a).

Steps 17 - 23: Preparation of the 3C library: first ligation and de-cross-linking

Timing: 20–22 h

17. Add 40 µl of 20% (wt/vol) SDS (final: 1.6% SDS) to the remaining sample from Step 15.

18. Incubate the mixture for 20–25 min at 65 °C while shaking (900 r.p.m.).

19. Transfer the digested nuclei to 50-ml centrifugation tubes and add 6.125 ml of 1.15× ligation buffer.

20. Add 375 µl of 20% (vol/vol) Triton X-100 (final: 1% Triton X-100).

21. Incubate the mixture for 1 h at 37 °C in a water bath while shaking gently.

22. Add 100 U of T4 DNA ligase (20 µl of a high-concentration stock) and incubate it at 16 °C for 4 h.

Pause point: The samples can be kept overnight at 16 °C if necessary.

23. Add 30 µl of 10 mg ml^{-1} proteinase K (300 µg in total) and incubate it overnight at 65 °C to de-cross-link the samples.

Steps 24 - 33: Preparation of the 3C library (DNA purification)

Timing: 7–8 h

24. Add 30 µl of 10 mg ml^{-1} RNase (300 µg in total) and incubate the mixture for 30–45 min at 37 °C.

25. Briefly cool the samples to room temperature and add 7 ml of phenol/chloroform/isoamyl alcohol (25:24:1) and shake the samples vigorously.

26. Centrifuge the samples for 15 min at 3,200g (room temperature).

27. Transfer the upper aqueous phase into a new tube and add 7 ml of Milli-Q H_2O. Add 1.5 ml of 2 M sodium acetate (pH 5.6), and then add 35 ml of 100% ethanol.

28. Mix the tubes thoroughly and place them at −80 °C for 2–3 h until the liquid is frozen solid.

29. Directly centrifuge the frozen samples for 45 min at 3,200g (4 °C).

30. Remove the supernatant and add 10 ml of 70% ethanol.

31. Centrifuge the mixture for 15 min at 3,200g (4 °C).

32. Remove the supernatant, air-dry the pellet for ~20 min at room temperature and dissolve the pellet in 150 µl of 10 mM Tris-HCl (pH 7.5) by incubating it for 30 min at 37 °C.

Pause point: This material is referred to as the '3C library' and can be stored at −20 °C for several months.

33. To determine ligation efficiency, run 0.5–1.0 µl of 3C material on a 0.6% (wt/vol) agarose gel. A successful ligation of six-cutter–digested 3C material should result in a single band, running at a similar height as the undigested control sample from Step 14 (Fig. 4b).

Steps 34 - 35: Preparation of the 3C-seq library (determination of DNA concentration and secondary digestion of 3C material)

Timing: 16–18 h

34. If primary digestion and ligation were successful, the 3C library (Step 32) can either be used for 3C-qPCR experiments (see Hagege et al.[19] for a detailed protocol) or be used to prepare the 3C-seq library as described here. First, run an aliquot (e.g., 1 µl) of 3C library DNA alongside a reference sample of species-matched genomic DNA to estimate DNA concentrations. To obtain sharp bands suitable for accurate gel densitometry quantification, a 1.5–2% (wt/vol) agarose gel is used. Optical density (OD) measurements do not provide an accurate estimation of DNA concentrations in 3C library samples.

35. Digest a preferred amount of the 3C library overnight (generally 25–50 µg) with a 4-base recognition restriction enzyme of choice (the four-cutter), at a DNA concentration of 100 ng μl^{-1}, using 1 U of enzyme per µg of DNA. Use buffers and incubation temperatures as recommended in the manufacturer's instructions.

Steps 36 - 56: Preparation of the 3C-seq library (Second ligation and DNA purification)

Timing: 12–13 h

36. Transfer the sample to a 1.5-ml Safe-Lock tube. Add an equal amount of phenol/chloroform/isoamyl alcohol (25:24:1) and mix it vigorously.

37. Centrifuge the mixture for 15 min at 15,800g (room temperature).

38. Transfer the upper phase to a new tube and add 2 µl of 20 mg ml^{-1} glycogen. Add a one-tenth volume of 2 M sodium acetate (pH 5.6), mix the contents and add 850 µl of 100% ethanol.

39. Mix the tubes thoroughly and snap-freeze them in liquid N_2.

40. Directly centrifuge the frozen tubes for 20 min at 15,800g (4 °C).

41. Remove the supernatant carefully and add 1 ml of 70% (vol/vol) ethanol.

42. Centrifuge the mixture for 5 min at 15,800g (4 °C).

43. Remove the supernatant carefully, air-dry the pellet for ~15 min and dissolve the pellet in 100 µl of Milli-Q H_2O by incubating it for 15 min at 37 °C.

44. Analyze 5 µl of the digested DNA on a 1.5% (wt/vol) agarose gel to check digestion efficiency. The resulting type of smear depends on the enzyme used, but the majority of fragments should be <1 kb and are usually between 300 and 500 bp (Fig. 4b).

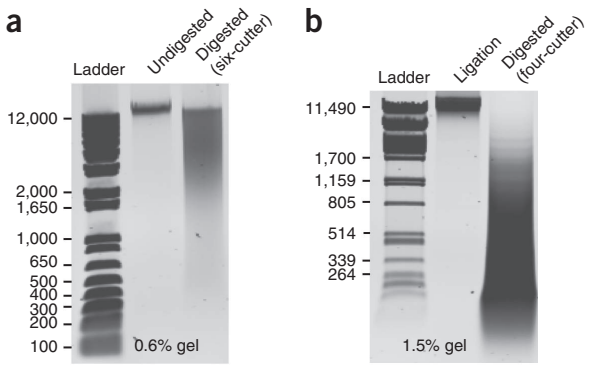


Figure 4: (a) Agarose gel (0.6%, wt/vol) on which an aliquot of undigested (left lane) and digested (right lane) sample (primary restriction digestion, Step 16) was run. A six-cutter was used, showing a typical smear of DNA fragments (a majority of DNA fragments residing between the 12 kb and 4 kb marker bands). (b) After ligation (left lane, Step 33), the DNA smear has returned to a sharp band (~12 kb). Secondary enzyme digestion (four-cutter) of the ligated 3C library typically results in a DNA smear of 2–0.1-kb fragments (1.5% (wt/vol) agarose gel).

45. Transfer the remaining sample to a 50-ml centrifugation tube. Add the components tabulated below and incubate the mixture at 16 °C for 4 h.

Component	Amount per reaction	Final
10× ligation buffer	1.4 ml	1×
T4 DNA ligase (5 U μl^{-1})	40 µl	200 U
Milli-Q H_2O	Up to 14 ml	

Pause point: The samples can be kept overnight at 16 °C if necessary.

46. Add 14 ml of phenol/chloroform/isoamyl alcohol (25:24:1) and shake the mixture vigorously.

47. Centrifuge the mixture for 10 min at 3,200g (room temperature).

48. Split the upper phase into two new 50-ml tubes. Add an equal amount of Milli-Q H_2O to each tube and add 1 µl of 20 mg ml^{-1} glycogen per ml.

Critical step: Increasing the volume before precipitation will greatly reduce the amount of coprecipitating DTT.

49. Add a one-tenth volume of 2 M sodium acetate (pH 5.6), mix the contents and add two volumes of 100% ethanol.

50. Place the tubes at −80 °C for 2–3 h until the liquid is frozen solid.

Pause point: The samples can be kept at −80 °C for several days.

51. Directly centrifuge the frozen tubes for 45 min at 3,200g (4 °C).

52. Remove the supernatant and add 15 ml of 70% (vol/vol) ethanol.

53. Centrifuge the mixture for 15 min at 3,200g (4 °C).

54. Remove the supernatant, air-dry the pellet for ~20 min and dissolve it in 75 µl of 10 mM Tris-HCl (pH 7.5 (per pellet)) by incubating it for 30 min at 37 °C. Thereafter, samples divided over two tubes can be recombined into a single tube.

55. Purify the DNA using the QIAquick gel purification kit according to the manufacturer's recommendations for direct cleanup from enzymatic reactions. Other DNA purification kits can be used, but we have obtained excellent purities with the QIAquick kit.

Critical step: One column can bind a maximum of 10 µg of DNA: use enough columns to avoid overloading and a subsequent loss of material.

56. Determine the DNA concentration of the resulting 3C-seq library using NanoDrop OD measurements.

Box 2 | 3C-seq PCR setup and optimization

As 3C-seq library fragments differ in length and abundance, we use the Expand long template system to minimize any biases resulting from these differences[11]. Bait-specific primers (without adapters) are first tested for proper linearity and efficiency.

1. Test the increasing amounts of 3C-seq library DNA (up to 200 ng) using a 50-µl PCR. Reaction components and conditions are described in PROCEDURE Step 57.
2. Analyze PCR products on a 1.5% (wt/vol) agarose gel, where they should appear as a reproducible smear of DNA fragments, usually showing two prominent bands[11]. These prominent bands are the result of recircularization of the bait fragment in the first ligation step, and of detection of the neighboring fragment owing to incomplete digestion of the primary restriction site on the bait fragment[11].
3. Assess the linear range of the individual primer pairs by quantifying prominent bands in each reaction of the dilution range.
4. Order versions of the primer pairs that perform well, including the P5 and P7 Illumina adapter sequences (**Box 1**). Test these new primers as described in steps 1–3 of **Box 2**.
5. Use successful P5 and P7 primers to prepare 3C-seq samples for sequencing (PROCEDURE Steps 57–60).

Steps 57 - 60: 3C-seq inverse PCR (preparing the sample for Illumina sequencing)
Timing: 5–6 h

57. Perform several PCR reactions (we generally amplify the equivalent of 500–1,000 ng input DNA per bait fragment) using the primers containing the P5/P7 Illumina adapters as overhang using the PCR reaction setup and program tabulated below. The amount of input 3C-seq library DNA used should be the maximum amount for which the PCR reaction is still linear and reproducible (see tables below and Step 58), not exceeding 200 ng per reaction.

Component	Amount per reaction	Final
10× buffer I	5 µl	1×
10 mM dNTPs	1 µl	0.2 mM
25 pmol µl^{-1} forward primer	1 µl	25 pmol
25 pmol µl^{-1} reverse primer	1 µl	25 pmol
Polymerase mix (5 U µl^{-1})	0.75 µl	3.75 U
3C-seq library DNA	Depends on concentration	25–200 ng
Milli-Q H_2O	Add up to 50 µl	

Cycle number	Denature	Anneal	Extend
1	94 °C, 2 min		
2–31	94 °C, 15 s	Primer-specific,1 min	68 °C, 3 min
32			68 °C, 7 min

Critical step: Inverse PCR primers first have to be tested for linearity and reproducibility as described in Box 2 (also see ref. 11), first without and then with the P5/P7 Illumina sequencing adapters attached.

➢ *Troubleshooting*

58. Verify PCR success by running small aliquots (10 µl) of each reaction on a 1.5% (wt/vol) agarose gel.

59. Pool all successful reactions from the same bait fragment and purify the DNA using 2 QIAquick gel purification columns. Elute the columns with 40 µl of Milli-Q H_2O and combine the samples.

60. Verify the purification procedure success by running an aliquot (5–10 µl) on a 1.5% (wt/vol) agarose gel. The sample is now ready to be used for Illumina high-throughput sequencing.

Pause point: The samples can be kept at −20 °C for several months.

Steps 61 - 64: 3C-seq sample pooling and Illumina high-throughput sequencing
Timing: 4 d

61. Quantify the DNA molarity of the individual samples on an Agilent Bioanalyzer with the DNA 7500 chip cartridge according to the manufacturer's instructions. Perform a 'smear analysis' quantification using the Bioanalyzer software.

Critical step: Make sure to use the DNA 7500 chip cartridge, as 3C material contains large (1–5 kb) DNA fragments that will influence DNA molarity and may not be detected using other DNA chip cartridges.

62. Design a pool of 3C-seq samples to be sequenced together in a single lane on the flow cell using the guidelines described in Box 3.

Box 3 | 3C-seq pooling guidelines

The Illumina sequencers use the first four sequenced bases to locate the DNA clusters on the flow cell. When too little variation is present in these first bases, the DNA clusters will not be correctly recognized and base calling will be compromised. The following pooling guidelines are used to ensure that the sequencing process proceeds correctly.

1. Pool *at least* six samples together in a single lane for multiplexing. As one sample can be sequenced in multiple lanes, there is no physical limit as to how many samples can be pooled. We have regularly pooled up to 12 samples in one lane.
2. Ensure that at least one adenine and one thymine base are present in each of the first four cycles of a sample pool. The cycles with the highest intensity of the adenine and thymine bases are used for cluster recognition by the sequencer. Without these specific nucleotides in the first four bases, base calling will be compromised and the sequencing run will fail.
3. Do not pool samples generated with the same bait-specific PCR primer, as sequences derived from these samples cannot be discriminated in the downstream analysis. If pooling of such samples is desired, short bar-code sequences (2–6 nt) will have to be added to the adapter-containing bait-specific primers in the final PCRs (Step 57).

63. Pool the selected samples in equal molarities in a single tube.

64. Proceed with the sequencing procedure as described by the manufacturer in the Illumina TruSeq SR cluster kit and TruSeq SBS manuals. The sequencing procedure can be outsourced to a sequence service provider. We generally use 76-bp single-read sequencing; paired-end sequencing is not required for 3C-seq.

Critical step: When loading the flow cell, aim for a cluster density of 750,000–850,000 clusters per mm^2. In our case, this is usually achieved with a final template DNA concentration of 9 pM.

Critical step: Ensure that the total number of sequencing cycles exceeds the sum of the bait-specific sequence length and a minimum of 36 bases for optimal alignment of the unknown interacting fragments.

Steps 65 - 79: Initial data processing
Timing: 1–2 d

65. Copy the whole run folder generated by the Illumina sequencer to the storage on the Linux

computer.

66. Open a terminal on the Linux computer and enter the commands described after the > signs.

67. Convert the binary output from the sequencer to text files in the Qseq format by using the BclToQseq scripts included in the Illumina Offline Basecaller (available at the Illumina website http://www.illumina.com/):

> cd Illumina_Run_Folder/Data/Intensities/BaseCalls

> /path_to_OLB/bin/setupBclToQseq.py --in-place –b.

> make –j 6

68. Determine the bait-specific sequences for de-multiplexing. Note that this also includes the primer, the primary restriction site and any sequence in between. To obtain the highest yield while still retaining high specificity, de-multiplexing is performed using only 6 bases instead of the entire bait-specific sequence. The first set of 6 bases that differ for 2 or more bases from the other bait sequences are used for de-multiplexing.

Critical step: Record the unique 6-bp bait-specific sequences (6-bp-bait) and their positions (6 bp-bait-pos) in the bait for each sample.

69. Determine the number of bases to trim from the 5′ and the 3′ ends of the reads as described in Steps 70–75. This procedure is performed in Microsoft Excel.

Critical step: The 5′ trimming is crucial, as the remaining bait-specific sequences will prevent the read from aligning to the reference sequence (Fig. 3). The 3′ trimming prevents the loss of short interacting fragments (see Experimental design).

70. First, extend the bait-specific primer sequence with the genomic sequence up to and including the primary restriction site.

71. Extend the bait-specific primer sequence with the genomic sequence up to and including the primary restriction site.

72. Subtract the forward Illumina P5 adapter sequence from the 5′ end of this sequence (Box 1).

73. Count the number of bases in the resulting sequence using the *len()* function to obtain the number of bases to trim from the 5′ end of the read (*n5trim*).

74. Subtract *n5trim* from the read length.

75. Subtract 36 bases from the result of Step 74 to obtain the number of bases to trim from the 3′ end (*n3trim*).

76. Create a NARWHAL[27] sample sheet (Supplementary Table 1) for the lanes that contain the 3C-seq samples. In this sample sheet, use any profile that runs BOWTIE[28] with the --best option. To de-multiplex, several options need to be set in the sample sheet: the bar code-read field is set to 1; the bar code-start field is set to the 6-bp-bait-pos; the bar code field is set to the 6-bp-bait sequence. For the trimming, the following options are added to the options field of the sample sheet to trim the sequences:
--trim5=*n5trim*,--trim3=*n3trim*.

77. Copy the NARWHAL sample sheet to the Linux computer.

78. (Optional) When the flow cell does not exclusively contain 3C-seq samples, it might be necessary to analyze only specific lanes. This can be achieved by setting up a directory with only the Qseq files for the specific lanes to be analyzed. This can be performed as follows, with *i* as the lanes to be analyzed:

```
> mkdir MyLanes/

> ln –s /full_path_to_qseq_folder/s_[i]_1_*_qseq.txt MyLanes/
```

79. Run NARWHAL using the following command:

```
> narwhal.sh –s samplesheet.txt Qseq_folder output_folder
```

TABLE 3 | Troubleshooting table.

Step	Problem	Possible reason	Solution
10	Formation of aggregates after addition of SDS to the restriction buffer	Too many nuclei are used or the nuclei are of poor quality	Dilute the material 2–4 times in 1.2× restriction buffer containing 0.3% (wt/vol) SDS. For future experiments, ensure gentle handling of the cells and nuclei. A more stringent lysis buffer and/or Douncing step can also be beneficial. If persistent, consider starting with fewer cells in future experiments
16	Poor primary digestion efficiency	Formaldehyde concentrations used are too high for the enzyme; the enzyme is not compatible with the 3C protocol and/or extensive nuclei aggregation	Lower formaldehyde concentrations (e.g., 1% instead of 2% (vol/vol)) or increase Triton X-100 concentration in Step 12. Alternatively, consider changing to a different enzyme. If nuclei are forming large aggregates, see Step 10 troubleshooting for advice
57	Poor PCR linearity, reproducibility or PCR failure	PCR conditions or design are suboptimal	Ensure that the correct primer T_m is used. Further optimizing the T_m using a gradient can be beneficial. Often, simply redesigning the 3C-seq primers will greatly improve PCR success
	Primer dimer formation	PCR conditions or design are suboptimal	See above. If primer dimer formation specifically occurs after addition of the P5/P7 adaptors, DNA purification kits with a >100-bp cutoff can be used to remove dimers before sequencing
79	Fewer than expected sequence yield for a particular sample	Unanticipated bait-specific sequence	Compare the list of expected barcodes to the most abundant sequences. To generate a list with the most abundant barcode sequences from a FastQ file, the following Linux command-line code can be used: `> grep '^[ACTGN]\+$' in.fastq | sed 's/^\(.\{6\}\).*/\1/g' | sort | uniq –c | sort –nr | head –n 30` Cross-reference unexpected highly abundant sequences with the expected primers and if possible assign these reads to a sample. Re-do de-multiplexing with the updated barcodes
	Low mapping percentage after sequencing	Primer dimers present in 3C-seq sample or the secondary restriction site occurs directly after the primary restriction site in the most abundant target fragments	Obtain all the non-aligning sequences from the BAM file: `> samtools view aln.srt.bam | grep -P '^\S+\t\d+\t\*.*$' > not_aligned.aln` Check these sequences for subsequences of the primers used in the amplification. Determine whether these sequences contain the restriction site for the secondary restriction enzyme. This issue occurs more frequently with increasing read-length. For this reason, we strongly recommend using the 3′ trimming procedure from Steps 70–75. If after trimming the target sequence is shorter than 25 bp, the secondary restriction enzyme needs to be changed in order for the read to be aligned properly

(continued)

TABLE 3 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
84	Complete absence of reads at expected sites of interaction	The fragment expected to interact with the bait is <36 bp	Further extend the 3′ trimming procedure or use a different six-cutter/four-cutter combination
		The genome assembly has changed (updated)	Reanalyze older data sets using the proper version of the genome assembly. This may be crucial when recent data sets need to be compared with older ones
	Weak 3C-seq interaction signals	Poor signal-to-noise ratio	Consider using a double cross-linking procedure by using ethylene glycol bis-succinimidylsuccinate treatment before formaldehyde as described in Lin *et al.*[34]

After the alignment, NARWAL will generate a PDF reporting the total number of reads generated, the percentage successfully aligned reads, the read distribution across the chromosomes, edit rates and duplication rates[27]. Successful 3C-seq experiments should have high duplication rates (>95%), with a majority of reads (>50%) mapped to the chromosome on which the bait is located.

➢ *Troubleshooting*

Steps 80 - 84: Bioinformatics and initial data visualization
Timing: 2 h

80. After the initial data processing, a restriction map of the genome needs to be generated as described in Steps 80–82. First, Search the genome for restriction sites using the findSequence.py script (Supplementary Data). This script will generate a BED file containing all the occurrences of a given sequence in the genome.

> python findSequence.py –f genome.fasta –s primary_restriction_sequence –b occurrences.bed

81. Create a BED file containing the regions between the restriction sites by using the regionsBetween.py script (Supplementary Data):

> python regionsBetween.py –i occurrences.bed –s chromsizes.txt –o regions.bed

82. Sort the regions with the BEDtools[29] sort command:

> bedtools sort –i regions.bed > sorted_regions.bed

83. Count the reads per target fragment using the alignCounter.py tool (Supplementary Data). The count result is a table that can be loaded into other tools such as R.

> python alignCounter.py –b aln.srt.bam –r sorted_regions.bed –o output_table.txt

84. Convert the read count tables to BED files using the command below. These BED files can be loaded into a variety of genome browsers including the UCSC Genome Browser (http://genome.ucsc.edu/).

> gawk ′/^[#]/{ if($4 > 0){print $1 ″\t″ $2 ″\t″ $3 ″\t″ $4 ;}; }′ output_table.txt > output_table.bed

➢ *Troubleshooting*

Troubleshooting

Troubleshooting advice can be found in Table 3.

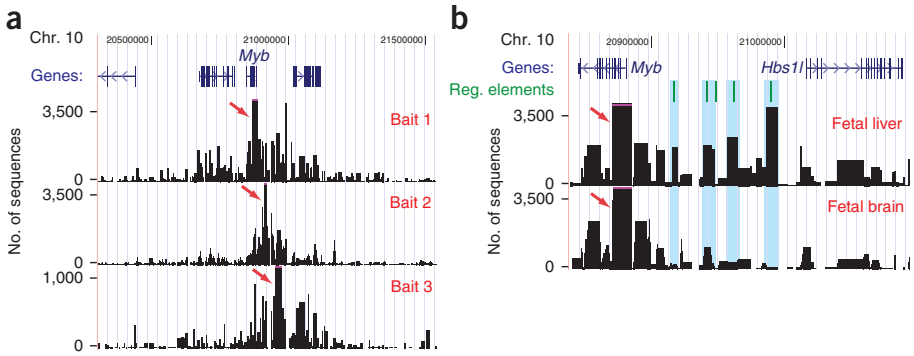


Figure 5: Typical interaction profiles obtained from a multiplexed 3C-seq experiment. (a) 3C-seq interaction profiles in mouse fetal liver cells shown for three bait fragments in the *Myb* locus[17] (1.2-Mb region shown). Bait signals are depicted by an arrow. (b) 3C-seq interaction profiles generated from both mouse fetal liver and brain using the *Myb* promoter as bait (shown is a ~250-kb region encompassing the Hbs1-like (*Hbs1l*) neighboring gene). *Myb* is highly expressed in fetal liver cells, but expression is much lower in fetal brain cells. Several fetal liver–specific interactions are located within an intergenic region containing several regulatory (Reg.) elements (green lines and blue shading)[17]. Bait signals are depicted by an arrow. Data were visualized using the UCSC genome browser. All animal work was approved by the Netherlands Animal Experimental Committee (DEC) and the Institutional Ethical Review Board of Erasmus Medical Center, and was carried out according to institutional and national guidelines.

Timing

Steps 1–3, single-cell preparation and cross-linking: 1–2 h
Steps 4–16, cell lysis, nuclei preparation and first restriction enzyme digestion: 18–20 h
Steps 17–23, preparation of the 3C library: first ligation and de-cross-linking: 20–22 h
Steps 24–33, preparation of the 3C library: DNA purification: 7–8 h
Steps 34 and 35, preparation of the 3C-seq library: determination of DNA concentration and secondary digestion of 3C material: 16–18 h
Steps 36–56, Preparation of the 3C-seq library: second ligation and DNA purification: 12–13 h
Steps 57–60, 3C-seq inverse PCR: preparing the sample for Illumina sequencing: 5–6 h
Steps 61–64, 3C-seq sample pooling and Illumina high-throughput sequencing: 4 d
Steps 65–79, initial data processing: 1–2 d
Steps 80–84, bioinformatics and initial data visualization: 2 h

Anticipated results

After sequencing and data processing, the resulting BED files (Step 84) can be visualized in a genome browser (e.g., UCSC genome browser, http://genome.ucsc.edu/). Careful attention should be given to the particular version of the genome that is used for analysis, especially when different experiments are compared. Several simple but important checks can provide information on whether the 3C-seq experiment was successful, which are automatically provided during initial data processing (Steps 65–79) by the NARWAL software[27]. The PDF file provided contains statistics on the chromosomal location of the aligned reads and the duplication percentage. These are important metrics for the initial validation of a 3C-seq experiment: the vast majority (>50%) of reads are usually found in cis (i.e., on the same chromosome), and as 3C-seq profiles consist of stacked reads the duplication percentage should be >95%. Typical alignment percentages are above 70%, although this can vary considerably between different primer sets. Lower percentages are often caused by the sequencing of primer dimers present in the PCR samples or failure to align reads coming from the (in general) most abundant interactions (the bait fragment itself and the neighboring fragment, see Box 2 and Table 3). However, low alignment percentages can still provide informative data, as long as the total number of aligned reads is high enough (>1 million reads[30]) and read distribution is as expected (see below and Fig. 5). After uploading the BED output file (Step 84) in a genome browser, interactions with the chosen bait

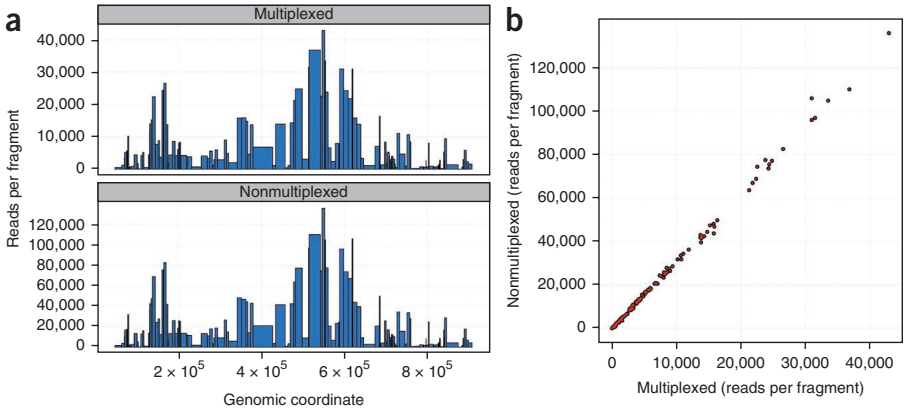


Figure 6: Comparison of interactions detected for the same 3C-seq sample after single or multiplexed library sequencing. (a) Interaction profiles around the bait fragment for a 3C-seq sample after multiplexed (top) or nonmultiplexed (bottom) library sequencing, showing highly similar profiles. (b) Scatter plot comparing read counts for 146 fragments around the bait fragment between nonmultiplexed and multiplexed data sets.

fragments can be observed. Signals are represented as bars (Fig. 5), the width of which is determined by the size of the actual restriction fragment. The height of the bars represents the number of reads found on the fragment and is a measurement of the frequency of interaction with the bait fragment. The highest signal density is always found around the viewpoint (typically ~40% of all reads are located within 1 Mb of the bait), with the two most abundant interactions being the bait and its neighboring fragment (Box 2). Signal intensity tends to rapidly decline with increasing genomic distance from the bait (a classic characteristic of 3C and its derivatives, see refs. 11,26), resembling a bell-shaped distribution around the bait (Fig. 5a). The majority (>75%) of cis interactions are normally found within a 1-Mb window around the bait, although bait fragments within highly complex genomic structures (e.g., immunoglobulin loci) can produce profiles that deviate from this general picture[18]. Interactions found in trans (generally about 40–50% of the reads) often show low interaction frequencies and appear to be randomly scattered around the genome. Trans-interaction signals therefore need to be interpreted with caution, as their reproducibility may appear questionable in a number of cases. However, several studies have begun to probe their functional relevance in specific cases, in particular in light of chromosomal translocations, and showed correlation between physical proximity and sites of recombination, indicating that physical proximity in trans may be relevant[31,32].

Multiplexing 3C-seq samples greatly increases the technique's throughput and results in a substantial cost reduction. Even though the total number of reads is lower in a multiplexed sample compared with a nonmultiplexed sample, interaction patterns remain almost identical (Fig. 6). Thus, multiplexing 3C samples seems to have little effect on the resulting interaction profiles (Fig. 6).

Further validation of detected interactions can be obtained by complementary experiments (e.g., 3C-qPCR, FISH) or by performing new 3C-seq experiments with these interactions as bait (a 'reverse experiment', see 'Controls' section of INTRODUCTION). Functional interpretation of 3C-seq profiles is often desired and requires correlation with other data sets, usually transcription factor binding and/or histone modification patterns for the locus of interest. When using 3C-seq to explore the regulatory elements in close proximity to a gene, strong interaction signals can often be positively correlated to the binding of transcription factors and the presence of specific histone modifications[17]. Performing 3C-seq experiments in different cell or tissue types can further provide valuable information on the tissue specificity of interactions and whether their presence can be correlated to differences in gene expression or protein binding (Fig. 5b). The 3C-seq data can also be further processed using dedicated tools and scripts (S.Thongjuea, R.S., F.G., E.S. and B. Lenhard, unpublished data, and ref. 12) for more in-depth analysis.

Acknowledgments

We thank A. van der Sloot, Z. Ozgur, E. Oole, M. van den Hout, F. Sleutels, S.Thongjuea and B. Lenhard for their help in sample processing, bioinformatics pipeline development and data analysis. R.S. received support from the Royal Netherlands Academy of Arts and Sciences (KNAW). P.K. was supported by grants from ERASysBio+/FP7 (project no. 93511024). E.S. was supported by grants from the Dutch Cancer Genomics Center, the Netherlands Genomics Initiative (project no. 40-41009-98-9082) and the French Alternative Energies and Atomic Energy Commission (CEA). This work was supported by the EU-FP7 Eutracc consortium.

Supplementary information

Supplementary information is available at the Nature Protocols website: Supplementary Data (4 python files) and Supplementary Table 1.

Contributions

R.S. and R.-J.P. adapted and optimized the protocol and library preparation for Illumina sequencing. R.S., P.K., A.v.d.H. and J.Z. used, developed and troubleshot the technique. C.K. optimized procedures for library sequencing, and R.B. developed the informatics pipeline for data processing and analysis. W.v.I., F.G., K.S.W. and E.S. supervised the projects, and participated in technology design and discussions. R.S., P.K., R.B., W.v.I., F.G., K.S.W. and E.S. drafted the manuscript.

References

1. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).
2. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385 (2012).
3. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature 489, 109–113 (2012).
4. Splinter, E. & de Laat, W. The complex transcription regulatory landscape of our genome: control in three dimensions. EMBO J. 30, 4345–4355 (2011).
5. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. Cell 144, 327–339 (2011).
6. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. 12, 283–293 (2011).
7. Stadhouders, R. et al. Transcription regulation by distal enhancers: who's in the loop? Transcription 3, 181–186 (2012).
8. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science 295, 1306–1311 (2002).
9. Gondor, A., Rougier, C. & Ohlsson, R. High-resolution circular chromosome conformation capture assay. Nat. Protoc. 3, 303–313 (2008).
10. Sexton, T. et al. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. Nat. Protoc. 7, 1335–1350 (2012).
11. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. 38, 1348–1354 (2006).
12. van de Werken, H.J. et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods 9, 969–972 (2012).
13. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat. Protoc. 2, 988–1002 (2007).
14. Fullwood, M.J. et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature 462, 58–64 (2009).
15. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009).
16. Soler, E. et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev. 24, 277–289 (2010).
17. Stadhouders, R. et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. EMBO J. 31, 986–999 (2012).
18. Ribeiro de Almeida, C. et al. The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. Immunity 35, 501–513 (2011).
19. Hagege, H. et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat. Protoc. 2, 1722–1733 (2007).
20. Naumova, N., Smith, E.M., Zhan, Y. & Dekker, J. Analysis of long-range chromatin interactions using chromosome conformation capture. Methods (2012).
21. Ecker, J.R. et al. Genomics: ENCODE explained. Nature 489, 52–55 (2012).
22. Dostie, J. & Bickmore, W.A. Chromosome organization in the nucleus—charting new territory across the Hi-Cs. Curr. Opin. Genet Dev. 22, 125–131 (2012).
23. Comet, I., Schuettengruber, B., Sexton, T. & Cavalli, G. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. Proc. Natl. Acad. Sci. USA 108, 2294–2299 (2011).
24. Jing, H. et al. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. Mol. Cell 29, 232–242 (2008).
25. Rippe, K., von Hippel, P.H. & Langowski, J. Action at a distance: DNA-looping and initiation of transcription. Trends Biochem. Sci. 20, 500–506 (1995).
26. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat. Methods 3, 17–21 (2006).
27. Brouwer, R.W., van den Hout, M.C., Grosveld, F.G. & van Ijcken, W.F. NARWHAL, a primary analysis pipeline for NGS data. Bioinformatics 28, 284–285 (2012).
28. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

29. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
30. van de Werken, H.J. et al. 4C technology: protocols and data analysis. Methods Enzymol. 513, 89–112 (2012).
31. Hakim, O. et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature 484, 69–74 (2012).
32. Zhang, Y. et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148, 908–921 (2012).
33. Visser, M., Kayser, M. & Palstra, R.J. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 22, 446–455 (2012).

Chapter 4

r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data

Supat Thongjuea[1,2]*, Ralph Stadhouders[3]*, Frank G. Grosveld[3,4], Eric Soler[3,4,5]† & Boris Lenhard[6,7,8]†

[1]Computational Biology Unit, Uni Computing, Uni Research AS, Bergen, Norway.
[2]Department of Molecular Biology, University of Bergen, Bergen, Norway.
[3]Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands.
[4]Center for Biomedical Genetics and Cancer Genomics Center, Erasmus Medical Center, Rotterdam, The Netherlands.
[5]Laboratory of Hematopoiesis and Leukemic Stem Cells (LSHL), CEA/INSERM U967, Fontenay-aux-Roses, France.
[6]Department of Molecular Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, London, UK.
[7]MRC Clinical Sciences Centre, Hammersmith Hospital Campus, London, UK.
[8]Department of Informatics, University of Bergen, Bergen, Norway

***These authors contributed equally.**
†Corresponding authors.

Published in:
Nucleic Acids Research
2013; 41:e132

Abstract

The coupling of chromosome conformation capture (3C) with next-generation sequencing technologies enables the high-throughput detection of long-range genomic interactions, via the generation of ligation products between DNA sequences, which are closely juxtaposed *in vivo*. These interactions involve promoter regions, enhancers and other regulatory and structural elements of chromosomes and can reveal key details of the regulation of gene expression. 3C-seq is a variant of the method for the detection of interactions between one chosen genomic element (viewpoint) and the rest of the genome. We present r3Cseq, an R/Bioconductor package designed to perform 3C-seq data analysis in a number of different experimental designs. The package reads a common aligned read input format, provides data normalization, allows the visualization of candidate interaction regions and detects statistically significant chromatin interactions, thus greatly facilitating hypothesis generation and the interpretation of experimental results. We further demonstrate its use on a series of real-world applications.

Introduction

The availability of complete sequenced genomes and increasingly deep coverage of transcriptomes has led to the successful annotation of protein-coding genes and a growing number of non-coding RNA genes in eukaryotic genomes. The mechanisms involved in regulating these genes in different cell types, in various developmental and differentiation processes, and under different environmental conditions are under intensive investigation, recently accelerated by high-throughput methods for the detection of promoters and regulatory elements (1–3). One of the key tasks in integrating data on gene expression with the location and activity of regulatory elements is to elucidate which regulatory elements interact with which gene promoters, and with which other regulatory elements, in a particular cellular context. Much of the early progress in studying the regulatory elements that act directly on distant target genes via physical interactions was made using DNA fluorescence in situ hybridization (4). However, DNA fluorescence in situ hybridization can only be used for a limited number of DNA loci at a time, and it provides only low-resolution data. The advent of the chromosome conformation capture (3C) technique (5), which generates novel ligation products between DNA sequences that are closely juxtaposed in the nuclear space *in vivo*, has led to many long-range genomic interactions detected at high resolution. A key study during the development of 3C showed that the looped conformation between the β-globin genes and the locus control region (LCR) was specific to erythroid cells where the genes are expressed, suggesting that promoter–enhancer contacts may be required for transcriptional regulation (6). The 3C method has been widely used to detect chromosomal interactions in mammalian cells. However, this technique is still low-throughput, as it relies on locus-specific polymerase chain reaction (PCR) primers and can only be used to interrogate chromatin interactions between pairs of pre-selected sequences. Therefore, many efforts have been made to develop protocols for high-throughput 3C-based analyses that allow the identification of many interactions in parallel [for review see (7)]. The resulting methods include for instance (i) 3C-on-chip (4C) (8) and 4C-seq (9), which can be used to identify the genome-wide interactions with a specific fragment of choice (a 'viewpoint'), (ii) 3C-carbon-copy (5C) (10,11), which probes interactions with many viewpoints within a confined genomic region (typically ~1 Mb), or (iii) Hi-C (12), the latter being able to identify interactions between all genomic sites. Each of these methods has advantages and disadvantages, and the specific choice of method depends on the type of question to be answered.

We have previously developed a 3C-seq protocol (13,14) based on an adaptation of the 4C-method (8) to next-generation Illumina sequencing. This protocol generates a vast amount of data consisting of millions of reads from regions of genomic interaction and requires a set of bioinformatics methods and tools to facilitate data preprocessing and data analysis, interpretation and visualization of candidate interaction regions. Currently, there are few tools available for 3C-seq data analysis (9,15). These tools only provide window-based analysis methods, which have the disadvantage of using an arbitrary window size that might limit the identification of interaction regions to within a certain size range. In addition, these tools do not facilitate the analysis of replicate experiments. To address these needs, we have developed an R/Bioconductor package called r3Cseq, a publicly available bioinformatics software package for 3C-seq studies, to perform the analysis of data generated by 3C-seq technology. The package provides a comprehensive workflow that starts with the aligned reads and ends with an interpretable visualization of regions of interaction. It can analyze data from various experimental designs, with or without a control experiment, and it supports in-depth data analysis of replicate experiments. It enables 3C-seq data normalization and statistical analysis for

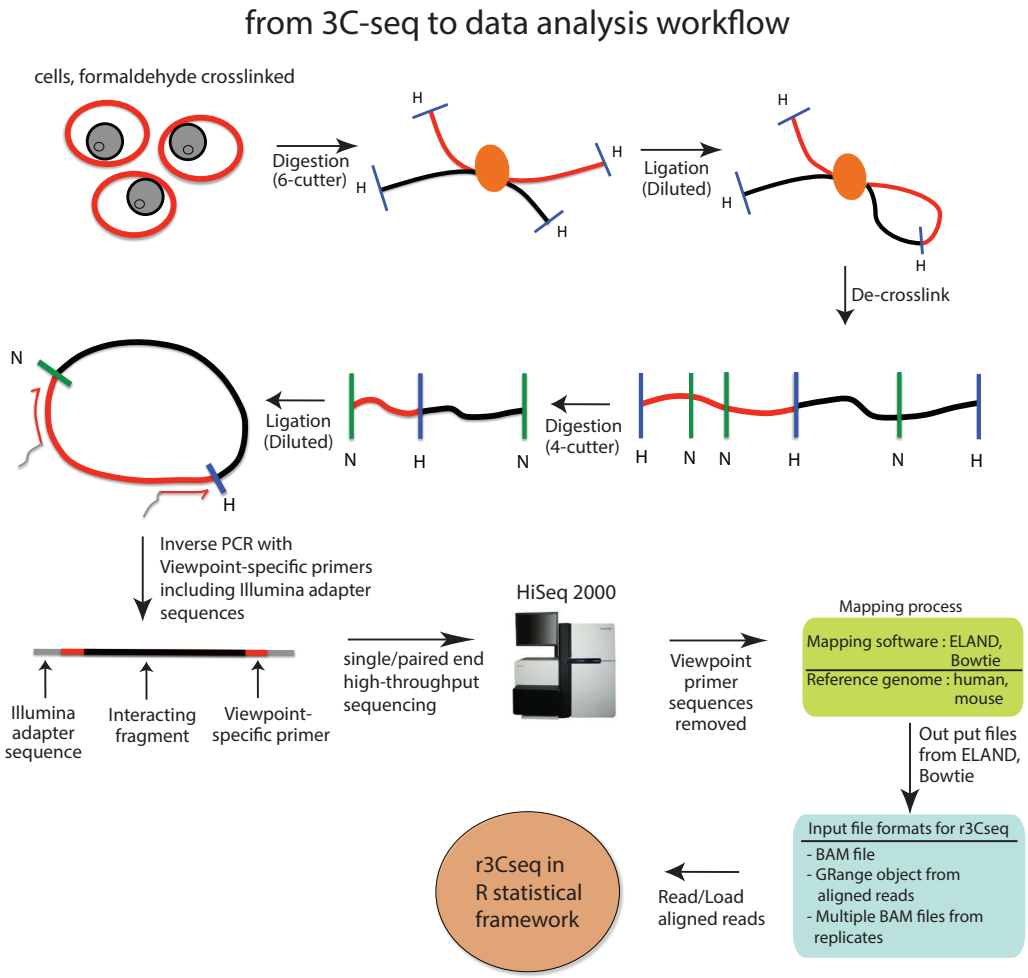


Figure 1. 3C-seq experimental procedures and data analysis workflow. Formaldehyde cross-linked chromatin is digested with a six-cutter restriction enzyme and ligated under dilute conditions. After de-cross-linking, DNA is digested with a four-cutter enzyme and again ligated under dilute conditions to create small circular fragments representing individual ligation events. Inverse PCR using viewpoint-specific primers containing Illumina sequencing adapters is used to generate a viewpoint-specific 3C-seq library. After high-throughput sequencing, reads are trimmed and mapped to the reference genome, after which they are loaded into the r3Cseq software.

the identification of cis and trans interactions (i.e. interactions between regions on the same chromosome and interactions between different chromosomes, respectively), using both restriction fragment-based and window-based methods. These functions will allow scientists to compare different ways of analyzing their data set and select the most suitable analysis for the interpretation of their data. Finally, r3Cseq produces a range of plots specifically designed for the visualization of genomic regions that physically interact with the selected genomic regions of interest. The output generated by r3Cseq consists of simple text and bedGraph (16) files compatible with visualization using other tools, such as the UCSC Genome Browser (16) and IGV (17).

Materials and methods

Principles of the 3C-seq procedure and r3Cseq data analysis workflow

The 3C-seq experimental procedure is outlined in Figure 1, and the subsequent r3Cseq data analysis workflow is shown in Figure 2. Isolated cells are treated with a cross-linking agent to preserve *in vivo* nuclear proximity between DNA sequences. The DNA isolated from these cells is then digested using a primary restriction enzyme, typically a 6-bp cutting enzyme, such as HindIII, EcoRI or BamHI. The digested products are then ligated under diluted conditions to favor intra-molecular over inter-molecular ligation events. This digested and ligated chromatin yields composite sequences representing (distal) genomic regions that are in close physical proximity in the nuclear space. The digested and ligated chromatin is then de-cross-linked and subjected to a second restriction digest using a four-cutter (e.g. NlaIII or DpnII) as a secondary restriction enzyme to decrease the fragment sizes. The resulting digested DNA is then ligated again under diluted conditions, creating small circular fragments. These fragments are inverse PCR amplified using primers specific for a genomic region of interest (e.g. promoter, enhancer or any other element potentially involved in long-range interactions), termed the 'viewpoint'. The amplified fragments are then sequenced using massively parallel high-throughput sequencing. The 3C-seq procedure produces DNA molecules consisting of viewpoint-specific primers followed by sequences derived from the ligated interacting fragments. These need to be trimmed *in silico* to remove the primer and viewpoint sequence, thus leaving only the captured sequence fragments for mapping (14). After trimming, reads are mapped against a reference genome using alignment software, such as Bowtie (18).

Our r3Cseq package has been developed in the R statistical framework (19) as part of Bioconductor (20). It uses binary alignment/map (BAM)-aligned read files as input (21), which are generated by commonly used alignment software and carries out operations, such as class initialization, counting aligned reads per restriction fragment or per window size, read count normalization, statistical analysis of interactions in both cis and trans, data visualization and data export of the identified contacting regions. Figure 2 shows the main features and the sequential steps of the r3Cseq pipeline.

Data normalization

Current normalization methods for next-generation sequencing data have shown that the read count distribution per region observed in RNA-seq, chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq) and Cap Analysis of Gene Expression (CAGE) experiments approximates a power-law distribution (22,23). To investigate whether this is also the case for 3C-seq data, we analyzed 11 published samples of 3C-seq data generated using different mouse cell types, restriction enzymes and viewpoints (22,23). We observed that, for all samples, read count distributions per restriction fragment and per 5-kb window size approximately fit a power-law distribution (Supplementary Figures S1A and S2A). The slopes of the power-law curves were similar across samples, whereas the read counts varied depending on the sequencing depth. We, therefore, adapted a method originally developed for normalizing deepCAGE data (23). For each sample, we fitted the reverse-cumulative distribution of reads per region to a power-law distribution. To do this, we first filtered out data to remove regions with <50 read counts. We also excluded the viewpoint from the analysis. A frequency table with the distribution of read counts per region was then generated for each sample. This frequency table served as the input for a simple linear regression model to obtain the slope and offset values for all samples. As expected, the offset values significantly vary depending on sequencing depth, whereas the fitted slope values vary within a small range. The fitted slope value across all samples was −1.35 (±0.2) on average. To normalize the read count per restriction fragment or per window size, we developed R functions to implement the formula described in the deepCAGE method by choosing a power-law reference distribution with an exponent of $\alpha = -1.35$ and an $n0 = 1$ million offset. The normalization functions are used to transform the read counts from all samples into normalized reads per million (RPM). Supplementary Figures S1C and S2C show the read count distribution after normalization. We also implemented a function to calculate a simpler defined RPM measure for genomic regions, using the number of aligned reads observed at the particular restriction fragment or window divided by the total number of the aligned reads, multiplied by 1 million. Supplementary Figures S1B and S2B show the read count distributions after normalization using this simple RPM calculation. As expected, the reverse-cumulative fitted values from the power-law distributions revealed a better fit of the normalized values for all samples as compared with the simple RPM calculation. In addition, plots depicting the log2 intensity ratio (M) versus the average log2 intensity values (A) between two samples in different experimental conditions exhibited better scaling. Here, the loess red line is close to $M = 0$ when using a reverse-cumulative

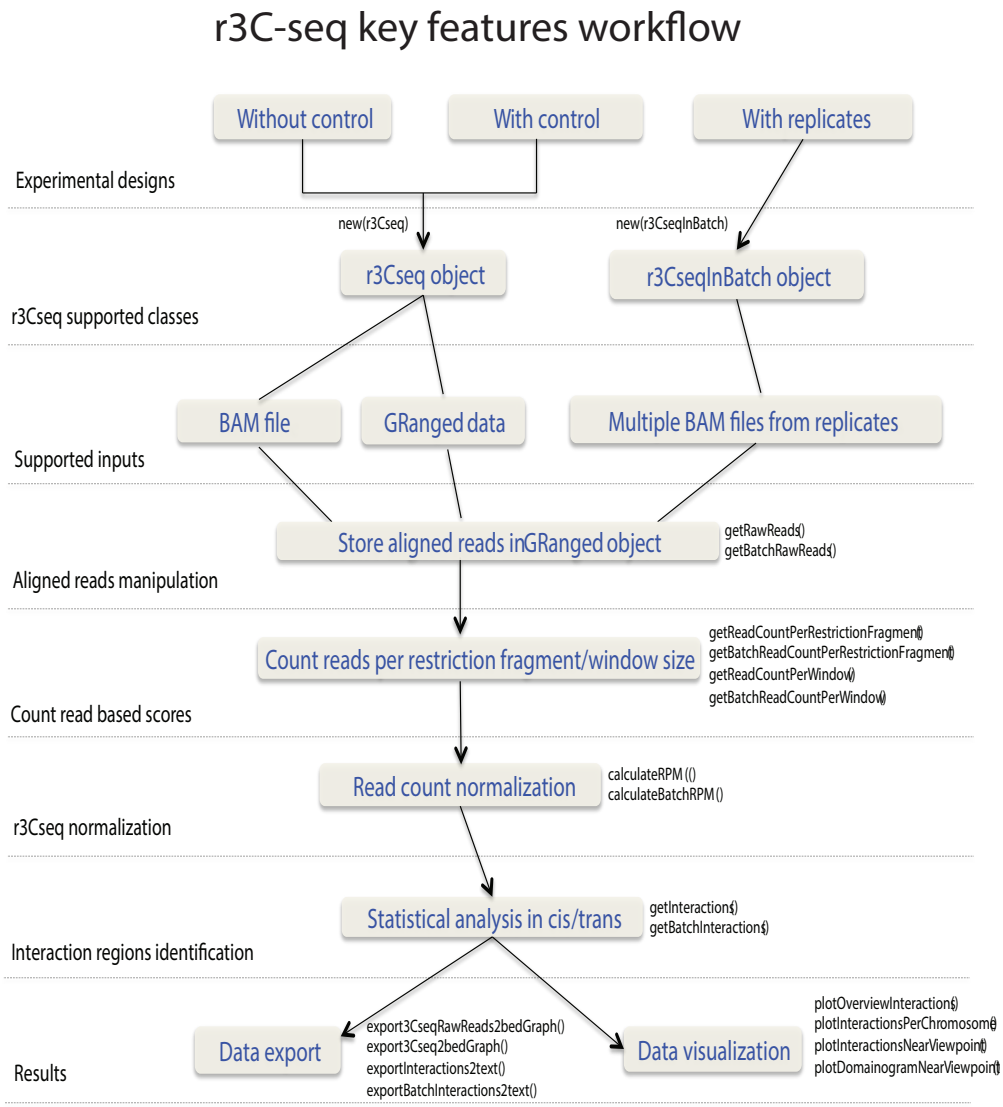


Figure 2. A summary of the r3Cseq analysis pipeline. The main features and the sequential order of operations are shown in the flow chart. In-depth discussion of the different operations and functions can be found in the 'Materials and Methods' and 'Results' sections.

fitting normalization to fit the global dependence between the M–A values, when compared with either no normalization or the simple RPM normalization (Supplementary Figure S3A and B). Furthermore, the log2 ratio of interaction regions between two samples located close to the viewpoint (within ±200 kb, red dots) is not strongly affected after applying the reverse-cumulative fitted values normalization, as the majority of these are in a range of ±3 of the log2 ratio. We, therefore, used the normalized values fitted by the reverse-cumulative of the power-law distribution as the quantitative interaction signals for the fold change calculations to compare interaction intensities between two experimental conditions (see later in the text). The reverse-cumulative fitted values of the power-law normalization method described in this study were implemented as an improvement over the simple RPM normalization that is most often used in count data analysis to remove bias because of unequal sequencing depth. In this study, we demonstrated

that this method performs better than those not applying normalization or those applying the simple RPM normalization. However, normalization techniques are still immature for most types of count data generated by next-generation sequencing technologies. Although in-depth development of such methods for 3C-seq is beyond the scope of this study, the r3Cseq package can easily be expanded with new normalization methods as they become available.

Identifying cis-interactions from 3C-seq experiments

Published 3C-based studies (8,9,24,25) have shown that interaction intensities are highest around the viewpoint, as DNA sequences near the viewpoint have an increased chance of being non-specifically tethered to the viewpoint during chromatin cross-linking. Interaction signals gradually decrease with increasing distances away from the viewpoint and can only be sporadically captured on other chromosomes. To determine the significant interaction regions of a given viewpoint, we applied a background scaling method to correct for interactions that are simply a consequence of short genomic distance to the viewpoint. We determined the relationship of 3C-seq signals of genomic regions located on the cis chromosome by ranking the read count per region based on the relative distance to the viewpoint. The non-parametric regression cubic smoothing spline algorithm implemented in R was then applied with smoothing parameter set between 0.06 and 0.4 (which can be changed by the user). The software uses the smoothing parameter 0.1 by default, as this value exhibits the most suitable steady degree of smoothing (Supplementary Figure S4). We assumed that a relatively small fraction of detected interactions would significantly interact with the given viewpoint. We thus used the average scaled interaction signals as the expected 3C-seq signal for a given genomic distance. 3C-seq signals in cis were then transformed into a Z-score using the '(obs-exp)/SD' formula, where obs is the observed interaction signal found on the cis chromosome, exp is the scaled interaction signal for a specific genomic distance and SD is the standard deviation of the residual values '(obs-exp)'. P-values can then be assigned to each Z-score and transformed into a q-value for false discovery rate (FDR) analysis using the qvalue package from Bioconductor (20), with a 0.05 FDR level (which can be changed by the user) and using bootstrap as the selected method [qvalue(interactions.p.values, fdr.level = 0.05, pi0.method = 'bootstrap')]. The method we applied to identify interactions in cis has successfully been used for 5C data analysis (26), and a similar method has been used for the detection of interactions in 4C data analysis (8,15). Figure 4 shows the analysis results of two data sets using the mouse *Myb* promoter as viewpoint, showing that this method successfully identifies 3C-seq interaction regions in cis. These experiments were performed under two experimental conditions: (i) fetal liver (FL) erythrocytes expressing high levels of *Myb* and (ii) fetal brain (FB) cells expressing low levels of *Myb* (24). See the 'Results' section for an in-depth discussion of this analysis.

Identifying trans-interactions from 3C-seq experiments

To identify interaction regions in trans, we applied a similar formula as described for the identification of interactions in cis. It is not necessary to scale the trans-signal data, as there is no proximity bias for the interaction signals found on the trans chromosomes. We assumed that captured trans-interactions would have higher interaction signals than the mean of global interaction signals. We, therefore, transformed the detected interaction signals into a Z-score using the '(obs-exp)/SD' formula, where obs is the observed interaction signal found in the whole data set (excluding regions located within ±100 kb around the viewpoint), exp is the mean interaction signal for the whole data set and SD is standard deviation of the whole data set. Procedures similar to those used for cis-interactions were then performed to transform the trans interaction signals into statistical interaction scores (P- and q-values). Applying this method to the *Myb* promoter data sets (24) (see 'Results' section) resulted in the identification of several significant interaction regions in trans (see Supplementary Figure S5C for an example set of interactions detected in trans).

Analysis of 3C-seq replicate experiments

To investigate 3C-seq data reproducibility among replicates, we performed additional 3C-seq experiments using the *Myb* promoter as the viewpoint in FL erythrocytes and FB cells (3C-seq data are available at http://r3cseq.genereg.net). When considering the entire data set, including signals with low read counts,

interacting regions (≥1 RPM, calculated from restriction fragment-based, 5- and 10-kb window-based sizes) across the whole genome in general show low reproducibility, as low intensity signals in 3C-based methods are likely the product of random contacts between DNA fragments (9,15). Remarkably, reproducibility (defined as the percentage of detected interactions present in both replicates against all detected interactions) is extremely low in trans (<1% of detected interactions were reproducible, Supplementary Table S1); interactions found in trans almost always exhibit very low read counts and are, therefore, likely to be caused by random ligation events. However, in cis, reproducibility is significantly higher (17–40%, Supplementary Table S1) and improves when larger window sizes are used for interaction detection. As the most robust interaction regions are invariably located in cis, we next checked the reproducibility of high signal interaction regions (≥500 RPM within ±500 kb relative to the viewpoint) and observed that they are highly reproducible (50–90%, Supplementary Table S2), indicating that 3C-seq reliably reveals local chromatin structure around the viewpoint. However, basic analysis of read count data across replicates, such as those implemented in DESeq (27) and edgeR (28), which require overall high reproducibility within the entire data set, is not suitable for data analysis on 3C-seq replicate data sets, especially if one is interested in very long-distance interactions (including inter-chromosomal interactions). To determine significant interactions among replicates, we first performed r3Cseq data analysis for each individual sample. We then combined the detected interactions across biological replicates, providing 'union' and 'intersection' options for this purpose. The union method combines all significant interactions across samples, whereas the intersection method takes only significant interactions present across all samples into account. Read counts and RPM values across samples are averaged to obtain representative values for the final list of detected interactions. The assigned P-values across samples are combined using Fisher's combined probability test as implemented in R (29), and q-values are calculated using the qvalue package with FDR level 0.05 (which can be changed by the user), using bootstrap as the selected method.

Results

Functionality available in r3Cseq

The r3Cseq package was built on and extends the functionality of the Bioconductor packages BSgenome, GenomicRanges, Rsamtools and rtracklayer (30). It contains functions for the following groups of tasks:

Importing aligned reads

r3Cseq can read BAM files and converts this file and its related information into an object-oriented core class for the r3Cseq package. r3Cseq can also load aligned reads from the GRanges object generated by the GenomicRanges package in R. A detailed description of input parameters can be found in the r3Cseq software documentation.

Data processing

After class initialization, processing functions getRawReads, getBatchRawReads, getReadCountPerRestrictionFragment, getReadCountPerWindow, getBatchReadCountPerRestrictionFragment and getBatchReadCountPerWindow are performed. The getRawReads function retrieves aligned reads from BAM files and transforms them to GRanges objects that can be stored in an r3Cseq object, whereas getRawReadsInBatch processes the data in batch mode and stores the aligned reads GRanges in R files (.rdata). To count the number of reads used for further analysis, r3Cseq provides two ways to count the number of reads per region; (i) count the number of reads per restriction fragment (using the getReadCountPerRestrictionFragment function) and (ii) count the number of reads per non-overlapping defined window (using the getReadCountPerWindow function), whereas the getBatchReadCountPerRestrictionFragment and getBatchReadCountPerWindow functions perform the same analysis for replicate data sets. These functions provide options for counting all reads or only the informative reads (i.e. those that are mapped exactly adjacent to restriction sites). The latter method will exclude reads generated from inappropriately digested chromatin by the restriction enzyme and randomly sequenced DNA fragments (see r3Cseq software documentation).

Data normalization and the identification of interacting regions

calculateRPM and calculateBatchRPM are functions that normalize the number of RPM for each restriction fragment or window. Users can select different RPM calculation methods, as described in the 'Materials and Methods' section, by defining the normalization method parameters (see r3Cseq software documentation). After normalization, the getInteractions and getBatchInteractions functions are used to calculate Z-scores, estimate P-values and assign q-values to detect significant interactions, respectively. For the getBatchInteractions function, users can define the selected combine-method parameter for the detection of interactions across replicates ('union' and 'intersection').

Visualization

The plotOverviewInteractions, plotInteractionsNearViewpoint, plotInteractionsPerChromosome and plotDomainogramNearViewpoint functions are provided for visualization of the interaction regions, taking advantage of the powerful plotting facilities in R. Supplementary Figure S4 shows examples of the plots generated by these functions, allowing users to explore the interaction regions of their data sets.

Data export

The exportInteractions2text, exportBatchInteractions2text and export3Cseq2bedGraph functions are used to export the analysis results to tab-delimited text files. The identified interaction regions can be exported into the bedGraph format, which can be easily uploaded to the UCSC Genome Browser (16) and IGV (17) for further visualization and exploration.

Annotation of interactions

The getExpInteractionInRefseq and getContrInteractionInRefseq functions provide a list of candidate genes, which contain significant interaction signals in their proximity. Here, proximity is defined by input parameters that specify the relative distances to the start and end positions of genes (see r3Cseq software documentation).

Preparing a final report

The generate3CseqReport function can be used to export all results of the analysis, including plots and text files, into a PDF file that can be used for data interpretation and publication.

A proof of principle analysis using r3Cseq

As a proof of principle analysis, we used the r3Cseq package to characterize long-range interactions at the mouse β-globin locus using our 3C-seq and ChIP-seq data from Soler et al. (13). The chromosomal architecture of the β-globin locus has been studied intensively and serves as an excellent test case for r3Cseq functionality. Previous studies have shown that on activation, the β-major gene (β-*maj*) from the β-globin locus engages in long-range interactions with upstream regulatory sites (forming the LCR) located 40–60 kb away (31–33). We obtained our 3C-seq data (13) using the β-major globin gene (β-*maj*) promoter as a viewpoint and applied the r3Cseq package for data analysis. 3C-seq experiments were performed in two cell types from the mouse 12.5 dpc embryo: (i) cells expressing the β-globin genes below detection level (FB) and (ii) cells expressing very high levels of β-globin (FL) (9,13). As interactions between the LCR and the β-globin genes have not been observed in FB cells, this experiment served as a negative control to allow the identification of erythroid-specific interactions within the β-globin cluster.

Short reads generated by Illumina sequencing of 3C-seq libraries were mapped to the mouse genome (NCBI37/mm9) using the Bowtie aligner. Mapping files were then analyzed using r3Cseq to identify candidate interacting regions and to generate plots for data visualization. As expected, regions identified as interacting in both tissues were found most frequently in cis on chromosome 7, where the β-*maj* viewpoint is located. The interactions predominantly map relatively close to the viewpoint fragment (within 125 kb up- and downstream). However, only in FL erythrocytes, robust interaction regions were detected in the

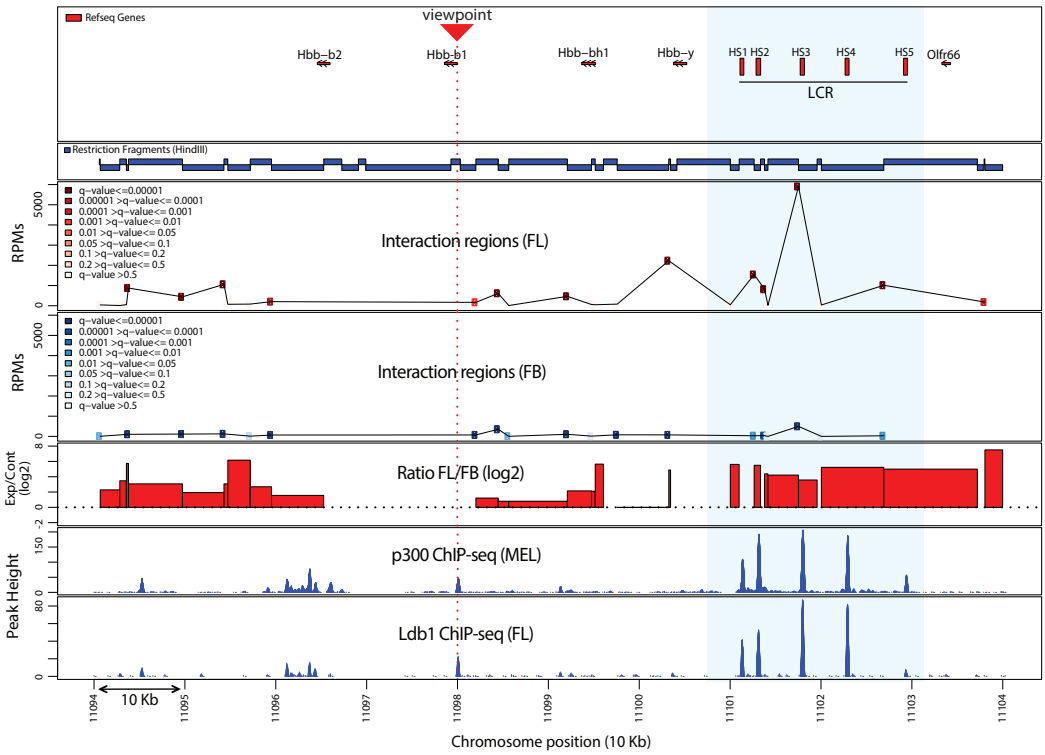


Figure 3. A proof of principle 3C-seq/r3Cseq analysis on the well-characterized β-globin locus. Gene locations are shown at the top followed below by a map of restriction fragments. The line plots show overall detected cis-interaction signals (40 kb up- and 60 kb downstream of the viewpoint) with the β-major promoter in both FL and FB cells. High signals on the viewpoint fragment or the immediately adjacent fragments were excluded. Color gradients represent the range of significant interaction signals (q-value). The bar plot represents the ratio (log2) of normalized signal between FL and FB. The light blue box highlights the LCR region with its hypersensitive sites (HS1-5), coinciding with several FL-specific significant interaction regions and binding sites of transcription factor complexes.

region 40–60 kb upstream of the β-*maj* gene, corresponding to the location of the LCR. We focused our analysis on this area: the outcome is summarized in Figure 3. In the LCR region, interaction signals detected in FL erythrocytes are much stronger than those observed in FB. These strong interaction regions are statistically significant, with $q \leq 0.01$. Reassuringly, they coincide with the binding regions of transcriptional co-activator p300 and the Ldb1 transcription factor, known to be involved in enhancer function and globin gene regulation. This confirms the presence of a looping structure, placing the LCR in close proximity to the β-*maj* gene promoter, and it shows that the sites of long-range interactions coincide with sites of regulatory factor binding (13). r3Cseq promptly provides results for both restriction fragment and window-based analysis. We observed that the significant interactions detected from individual restriction fragments and 5- and 10-kb window-based regions in FL are similar. However, the strongest interaction signal is positioned slightly differently depending on the window range; fragment-based analysis detects the strongest signal at the third hypersensitive site (HS3) of the LCR, whereas 5- and 10-kb window-based analysis positions the peak interaction signal at HS2 of the LCR (Supplementary Figure S6). This discrepancy is an obvious consequence of the choice of window size. On the other hand, log2 fold change calculations (FL/FB) show a robust FL-specific interaction signal covering all five HSs of the LCR (Figure 3). Although all methods detect an erythroid-specific β-*maj*–LCR interaction, the differences between these methods can produce subtle differences in interaction profiles. Additionally, r3Cseq provides an option for users to use the getReadCountPerWindow function with an 'overlapping' window option that may correct for any

bias from the arbitrary starting position of windowing. Users will, therefore, have to carefully consider this when selecting their analysis parameters. Because of the generally maximized resolution (~4 kb) and their unbiased nature, we suggest that the interaction signals detected per individual fragment are the most suitable starting point. Window-based approaches will reduce the detected resolution (depending on the selected size range), but they will improve reproducibility, which can be convenient when long-distance cis-interactions or trans-interactions are of particular interest.

The r3Cseq interaction detection method provides a better interaction score (q-values) than was used in our previous analysis (13). This analysis did not include background signal correction, which resulted in assigning overly significant interaction scores to low interaction signals (Supplementary Figure S6). Although our previous method detected ~4000 contacts, our current methods detected ~600 significant contacts in FL ($q \leq 0.05$) using the same cut-off.

To show how the r3Cseq package can also be used to analyze data sets generated by other laboratories, we used r3Cseq to identify long-range interactions at the mouse β-globin locus using data obtained from a recent 4C-seq study (9). We analyzed those data sets for which the β-*maj* gene promoter was used as the viewpoint (in FL cells), which was produced using a slightly different protocol (9). In this data set, we were able to demonstrate that our detection method can capture strong interaction regions using individual restriction fragments, 2- and 5-kb window-based analysis in the 100-kb β-globin locus domain (Supplementary Figure S7). Preferred contacts within the locus start from the active β-globin gene toward the most distal HS of the LCR (HS5), and the strongest interaction signals in the LCR are predominantly located between HS1 and HS2. These results coincide and highly correlate with the interaction profile reported in the 4C-seq study (9) (Supplementary Figure S7).

Taken together, we have shown that our r3Cseq package can be used successfully to analyze 3C-seq data to reveal long-range chromatin interactions that play critical roles in gene regulation.

Application to genomic regions with previously uncharacterized interactions

To demonstrate that 3C-seq/r3Cseq can be further applied to study chromatin interactions in a structurally unexplored locus, we outline how r3Cseq was used to analyze 3C-seq data generated to study the chromatin conformation of the mouse *Myb* locus during erythroid development. The 3C-seq and ChIP-seq data used in this demonstration were taken from previously published data (24).

Myb, encoding the c-Myb transcription factor, is a key hematopoietic regulator and plays a pivotal role in maintaining a proper balance between erythroid cell proliferation and differentiation (34–36). Previous reports have shown that erythroid transcription factor complexes occupy distinct sites near *Myb* in the *Myb-Hbs1l* intergenic region (13,37,38). ChIP-seq data obtained for the Ldb1 transcription factor complex, which is a key regulator of erythroid development, revealed a binding cluster in a region spanning 60 kb in the *Myb-Hbs1l* intergenic region. *Myb* expression is highly dynamic during the course of erythroid differentiation (39,40). Considering that the Ldb1 complex is required for proper erythroid maturation and is known to regulate genes in a long distance manner, it was hypothesized that the intergenic Ldb1-complex–binding sites represent distal regulatory elements that control *Myb* expression during erythroid differentiation.

We previously performed 3C-seq experiments using the *Myb* promoter as a viewpoint to investigate whether the Ldb1-complex–binding sites in the intergenic region interact with the *Myb* gene via chromatin looping (24). We used r3Cseq to identify and explore the candidate regions, which are interacting with the *Myb* promoter. As for the β-globin locus experiment, 3C-seq was performed on FL erythrocytes (expressing high levels of *Myb*) and FB cells (expressing undetectable levels of *Myb*). The latter was used as a negative control to appropriately link locus structure to gene expression. After mapping, we used r3Cseq to analyze the 3C-seq data, and we were able to identify candidate interaction regions, which are shown in Figure 4. In FL erythrocytes, these high interaction signals were found to coincide with the intergenic Ldb1–complex– and p300-binding sites (Figure 4, indicated by the purple dashed box). Interaction frequencies of these regions in FL erythrocytes were statistically significant ($q \leq 0.01$) and substantially higher than in FB (fold change ($\log2 \geq 2$), where they were either absent or low. Domainogram plots of interaction regions generated by the plotDomainogramNearViewpoint function (using a window-based analysis running from 2 to 30 kb, increasing 1 kb per run) clearly revealed the different intergenic interaction intensity between FL and FB. We also confirmed that these significantly different intergenic interactions between FL and FB can be detected using a different analysis method (9) (Supplementary Figure S8), suggesting that the data

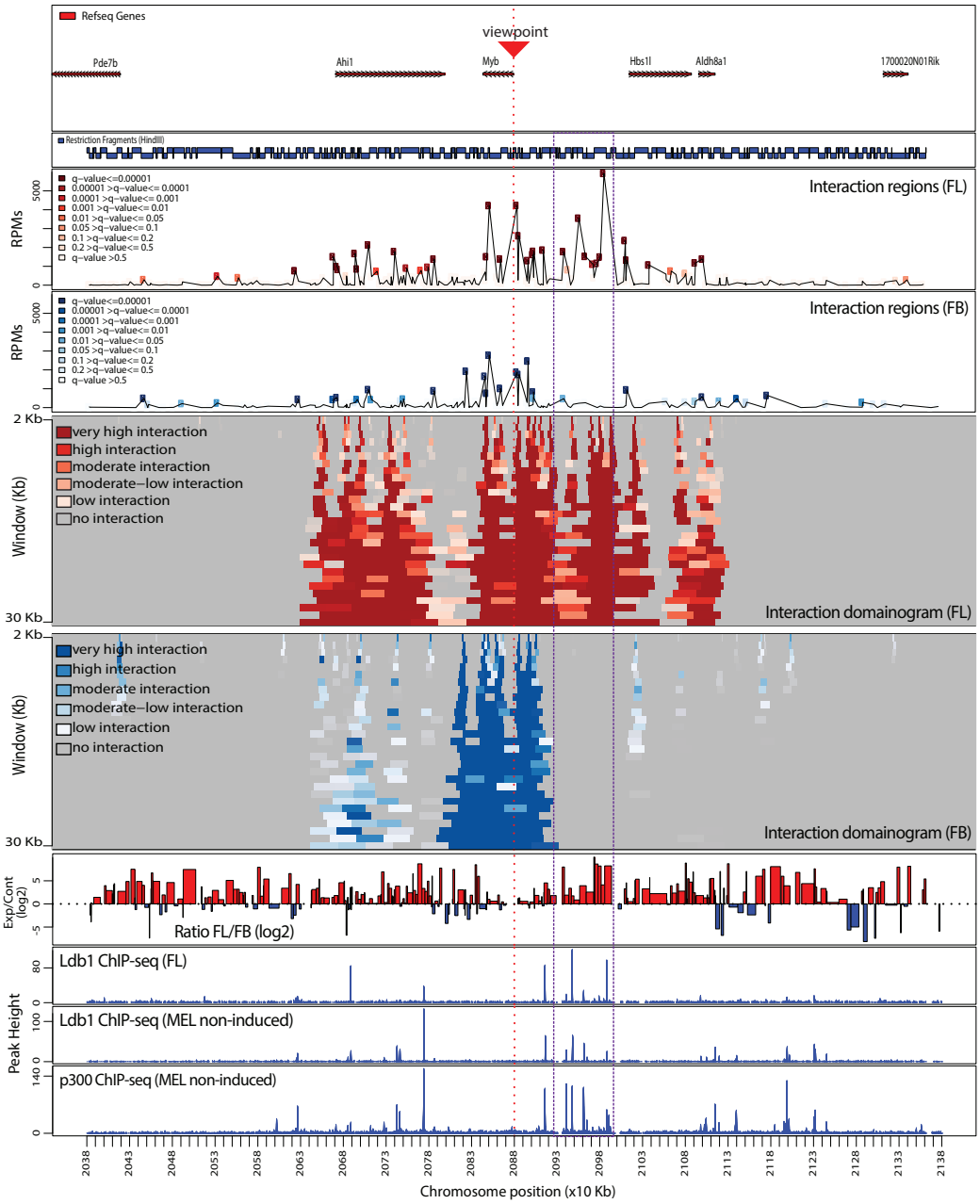


Figure 4. Application of 3C-seq/r3Cseq analysis at the *Myb* locus. Gene locations are shown at the top followed below by a map of restriction fragments. The line plots show detected cis-interaction regions 500 kb up- and 500 kb downstream of the *Myb* promoter viewpoint in both FL and FB cells. High signals on the viewpoint fragment or the immediately adjacent fragments were excluded. The domainograms show the detected interactions after a window-based analysis (running from 2 to 30 kb) in FL and FB cells. Color gradients of the domainograms represent the interaction signal strength detected for each run of the defined window (transformed q-value). The purple dashed box highlights the *Myb-Hbs1l* intergenic region, which shows strong interaction signals coinciding with binding sites of the Ldb1 and p300 transcription factor complexes.

generated by our protocol can also be analyzed by other existing tools. Although both analysis methods assign tissue-specific interactions to the intergenic region, r3Cseq interactions are more robust when compared with the 4C-seq pipeline (using default parameters, interactions are detected at ~0.01–0.1 of the median of window coverage). This suggests that r3Cseq possesses increased detection sensitivity, at least in this particular case. These results show that analyzing 3C-seq data with r3Cseq can identify candidate tissue-specific regulatory regions within a structurally unexplored locus for further experimental investigation. The example data sets and the R codes used to perform the key steps of this analysis are provided at the r3Cseq website (http://r3cseq.genereg.net).

We next wanted to test whether our method and software could be used to study the dynamics of long-range chromatin interactions during cellular differentiation. We analyzed 3C-seq data obtained from mouse erythroleukemia (MEL) cells before and after treatment with a differentiation-inducing agent, again using the *Myb* promoter as a viewpoint (24). The 3C-seq data were analyzed using r3Cseq. The identified interaction regions are shown in Supplementary Figure S9. In non-induced MEL cells (expressing high levels of *Myb*), interaction regions are similar to those found in FL erythrocytes (Figure 4), often overlapping with Ldb1- and p300-complex–binding sites. Strikingly, on induction of differentiation, these interaction regions showed much lower interaction signals. These results reveal diminished interactions between the promoter and intergenic regulatory regions upon cellular differentiation, coinciding with the downregulation of *Myb* expression and suggesting that these interactions are involved in the regulation of *Myb*. Studying long-range interactions within developmentally regulated loci, exemplified here by the *Myb* locus, is an important application of 3C-seq/r3Cseq in exploring the mechanisms of gene regulation.

We next performed 3C-seq replicates for the *Myb* promoter experiment to investigate the reproducibility of the detected interactions across independently prepared samples. As described in the 'Materials and Methods' section, we observed that genome-wide detected interaction regions show low reproducibility at the single-fragment level, especially for low signals and inter-chromosomal interactions (Supplementary Table S1). To further investigate such interactions in a set of biological replicate experiments, we used a 20-kb window-based analysis to detect significant interaction across the replicates. Indeed, the plotOverviewInteractions function and 'intersection' method can remove signals originating from random ligation events present in the single-data set analysis, which mostly occur in trans and at long distance sites in cis (Supplementary Figure S10A and B). We next determined the genes that are located within the FL significant interaction regions. Using the getExpInteractionInRefseq function, we detected 11 genes in the proximity of significant interaction regions (50 kb upstream of the gene start and 5 kb downstream of the gene end), whereas 192 genes were detected in a single-data set analysis in FL erythrocytes (Supplementary Figure S10C and D). As signal reproducibility around the viewpoint is high, genes close to *Myb* where found within this list (*Ahi1*, *Hbs1l* and *Aldh8a1*). Interestingly, of the other genes in relatively close proximity to *Myb* (within 3 Mb), only those that are highly expressed in erythroid cells (*Bclaf1* and *Fam54a*, as determined by RNA-seq in MEL cells, data not shown) are found to interact reproducibly with *Myb*. Five genes located on trans chromosomes (*Gm14496*, *Eif4enif1*, *Sfi1*, *Spata5* and *Rit2*) were consistently found in proximity of *Myb*. Whether this observed gene clustering is of any relevance to *Myb* and/or erythroid biology remains unclear, although these observations might prove to be interesting in the context of genetic translocations and transcription factories (41).

Discussion

We developed the R/Bioconductor package r3Cseq, and in this study, we describe its functionality and demonstrate its use and power for the identification of chromatin interaction regions generated by 3C-seq experiments. The software provides functionality for pre-processing, analyzing and visualizing interaction regions with any given viewpoint of interest. The package can process BAM files, which are generated by mapping software, such as Eland and Bowtie. We provided r3Cseq with functions to support the BAM file format, as it is a compressed binary file (21), which allows users to perform 3C-seq data analysis on a regular personal computer (CPU ~2 GHZ with ~4 GB of random access memory, Supplementary Table S3). However, it is recommended, when possible, to use more powerful computer hardware when performing 3C-seq data analysis on replicate data sets, as this will require more random access memory and storage space.

Our work focused on building the functionality to support data analysis of 3C-seq experiments. We adapted and applied methods used in high-throughput sequencing data analysis for normalization and the detection of significant interaction regions. We applied a fitted reverse-cumulative distribution of reads per

region to a power-law distribution as the normalization method, increasing the statistical power of 3C-seq signal detection. This method reveals a better fit of normalized values for 3C-seq data than a simplified RPM calculation. r3Cseq still provides functions to support both methods, offering users the choice to use normalized values obtained from each separate method for further analysis. We adapted methods used in previous 4C and 5C studies to detect significant interactions in both cis and trans. Our method corrects any bias resulting from background interaction signals and assigns an interaction score (q-value) directly to a certain restriction fragment or a defined window. Selecting an appropriate window setting for counting reads is critical for 3C-seq data analysis. Both fragment-based and window-based methods have advantages and disadvantages. A fragment-based strategy generally maximizes resolution (~4 kb) and can identify direct interaction sites in an unbiased way, which might be preferred when specific interaction regions are to be compared with other types of high-resolution data, such as the transcription factor-binding sites identified by ChIP-seq. The outcome of window-based methods depends on the choice of the arbitrary selected window size, subsequently limiting the identification of interaction regions to within that specific size, often reducing the effective resolution. However, the window-based strategy is helpful for detecting large interaction domains, which can significantly promote the identification of novel interaction regions. Larger window sizes show a higher reproducibility (Supplementary Table S1), especially at large distances (>500 kb) from the viewpoint and may also be preferred for replicate data analysis. To facilitate both fragment-based and window-based analysis, r3Cseq enables users to easily switch between both types of analysis and promptly provides these results as text files and interpretable plots (see http://r3cseq.genereg.net for more details). r3Cseq also supports data analysis of replicate 3C-seq experiments, as it can combine detected interactions across biological replicates to produce a final list of significant interactions. Users can select a union or an intersection operation to obtain the final list of interactions, as described in the 'Materials and Methods' section. Both these options are useful, although only the intersection method allows for the detection of truly consistent interaction regions across samples. In summary, we have shown that r3Cseq can remove signals originating from random ligation events and provide data normalization, the accurate detection and powerful visualization of both existing and novel significant interaction regions present across multiple biological replicates.

Availability and implementation

The r3Cseq package has been implemented in R and is available as part of the Bioconductor (www.bioconductor.org) distribution, as version 2.9. As such, it also gives users and software developers the opportunity to extend and customize the pipeline to their needs. We have developed a website to host the r3Cseq package, which can be found at http://r3cseq.genereg.net. The website provides downloadable data sets presented in this article and the current version of the r3Cseq package (version 1.5.0). The website also describes the R code examples for the 3C-seq data analysis pipeline. Additional guidelines and typical workflows can be found in the package's vignette in R. The ChIP-seq and 3C-seq data used here were deposited in the sequence read archive (SRA) database. Accession numbers for these data were previously published (9,13,23,24).

Future directions

In the next version of r3Cseq package, more functions will be implemented to allow users to incorporate external data sets, such as ChIP-seq and expression data into the analysis, which will be of great assistance in studying long-range gene regulation.

Supplementary Data

Supplementary Data are available at NAR Online: Supplementary Tables 1–3 and Supplementary Figures 1–10.

Funding

EU-FP7 integrated project EuTRACC LSHG-CT-2007-037445 (S.T., R.S., E.S. and F.G.); Dutch Cancer Genomics Center (CGC) and the French Alternative Energies and Atomic Energy Commission (CEA) (to E.S.); Norwegian Research Council (YFF) and Bergen Research Foundation (BFS) (to B.L.). Funding for open access charge: Department of Informatics, University of Bergen (to B.L.).

Conflict of interest statement. None declared.

Acknowledgements

The authors thank Gemma Danks, Nathan Harmston and Chilamakuri C. S. Reddy for critical reading of the manuscript, Petros Kolovos for the discussion on data analysis and Wilfred van IJcken and Mirjam van den Hout for Illumina sequencing and help with the initial bioinformatic processing.

References

1. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl Acad. Sci. USA. 2003;100:15776–15781.
2. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316:1497–1502.
3. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132:311–322.
4. Rudkin GT, Stollar BD. High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence. Nature. 1977;265:472–473.
5. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–1311.
6. Tolhuis B, Palstra R-J, Splinter E, Grosveld F, de Laat W. Looping and Interaction between Hypersensitive Sites in the Active [beta]-globin Locus. Mol. Cell. 2002;10:1453–1465.
7. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Gene Dev. 2012;26:11–24.
8. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C) Nat. Genet. 2006;38:1348–1354.
9. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BA, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods. 2012;9:969–972.
10. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–1309.
11. Tiwari VK, Cope L, McGarvey KM, Ohm JE, Baylin SB. A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. Genome Res. 2008;18:1171–1179.
12. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–293. [
13. Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Stevens M, Kockx C, van Ijcken W, et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev. 2010;24:277–289.
14. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, Palstra RJ, Wendt KS, Grosveld F, van Ijcken W, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat. Protoc. 2013;8:509–524.
15. Splinter E, de Wit E, van de Werken HJG, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods. 2012;58:221–230.
16. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
17. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat. Biotechnol. 2011;29:24–26.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
19. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J. Comput. Graph. Stat. 1996;5:299–314.
20. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–2079.
22. Zhang ZD, Rozowsky J, Snyder M, Chang J, Gerstein M. Modeling ChIP sequencing in silico with applications. PLoS Comput. Biol. 2008;4:e1000158.
23. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 2009;10:R79.
24. Stadhouders R, Thongjuea S, Andrieu-Soler C, Palstra RJ, Bryne JC, van den Heuvel A, Stevens M, de Boer E, Kockx C, van der Sloot A, et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. EMBO J. 2012;31:986–999.
25. Ribeiro de Almeida C, Stadhouders R, de Bruijn MJ, Bergen IM, Thongjuea S, Lenhard B, van Ijcken W, Grosveld F, Galjart N, Soler E, et al. The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. Immunity. 2011;35:501–513.
26. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–113.
27. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
28. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–140.
29. Tukey J. Statistical methods for research workers. Econometrica. 1952;20:511–512.
30. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–1842.
31. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell. 2002;10:1453–1465.
32. Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W. The beta-globin nuclear compartment in development and erythroid differentiation. Nature Genet. 2003;35:190–194.

33. Song SH, Hou C, Dean A. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. Mol. Cell. 2007;28:810–822.
34. Lieu YK, Reddy EP. Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation. Proc. Natl Acad. Sci. USA. 2009;106:21689–21694.
35. Ramsay RG, Gonda TJ. MYB function in normal and cancer cells. Nat. Rev. Cancer. 2008;8:523–534.
36. Vegiopoulos A, Garcia P, Emambokus N, Frampton J. Coordination of erythropoiesis by the transcription factor c-Myb. Blood. 2006;107:4703–4710.
37. Tallack MR, Whitington T, Yuen WS, Wainwright EN, Keys JR, Gardiner BB, Nourbakhsh E, Cloonan N, Grimmond SM, Bailey TL, et al. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res. 2010;20:1052–1063.
38. Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C. Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells. Genome Res. 2010;20:1064–1083.
39. Gonda TJ, Metcalf D. Expression of Myb, Myc and Fos proto-oncogenes during the differentiation of a murine myeloid-leukemia. Nature. 1984;310:249–251.
40. Emambokus N, Vegiopoulos A, Harman B, Jenkinson E, Anderson G, Frampton J. Progression through key stages of haemopoiesis is dependent on distinct threshold levels of c-Myb. EMBO J. 2003;22:4478–4488.
41. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. 2006;4:e138.

Chapter 5

Dynamic long-range chromatin interactions control *Myb* proto-oncogene transcription during erythroid development

Ralph Stadhouders[1]*, Supat Thongjuea[2]*, Charlotte Andrieu-Soler[1,3], Robert-Jan Palstra[1], Jan Christian Bryne[2], Anita van den Heuvel[1], Mary Stevens[1], Ernie de Boer[1], Christel Kockx[4], Antoine van der Sloot[4], Mirjam van den Hout[4], Wilfred van IJcken[4], Dirk Eick[5], Boris Lenhard[2,6], Frank Grosveld[1,7]† & Eric Soler[1,7]†

[1]Department of Cell Biology, Erasmus MC, Rotterdam, The Netherlands.
[2]Computational Biology Unit, Uni Computing AS, University of Bergen, Bergen, Norway.
[3]Institut National pour la Santé Et la Recherche Médicale (INSERM) U872, Physiopathology of Ocular Diseases: Therapeutic Innovations, Paris, France.
[4]Center for Biomics, Erasmus MC, Rotterdam, The Netherlands
[5]Department of Molecular Epigenetics, Helmholtz Zentrum München, Center of Integrated Protein Science (CIPSM), Munich, Germany.
[6]Department of Biology, University of Bergen, Bergen, Norway.
[7]Cancer Genomics Centre, Erasmus MC, Rotterdam, The Netherlands.

***These authors contributed equally.**
†Corresponding authors.

Published in:
EMBO Journal
2012; 31:986-99

Abstract

The key haematopoietic regulator *Myb* is essential for coordinating proliferation and differentiation. ChIP-Sequencing and Chromosome Conformation Capture (3C)-Sequencing were used to characterize the structural and protein-binding dynamics of the *Myb* locus during erythroid differentiation. In proliferating cells expressing *Myb*, enhancers within the *Myb-Hbs1l* intergenic region were shown to form an active chromatin hub (ACH) containing the *Myb* promoter and first intron. This first intron was found to harbour the transition site from transcription initiation to elongation, which takes place around a conserved CTCF site. Upon erythroid differentiation, *Myb* expression is downregulated and the ACH destabilized. We propose a model for *Myb* activation by distal enhancers dynamically bound by KLF1 and the GATA1/TAL1/LDB1 complex, which primarily function as a transcription elongation element through chromatin looping.

Introduction

The differentiation of stem and progenitor cells into mature differentiated cells requires a tight control of progenitor cell expansion, proliferation arrest and terminal differentiation. The *Myb* proto-oncogene encoding the c-Myb transcription factor (TF) is expressed in stem and progenitor cells of all haematopoietic lineages and plays a central role in the control of their proliferation (Mucenski et al, 1991; Sandberg et al, 2005; Vegiopoulos et al, 2006; Ramsay and Gonda, 2008; Lieu and Reddy, 2009). Lack of *Myb* is lethal (E15) due to the complete absence of definitive erythroid cells (Mucenski et al, 1991). Conditional knockout models revealed additional essential non-erythroid roles of *Myb*, mainly in the lymphoid system (Bender et al, 2004; Thomas et al, 2005), and the self-renewal and multi-lineage differentiation potential of adult haematopoietic stem cells (Lieu and Reddy, 2009). *Myb* is highly expressed in immature proliferating haematopoietic cells and is strongly downregulated in terminally differentiating cells (Gonda and Metcalf, 1984; Emambokus et al, 2003), suggesting that *Myb* is linked to the transition between proliferation and differentiation. Aberrant *Myb* expression in leukemic cells is consistent with this idea (Ramsay and Gonda, 2008), correlating with increased proliferation and a loss of differentiation. Despite its importance, the control of *Myb* expression during haematopoiesis is poorly understood. Early work suggested a regulatory role for sequences in the first intron, primarily in blocking transcription elongation (Bender et al, 1987; Reddy and Reddy, 1989; Hugo et al, 2006). Recently, microRNAs were shown to be involved in regulating c-Myb protein levels (Xiao et al, 2007; Lu et al, 2008). However, the transcriptional regulatory elements and associated trans-acting factors controlling *Myb* expression during development remain mostly uncharacterized.

The mouse *Myb* gene on chromosome 10 is flanked by the *Ahi1* and *Hbs1l* genes, which have no known function during haematopoiesis. Several studies pointed out a potential role for the 135 kb *Myb-Hbs1l* intergenic region in the regulation of *Myb*: (i) transgene integration within the intergenic region led to severe downregulation of *Myb* expression (Mukai et al, 2006); (ii) ChIP-on-chip data showed an open chromatin structure (i.e., H3Ac and H4Ac) of the region in human erythroid cells expressing *MYB* (Wahlberg et al, 2009); and (iii) several studies showed that SNPs in the human *MYB-HBS1L* intergenic region (possibly affecting *MYB* expression) were strongly associated with variation in several clinically relevant erythrocyte traits (Thein et al, 2007; Lettre et al, 2008; Ganesh et al, 2009). For example, specific SNPs associate with elevated fetal haemoglobin (HbF), which ameliorates β-hemoglobinopathy severity and has therapeutic potential. Thus, important regulatory elements appear to reside in the *Myb-Hbs1l* intergenic region, but they have not been localized precisely or characterized in any way.

Erythroid development is controlled by an array of TFs, including GATA1, its associated partners LDB1, TAL1, KLF1 and c-Myb (Cantor and Orkin, 2001). A complex of the haematopoietic TFs GATA1/TAL1/LDB1 together with the ETO2/MTGR1 cofactors (the 'LDB1 complex') binds regulatory regions of developmentally regulated genes (Fujiwara et al, 2009; Yu et al, 2009; Kassouf et al, 2010; Soler et al, 2010; Tallack et al, 2010) and controls their activation upon terminal erythroid differentiation (Soler et al, 2010). The LDB1 complex preferentially binds at large distances from promoters (up to 300 kb) in intergenic regions, providing long-range candidate regulatory elements. An example is the long-range control of the β-globin genes by cis-regulatory elements spread over 100 kb, forming the locus control region (LCR). When β-globin is expressed, the LCR folds into a three-dimensional (3D) active chromatin hub (ACH) (Tolhuis et al, 2002; Palstra et al, 2003), where distal enhancers reside in close proximity to the expressed genes. Structural proteins such as CTCF and Cohesin are known to participate in such 3D interactions (Ong and Corces, 2011). TFs also have a role in long-range gene regulation, for example, LDB1, GATA1, FOG1 and KLF1 are required to maintain such interactions within the β-globin locus and other loci (Drissen et al, 2004; Vakoc et al, 2005;

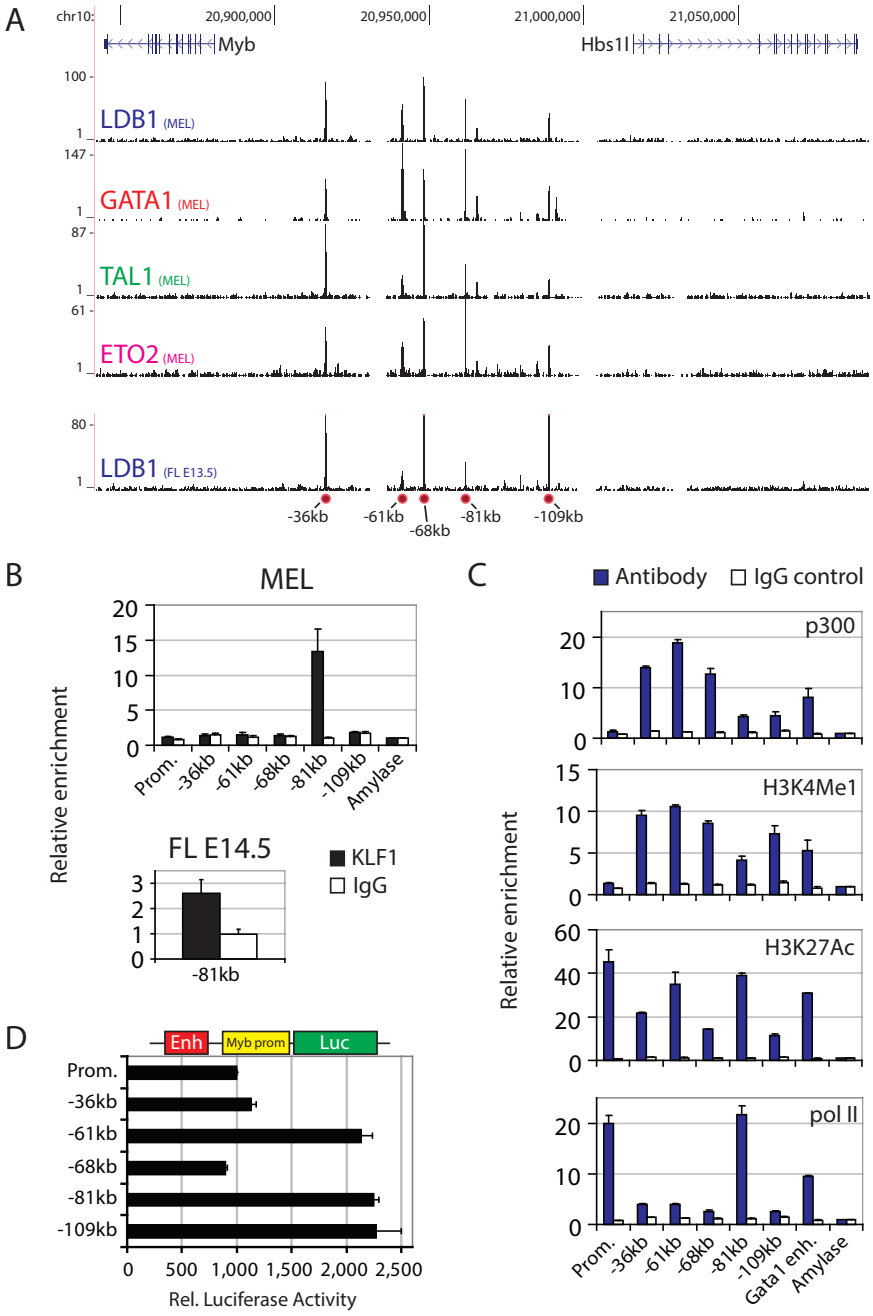


Figure 1. The *Myb-Hbs1l* intergenic region contains transcriptional enhancers. (A) ChIP-Seq of the LDB1 complex components LDB1, GATA1, TAL1 and ETO2 in the *Myb-Hbs1l* locus in MEL cells (MEL). LDB1 binding in primary E13.5 fetal liver erythroid progenitors is also shown (FL E13.5). The position of the intergenic binding sites relative to the *Myb* TSS is indicated at the bottom (red circles). (B) ChIP analysis showing intergenic KLF1 occupancy in MEL cells and in E14.5 fetal liver cells. (C) ChIP analysis showing the binding of p300, polII and the presence of the enhancer-associated histone modifications H3K4me1 and H3K27ac at the intergenic region in MEL cells. (D) Luciferase reporter assays in MEL cells showing the enhancer activity of the different intergenic elements. ChIP enrichments were calculated versus a negative control region (amylase). Results are presented as the mean±s.e.m. of at least two independent experiments.

Song et al, 2007; Jing et al, 2008).

We report here that the activating LDB1 complex, KLF1 and CTCF occupy multiple regulatory elements within the *Myb-Hbs1l* intergenic region, which have the chromatin hallmarks of active enhancers. Chromosome Conformation Capture (3C) and high-throughput sequencing (3C-Seq) show that these elements and the actively transcribed *Myb* gene cluster together in the nuclear space to form an ACH *in vivo*, bringing the enhancers in close proximity to the *Myb* gene promoter and first intron. The latter contains a highly conserved CTCF binding site around which productive transcription elongation starts. The ACH is lost when cells terminally differentiate, concomitant with the downregulation of *Myb* and a decreased binding of TF complexes at the distal enhancers.

Results

The LDB1 complex binds distal enhancers in the Myb-Hbs1l intergenic region

ChIP-Sequencing (ChIP-Seq) was used to identify the genome-wide binding sites of key erythroid TFs in mouse erythroleukaemia (MEL) cells and in primary mouse fetal liver (FL) cells (Soler et al, 2010). This showed preferential intragenic and intergenic binding of the LDB1 complex away from promoter sequences, suggesting it is involved in long-range gene regulation, a hypothesis supported by other studies (Song et al, 2007). Five LDB1 complex binding sites were detected in the *Myb-Hbs1l* intergenic region, −36, −61, −68, −81 and −109 kb upstream of the *Myb* transcription start site, in MEL cells and primary mouse erythroid progenitors from E13.5 FL (Figure 1A; Supplementary Figure S1A and B). These intergenic binding sites harboured all components of the activating LDB1 complex (GATA1/LDB1/TAL1/ETO2) in erythroid progenitors, consistent with active transcription of both *Myb* and *Hbs1l* genes (Supplementary Figure S1C). Additionally, in MEL and primary FL cells, the −81 kb binding site was found co-occupied by KLF1 (Figure 1B), a key erythroid TF primarily associated with gene activation, in agreement with a recent KLF1 ChIP-Seq experiment performed using primary mouse erythroid progenitors (Tallack et al, 2010; Supplementary Figure S2B and C). None of these TFs were found to bind the *Myb* or *Hbs1l* promoters. Next, all intergenic sites were shown to possess characteristic features supporting enhancer activity, that is, the presence of the histone acetyl transferase p300 (Visel et al, 2009), RNA polymerase II (polII), monomethylated histone 3 Lysine 4 (H3K4me1), and acetylated H3K27 (Heintzman et al, 2009; Figure 1C). PolII occupancy was especially abundant on the LDB1/KLF1 bound −81 kb sequence, showing similar enrichments to the highly active *Myb* promoter. In order to show that these LDB1 binding sites can indeed act as enhancers, they were cloned upstream of a minimal *Myb* promoter controlling a firefly luciferase reporter gene. Transfection into MEL cells showed that the −61, −81 and −109 kb elements are able to enhance luciferase activity (Figure 1D). In summary, these results suggest that the intergenic LDB1 complex binding sites represent active regulatory elements in erythroid progenitors, some possessing enhancer activity *in vitro*.

In-vivo conformation of the Myb-Hbs1l locus

We next performed 3C-Seq (Soler et al, 2010) experiments (Supplementary Figure S3) to investigate whether the *Myb* promoter was interacting with the intergenic regulatory elements. 3C-Seq was first performed on fresh mouse E12.5 FL tissue (primarily containing erythroid progenitors) using the *Myb* promoter as the viewpoint. Fetal brain (FB) samples were processed in parallel as a control, since *Myb* expression is much lower in brain tissue and it lacks the erythroid-specific LDB1 complex. Furthermore, a previous p300 ChIP-Seq performed in FB tissue showed no enrichments within the *Myb-Hbs1l* intergenic region (Supplementary Figure S2A). Multiple promoter-interacting elements located in the intergenic region were detected in FL, of which most were either absent or showed a low signal in FB (Figure 2A and B), thus revealing erythroid-specific long-range communication between the *Myb* promoter and intergenic elements. In addition, 3C-Seq signals were shown to correlate with binding of the LDB1 complex, KLF1 and CTCF, which have all been implicated in mediating long-range chromatin interactions (Drissen et al, 2004; Vakoc et al, 2005; Splinter et al, 2006; Song et al, 2007; Figure 2B). Statistical analysis (Poisson distribution/running-mean comparison, $P \leq 0.001$) of the FL and FB 3C-Seq data sets confirmed the erythroid specificity of the majority of intergenic interactions (Supplementary Figure S4). Quantitative 3C-qPCR experiments were carried out to confirm these results. This shows a very similar long-range interaction pattern (Figure 2C), with the exception of the −68-kb LDB1 complex binding site that was not detected by 3C-Seq but was

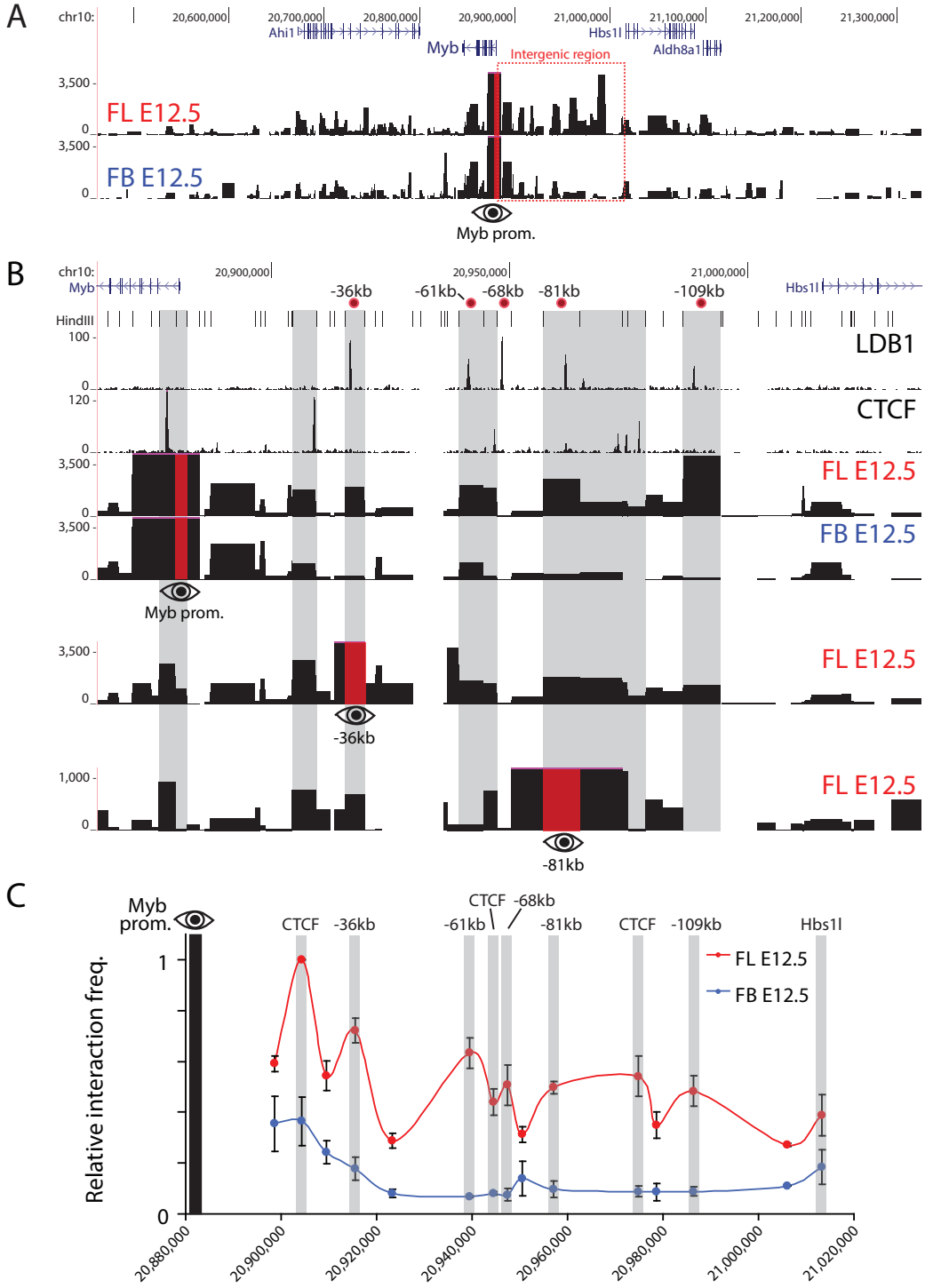


Figure 2. *(Legend at the bottom of the next page)*

found interacting by 3C-qPCR. These data show *in vivo* nuclear proximity between LDB1 complex, KLF1 and CTCF-bound intergenic sequences and the *Myb* promoter, further implying they represent regulatory elements involved in *Myb* transcriptional regulation.

To further confirm the *Myb* promoter 3C(-Seq) data, the 3C-Seq was repeated using the −36 and −81 kb LDB1 complex binding sites as viewpoints. This showed that both sites interact with the *Myb* promoter and the adjacent CTCF-bound intron 1 fragment (Figure 2B). Additionally, there were multiple interactions detected between the −36 kb/−81 kb LDB1 complex binding sites and other TF and CTCF binding sites (Figure 2B). Collectively, the 3C data show that the active erythroid *Myb* promoter and intron 1 cluster with intergenic TF-bound elements to form a complex higher order chromatin structure. Of note, these data indicate that whereas the *Myb* gene promoter is found in close proximity to the distal enhancers, both the −36 and −81 kb regions also show a strong interaction with the intron 1 CTCF site as well.

The intron 1 CTCF element marks the start of productive transcription elongation and interacts with elongation factor-bound distal enhancers

Several studies have shown that *Myb* expression is regulated at the level of transcription elongation through an attenuation site in the first intron (Bender et al, 1987; Watson, 1988; Reddy and Reddy, 1989; Hugo et al, 2006) ~2 kb downstream of *Myb* exon 1, in the vicinity of the CTCF binding site identified in our study. Since this region interacts with the distal −36 and −81 kb elements, the intronic CTCF-bound element may actually mark the site of productive transcription elongation. Hence, ChIP experiments were carried out in erythroid progenitors to map the appearance of Serine 2 (Ser2)-phosphorylated polII (polII Ser2-P) and the H3K36 trimethylation (H3K36me3) mark, which are specifically associated with transcription elongation and peak within the transcribed region of genes (Brookes and Pombo, 2009; Buratowski, 2009; Figure 3A and C). As expected, no polII Ser2-P or H3K36me3 enrichments were detected at the promoter and upstream regions, whereas a sharp increase was seen starting around the CTCF binding site and increasing into the gene body (Figure 3A and C). Ser5-P polII on the other hand, representing the initiating polII state, specifically accumulated upstream of the CTCF site. In order to more precisely localize the transition from transcription initiation to productive elongation, ChIP-Seq for polII Ser5-P and H3K36me3 was performed in MEL cells. As shown in Figure 3B and D, the initiating polII signal covers the 5' end of the gene and extends up to the intronic CTCF site. In contrast, H3K36me3 starts to appear after the CTCF binding site in MEL cells. In addition, a recently published H3K36me3 data set from mouse primary erythroid progenitors (Wong et al, 2011) and data obtained from the human erythroid cell line K562 show a similar pattern (Supplementary Figure S5). These data suggest that the transition to productive transcription elongation occurs around the intronic CTCF site. ChIP experiments were used next to analyse the presence of the elongation factors CDK9 and TIF1γ at the *Myb* locus. CDK9 is a kinase that phosphorylates the Ser2 residue of the polII C-terminal domain (CTD), and is known to bind the LDB1 complex (Meier et al, 2006). TIF1γ was recently identified as a component of the LDB1 complex, regulating transcription elongation in haematopoietic cells, at least in part by allowing CDK9 recruitment to its target sites (Bai et al, 2010). CDK9 and TIF1γ showed only minor enrichments at the promoter and first intron (where polII Ser2-P appears), but surprisingly showed a much stronger occupancy at the upstream regulatory elements (Figure 3E and F). These experiments suggest that

Figure 2. Long-range genomic interactions within the *Myb-Hbs1l* locus. (A) 3C-Seq analysis of the *Myb* promoter-associated regions *in vivo*, using E12.5 mouse fetal liver (FL E12.5) and fetal brain (FB E12.5). Signals are presented as reads per millions per HindIII restriction fragment (vertical axis). The viewpoint (*Myb* promoter) is indicated by a red bar with an eye symbol. (B) Zoom-in view of the *Myb-Hbs1l* intergenic region. The ChIP-Seq profiles for LDB1 and CTCF (MEL) are shown together with the 3C-Seq signals obtained using the *Myb* promoter (top), the -36 kb (middle) and the -81 kb elements (bottom) as viewpoints (indicated by a red bar and eye symbol). Grey shading of HindIII fragments indicates sites where long-range interactions and transcription factor binding colocalize. The position of the HindIII restriction sites and the intergenic enhancers (relative to the *Myb* TSS) is indicated at the top. (C) Locus-wide crosslinking frequencies analysed by 3C-qPCR using the *Myb* promoter as viewpoint. Relative crosslinking frequencies observed in E12.5 FL (red) and FB (blue) are shown. Highest crosslinking frequencies per FL/FB pair tested were set to 1. The x axis shows the genomic coordinates of the interacting fragments in the locus. Data are plotted as mean±s.e.m. of at least three independent experiments.

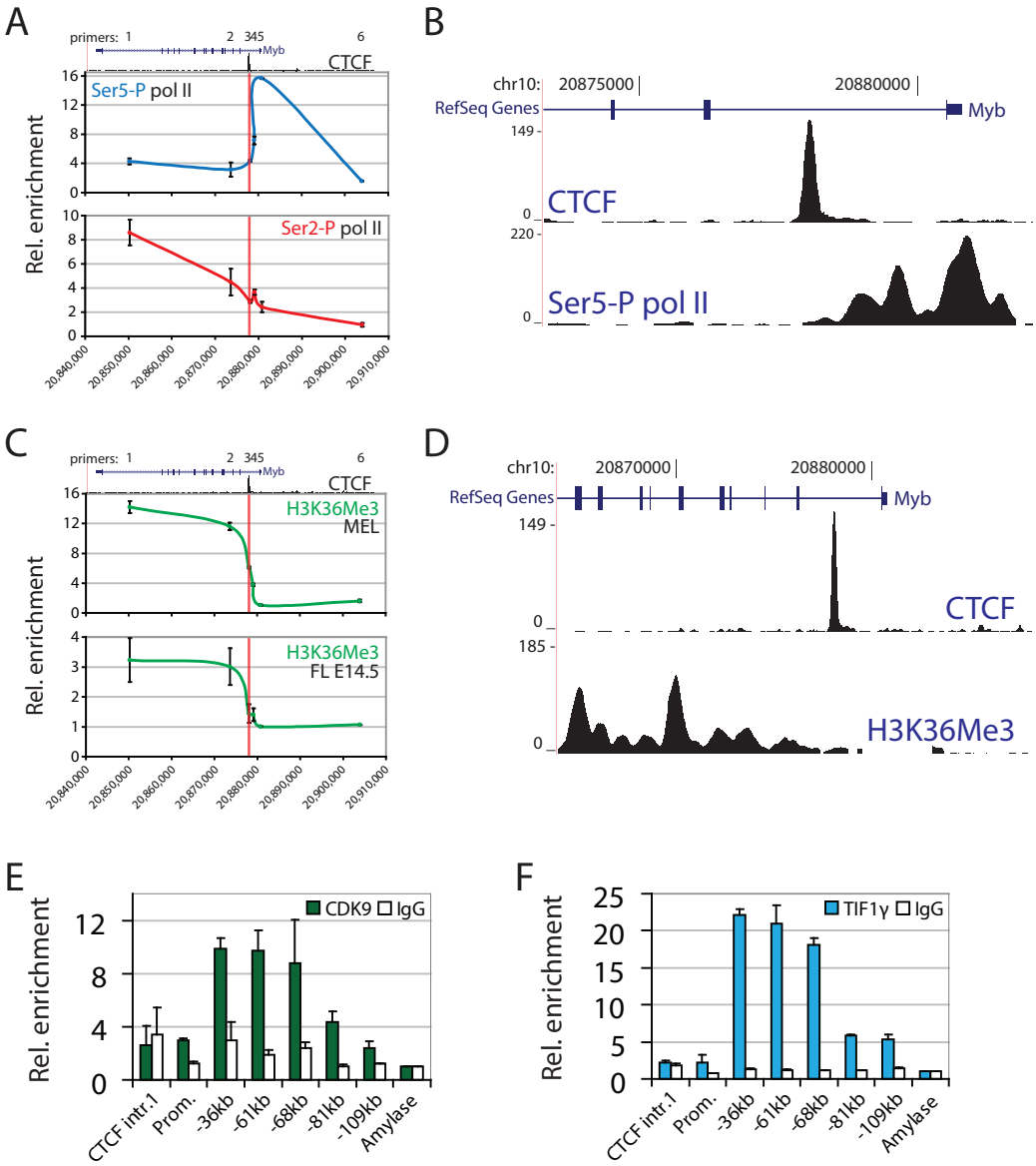


Figure 3. Transcription elongation starts in the vicinity of the *Myb* first intron CTCF element. (A, C) ChIP analysis showing the distribution of (A) polII phosphorylated at Ser5 (Ser5-P) and Ser2 (Ser2-P) in MEL cells, and (C) H3K36me3 in MEL and E14.5 fetal liver cells. UCSC Genome browser pictures depicting the *Myb* gene and CTCF binding in the first intron (ChIP-Seq) are shown above the graph. Primer pairs (1-6) used for PCR are indicated above the gene. The x-axis shows the genomic coordinates and the position of the CTCF binding site is indicated at the top (red vertical line). (B, D) ChIP-Seq profiles of CTCF, (B) Ser5-P pol II and (D) H3K36me3 obtained from MEL cells. (E, F) Occupancy of the *Myb-Hbs1l* intergenic region by the elongation factors (E) CDK9 and (F) TIF1γ in MEL cells as shown by ChIP. Enrichments were calculated versus a negative control region (amylase) and presented as the mean±s.e.m. of at least two independent experiments.

productive elongation is stimulated around the intronic CTCF site by positive elongation factors bound at the distal enhancer elements. These factors are likely to be brought in physical proximity to the elongation site by dynamic chromatin looping, where they can transiently carry out their enzymatic function. In support

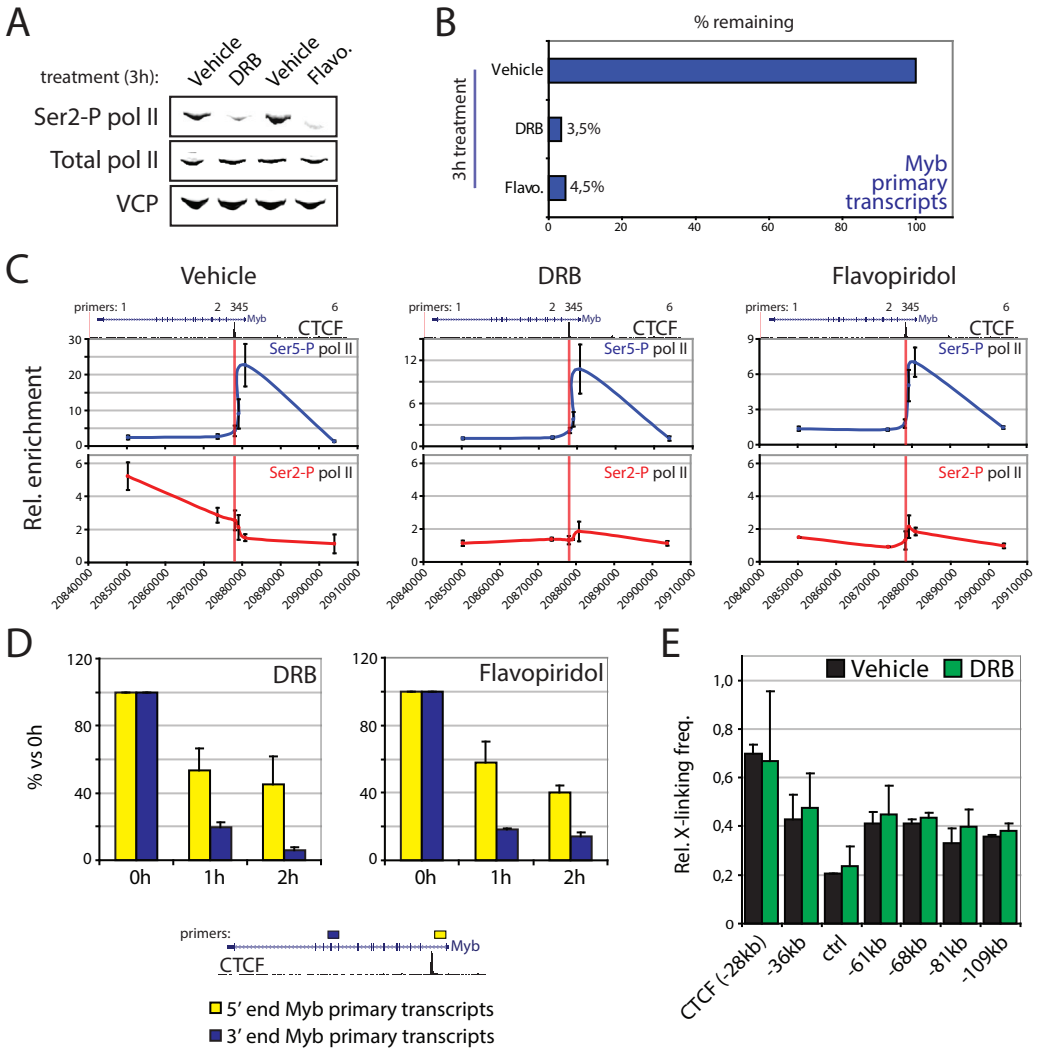


Figure 4. Effect of CDK9 inhibition on phosphorylated pol II occupancy, transcription and chromatin looping. (A) Western blot analysis of Ser2-P pol II and total pol II levels in vehicle-, DRB- or Flavopiridol (Flavo.)-treated MEL cells. Valosin Containing Protein (VCP) served as a loading control. (B) *Myb* primary transcripts measured by RT-qPCR after treatment with the indicated compounds. Signals were normalized to *18S* rRNA expression and transcript levels in vehicle-treated cells were set to 100%. (C) ChIP analysis of Ser5-P and Ser2-P pol II binding at the *Myb* transcriptional unit in vehicle-, DRB- and Flavopiridol-treated MEL cells. Genomic coordinates, gene location, CTCF occupancy and PCR primers are indicated above each graph (as in Figure 3). (D) Time-course CDK9 inhibition in MEL cells using DRB and Flavopiridol. *Myb* 5′ end and 3′ end transcripts were measured by RT-qPCR. Primer locations within the gene are depicted by coloured rectangles in a schematic below the graphs. (E) 3C-qPCR analysis on vehicle- or DRB-treated MEL cells. The *Myb* promoter HindIII fragment was used as a viewpoint. Data are plotted as mean±s.e.m. of at least two independent experiments.

of this notion, depletion of TIF1γ in primary human erythroid cells resulted in a severe reduction of *Myb* mRNA levels (Bai et al, 2010). In addition, in order to prove that the Ser5-P polII enrichments observed in the Myb first intron were specific and independent from productively elongating polII (Ser2-P), MEL cells were treated with the Cdk9 inhibitors DRB or flavopiridol. Under these conditions, a global loss of phosphorylated

Ser2 RNA polII was observed (Figure 4A) and *Myb* transcription was almost completely abolished (primary transcripts are decreased by >95%, Figure 4B). Importantly, ChIP experiments showed that the Ser5-P polII pattern on the *Myb* promoter and first intron was similar in vehicle-treated cells and cells treated with CDK9 inhibitors, while Ser2-P polII enrichments were lost (Figure 4C). Thus, the Ser5-P polII occupancy of the *Myb* promoter and first intron up to the CTCF site is independent of ongoing transcriptional elongation.

In support of this, while CDK9 inhibition results in a loss of full-length transcripts, transcription of the 5′ end of the gene is maintained and much less sensitive to CDK9 inhibition (Figure 4D). The 40–50% decrease in 5′ transcripts compared with vehicle-treated cells can be accounted for by the general ~50% decrease of Ser5-P polII at the promoter under these conditions (Figure 4C). Importantly, these data show that in the absence of Ser2 phosphorylation, RNA polII is still able to engage at the *Myb* gene and is still transcribing the first ~2 kb (i.e., up to the CTCF site) but is unable to bypass this site and progress throughout the gene efficiently. Interestingly, the long-range interactions are maintained upon DRB treatment (Figure 4E).

Erythroid differentiation is accompanied by decreased Myb expression and a loss of chromatin looping

In order to correlate the long-range interactions observed in the *Myb-Hbs1l* locus with *Myb* transcriptional activity, *Myb* expression and locus structure were analysed during erythroid differentiation. Differentiation of MEL cells or mouse E13.5 FL primary erythroid progenitors resulted in a strong decrease in *Myb*, but not *Hbs1l* or *Ahi1* primary transcription (Figure 5A). Erythroid maturation of MEL and FL cells was monitored by analysing the activation of the two terminal differentiation markers Glycophorin A (*Gypa*) and Beta-Major (*Hbb-b1*) (Supplementary Figure S6A), as well as the characteristic decrease in cell size of the primary progenitors (Supplementary Figure 6B). Thus, significant downregulation of *Myb* transcription occurs upon terminal erythroid differentiation, while the flanking genes show stable or modestly increasing expression levels. 3C-Seq was subsequently carried out using the *Myb* promoter as viewpoint in MEL cells before and after differentiation, representing stages of high and low *Myb* expression, respectively (Figure 5B). In non-differentiated MEL cells, the *Myb* promoter showed a long-range interaction pattern very similar to that seen in primary erythroid progenitors (Figure 2B). However, opposite to what was observed for the β-globin locus (Palstra et al, 2003), the frequency of most intergenic contacts was strikingly diminished upon differentiation. This loss of interaction was observed essentially for all LDB1 complex, KLF1- and CTCF-bound fragments of the locus (Figure 5B). 3C-Seq experiments using mouse dermal fibroblasts (MDF, which do not express *Myb*) confirmed the erythroid-specific nature of the interactions (Figure 5B). The loss of chromatin looping in both MEL cells and primary erythroid progenitors upon erythroid differentiation was confirmed by the more quantitative 3C-qPCR method (Figure 5C). These data show that *Myb* downregulation upon erythroid differentiation is accompanied by a loss of communication between the *Myb* promoter and the intergenic TF-bound enhancers.

Decreased transcription and elongation factor occupancy at the intergenic enhancers upon erythroid differentiation

The long-range interactions between *Myb* and the intergenic enhancers are lost upon differentiation, clearly paralleling *Myb* downregulation. However, it is unclear what underlies the loss of looping and expression. To address this question, quantitative ChIP experiments were performed in MEL cells before and after differentiation to analyse intergenic TF occupancy during erythroid maturation. An overall decrease in LDB1 complex (Figure 6A), KLF1 (Figure 6B) and elongation factor (Figure 6C) occupancy was seen at the intergenic binding sites upon differentiation. In agreement with this, the levels of enhancer-associated histone modifications and proteins often decrease as well (Supplementary Figure S7A). Furthermore, ChIP-Seq showed changing polII occupancy of the *Myb* transcription unit during differentiation (Supplementary Figure S7B). In both undifferentiated and differentiated cells, polII accumulates at the promoter and first intron (high signals), up to the CTCF site. In undifferentiated cells, polII actively bypasses this site and progresses into the gene, whereas in differentiated cells a strong reduction of polII beyond the CTCF site is seen (Supplementary Figure S7B). Thus, a loss of activating proteins at the intergenic regulatory elements upon differentiation coincides with losses of long-range interactions, polII progression into the gene body, and expression. Interestingly, initiation still appears to take place, as previously suggested (Bender et al,

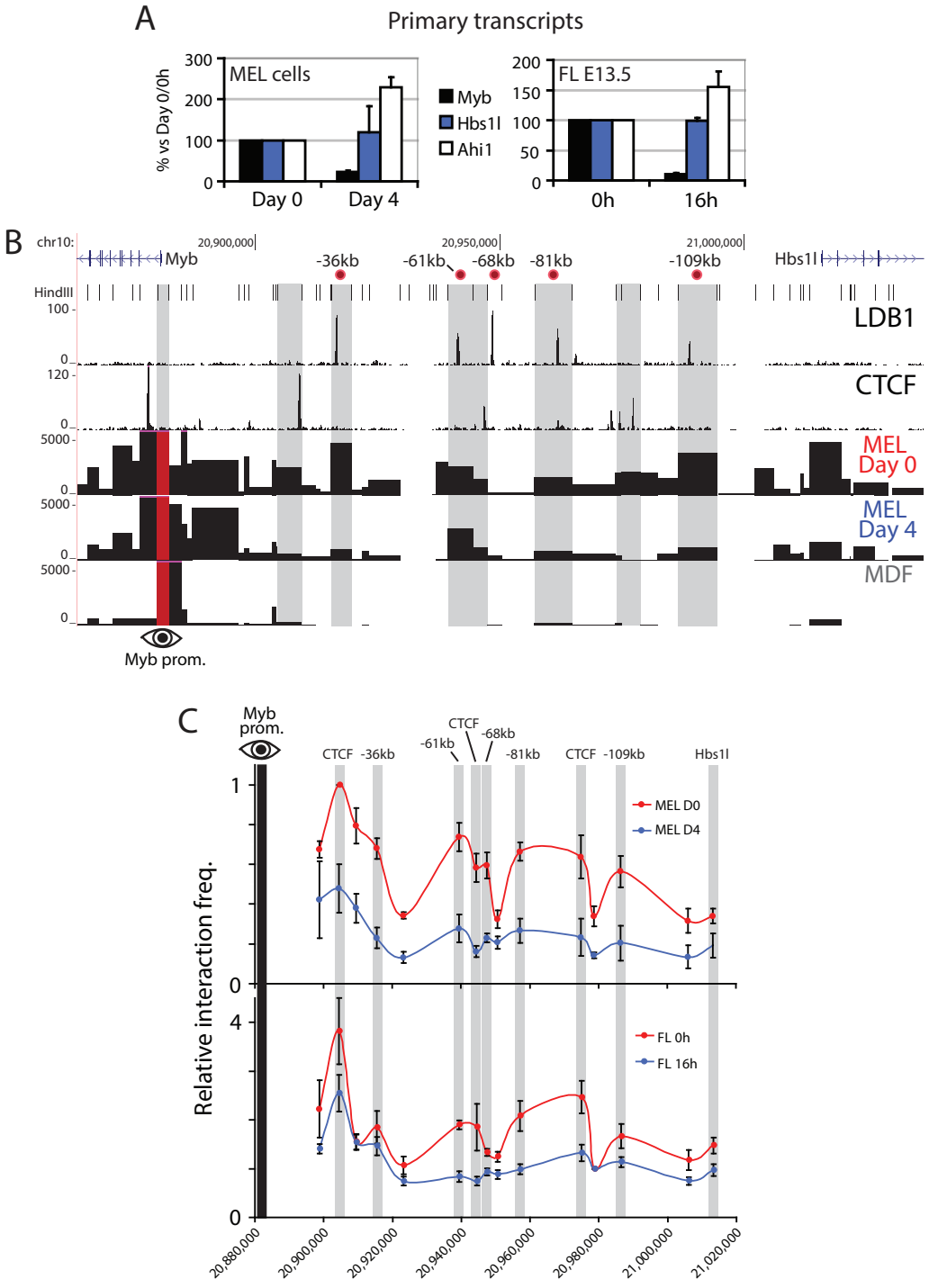


Figure 5. *(Legend at the bottom of the next page)*

1987).

LDB1 and KLF1 are essential for high Myb expression in erythroid progenitors

LDB1 and KLF1 have recently been implicated in long-range gene regulation (Drissen et al, 2004; Song et al, 2007; Tallack et al, 2010). KLF1 selectively occupies the –81-kb site in the *Myb-Hbs1l* intergenic region, while LDB1 binds all five regulatory elements (Figure 1). Since intergenic binding of both proteins decreased as *Myb* transcription is downregulated, we hypothesized that loss of KLF1 or LDB1 in erythroid progenitors would result in decreased *Myb* expression. To verify this, short hairpin RNAs (shRNAs) against *Klf1* or *Ldb1* mRNA were used to reduce their respective protein levels in MEL cells. A 50–80% decrease in mRNA and protein was observed when compared with cells transduced with a control lentivirus (Figure 7A and B). Both knockdowns resulted in a 50% decrease of *Myb* transcription, while the flanking *Hbs1l* and *Ahi1* genes were not significantly affected (Figure 7C and D). Similarly, knocking down the expression of CTCF also results in a significant reduction of *Myb* transcription, without affecting *Hbs1l* or *Ahi1* (Supplementary Figure S8C). The decrease in *Myb* primary transcripts was not caused by cellular differentiation or a change in TF levels due to LDB1 or KLF1 depletion, as we observed no changes compatible with erythroid maturation in the expression of late erythroid markers (*Gypa* and *Hbb-b1*) or key erythroid TFs (Supplementary Figure S8A and B). These results are in agreement with reports showing reduced *Myb* expression *in vivo* in *Klf1*$^{-/-}$ FL (Pilon et al, 2008), and a 50% decrease of *Myb* expression in bone marrow haematopoietic progenitors conditionally depleted for LDB1 (Li et al, 2011). Since LDB1 is a scaffold-like protein important for TF complex assembly and chromatin looping, locus conformation was analysed by 3C-qPCR after LDB1 knockdown (Figure 7E). This showed that LDB1 depletion indeed results in reduced long-range promoter–enhancer contacts (Figure 7E), further emphasizing its key role in chromatin loop formation.

Discussion

The expression of the *Myb* proto-oncogene in haematopoietic cells is subjected to very tight control to properly coordinate cellular proliferation and differentiation. Given that enforced *Myb* expression impairs haematopoietic differentiation and that aberrant *Myb* expression associates with haematopoietic malignancies, deciphering *Myb* transcriptional control is crucial for a better understanding of both normal haematopoietic development and associated disorders.

TF binding and long-range interactions at the Myb-Hbs1l locus

A combination of ChIP-Seq and 3C-Seq was used to map the genome-wide binding sites of critical transcription and structural factors, and to characterize the spatial interactions within the *Myb* locus. 3C-Seq offers an advantage over array-based 4C technology to map long-range genomic interactions at the level of a single locus (in addition to a genome-wide level), since it does not suffer from saturating signals surrounding the viewpoints (Simonis et al, 2006; Soler et al, 2010). It is therefore well suited to analyse locus-wide chromatin looping within tens of kilobases up to megabases without prior knowledge of the interaction sites. Combining ChIP-Seq and 3C-Seq shows that the *Myb-Hbs1l* intergenic region harbours important regulatory elements controlling *Myb* expression, that bind either the structural protein CTCF or the essential erythroid TFs GATA1, LDB1, TAL1 and KLF1. The sites that bind KLF1 and the GATA1/TAL1/

Figure 5. Erythroid differentiation is accompanied by a loss of *Myb* transcription and long-range genomic interactions. (A) Primary transcript levels of *Myb*, *Hbs1l* and *Ahi1* during terminal differentiation of MEL (left panel) and E13.5 fetal liver (FL) erythroid progenitors (right panel). MEL cells were induced for 4 days in the presence of 2% DMSO. Fetal liver cells were cultured *ex vivo* for 16 h in differentiation medium. Data are expressed as percentages of expression versus day 0 (MEL) or 0 h (FL) of differentiation. Signals were normalized to *Rnh1* or *Calr* expression, and day 0 (MEL) or 0 h (primary cells) values were set to 100 (i.e., undifferentiated cells). Results are plotted as mean±s.e.m. of three independent experiments. (B) 3C-Seq analysis of the *Myb* promoter-associated regions in undifferentiated MEL cells (MEL day 0) and differentiated MEL cells (MEL day 4). Mouse dermal fibroblasts (MDFs) were used as a negative control (no *Myb* expression). Results are represented as in Figure 2. (C) Analysis of the *Myb-Hbs1l* locus conformation by 3C-qPCR in differentiating MEL and FL cells. See Figure 2C for details.

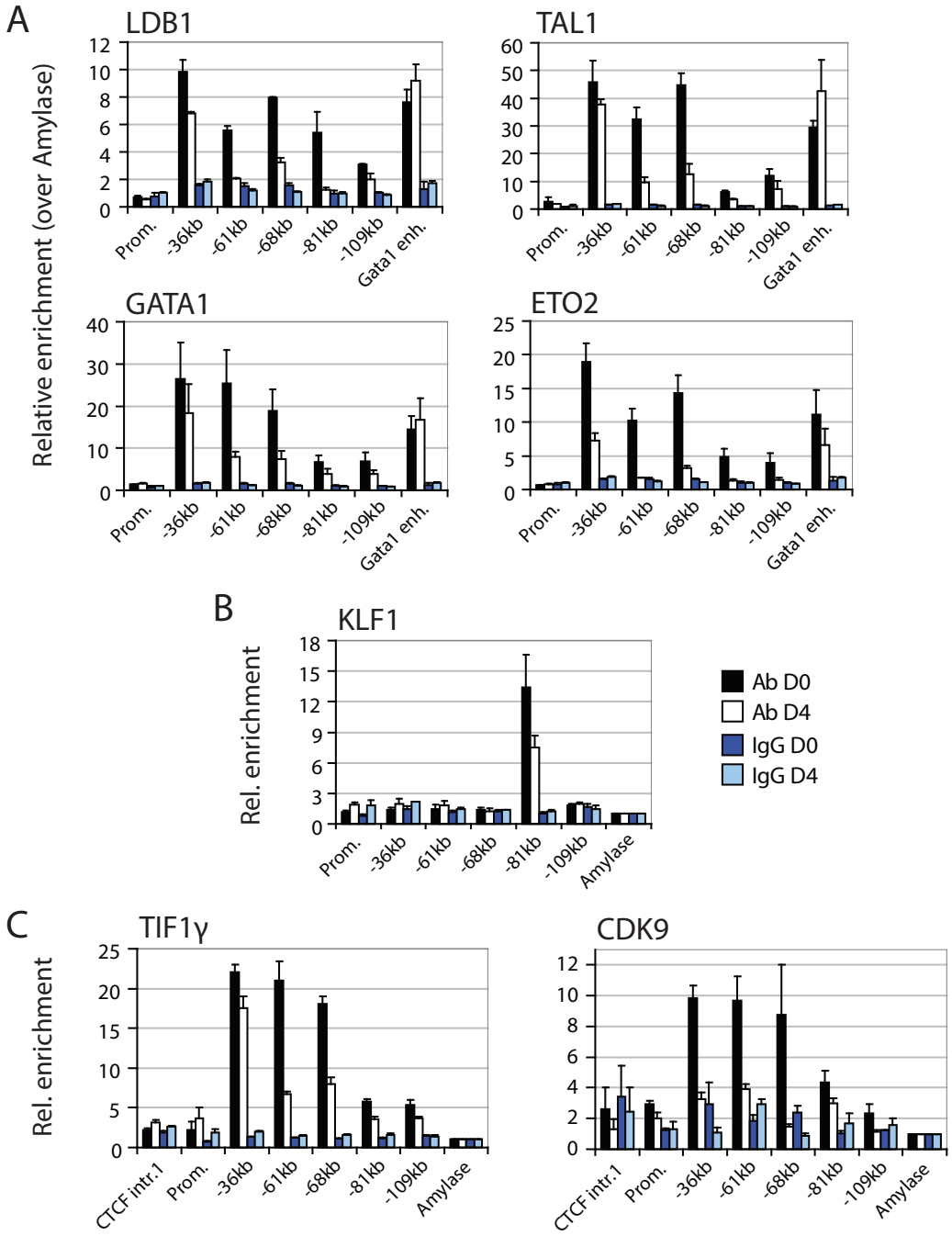


Figure 6. Erythroid differentiation induces a dramatic decrease of transcription factor occupancy within the *Myb-Hbs1l* intergenic region. (A–C) MEL cells were treated with 2% DMSO for 4 days to induce erythroid differentiation. Chromatin occupancy of (A) LDB1, TAL1, GATA1, ETO2, (B) KLF1, (C) TIF1γ and CDK9 was examined by ChIP in undifferentiated (D0) and differentiated (D4) MEL cells. Amylase served as a negative control region. Data are presented as mean±s.e.m. of 2–4 independent experiments.

LDB1 complex are transcriptional enhancers, confirming the positive role of these factors on erythroid gene expression (Figure 1; Soler et al, 2010; Tallack et al, 2010).The 3C-Seq genomic interaction profiles show an erythroid-specific pattern of interactions between the *Myb* promoter, first intron and intergenic enhancers (Figure 2), which is highly similar for primary erythroid progenitors and MEL cells. CTCF, KLF1 and GATA1/TAL1/LDB1 binding sites were shown to mark the sites of long-range genomic interactions. The reproducibility between different biological materials, 3C-qPCR validations and the clear overlap between long-range interactions and TF binding further validate the specificity of the 3C-Seq profiles.

Both KLF1 and LDB1 activate Myb expression, and the LDB1 complex is required to establish spatial proximity between Myb and the distal intergenic enhancers

We show here a requirement for KLF1 and LDB1 in maintaining high levels of *Myb* expression in erythroid progenitors (Figure 7). Reducing the level of either of these factors results in a 50% decrease of *Myb* transcription without inducing erythroid differentiation (Supplementary Figure S8). This suggests that *Myb* downregulation coincides with, but is not a driver of differentiation. The DNA-binding erythroid Kruppel-like factor KLF1 is the founding member of the mammalian Kruppel-like family of zinc-finger TFs. It recognizes CACCC-box motifs often found in erythroid-specific gene promoters and is required for their activation. KLF1 binds a single location in the *Myb-Hbs1l* locus at the –81-kb enhancer, which contains a conserved CACCC-box motif. The positive role of KLF1 on erythroid gene expression is confirmed by our finding that KLF1 activates *Myb* transcription. Interestingly, *Klf1*$^{-/-}$ mouse embryos die around E15 from a lack of definitive erythropoiesis, resulting in severe anaemia (Nuez et al, 1995; Perkins et al, 1995). This phenotype shares similarities with the lethal anaemia of *Myb*$^{-/-}$ embryos which die around E15 (Mucenski et al, 1991). It has been shown that E13.5 *Klf1*$^{-/-}$ FL-derived erythroid cells fail to progress through the last cell cycles of terminal erythroid differentiation, in part due to misregulation of the G1-to-S phase transition TFs E2F2 and E2F4 (Pilon et al, 2008; Tallack et al, 2009). The phenotypic similarities between *Klf1*$^{-/-}$ and *Myb*$^{-/-}$ mouse models, the strong downregulation of *Myb* in *Klf1*$^{-/-}$ FL cells (Pilon et al, 2008) and the implication of *Myb* in the G1-to-S transition (Oh and Reddy, 1999) suggest that *Myb* misregulation in *Klf1*$^{-/-}$ cells also contributes significantly to the observed proliferative defect.

The widely expressed nuclear adaptor LDB1 functions as a core component of multiprotein complexes, regulating the development of many tissues. The LDB1 protein itself has no known DNA-binding or enzymatic activities. In erythroid cells, LDB1 forms a complex with the DNA-binding TFs GATA1, TAL1 (SCL), E2A and the cofactors LMO2/LMO4 and ETO2/MTGR1. In addition, the LDB1 complex interacts with transcription elongation factors, like TIF1γ and CDK9, a kinase known to regulate transcription elongation through phosphorylation of the polII CTD at Ser2. Consistent with its essential functions, *Ldb1*$^{-/-}$ mouse embryos do not develop beyond the E10 stage and show dramatic developmental defects including a lack of haematopoiesis (Mukhopadhyay et al, 2003; Li et al, 2010). Due to the early lethal phenotype, the role played by LDB1 during haematopoiesis *in vivo* remained largely unexplored. Recent data, however, showed a continuous requirement for LDB1 in the maintenance and differentiation of haematopoietic stem cells, and in the development of the lymphoid, erythroid and megakaryocytic lineages (Li et al, 2010, 2011). LDB1 is required to activate the late erythroid gene expression program (Li et al, 2010; Soler et al, 2010) and it exerts this function at least in part by facilitating long-range interactions between remote enhancers and their target genes (Song et al, 2007). Our analysis of the *Myb-Hbs1l* locus conformation shows that in erythroid progenitors expressing *Myb* at high levels, the enhancers are clustered in the nuclear space to form an ACH structure resembling the one observed within the active β-globin locus. We show here that LDB1 is required for the maintenance of the long-range interactions between the *Myb* gene and the upstream enhancers. Reducing the level of LDB1 in erythroid progenitors results in a decrease of *Myb* promoter–enhancer interactions and transcription (Figure 7). Interestingly, transcription of the neighbouring genes *Hbs1l* and *Ahi1* remains unaffected under these conditions, even though *Ahi1* harbours a binding site for the LDB1 complex in its first intron (Soler et al, 2010). During the course of erythroid differentiation, when *Myb* transcription is downregulated dramatically, the long-range interactions are reduced, resulting in a loss of the ACH (Figure 5). This loss of long-range communication is explained by the decreased occupancy of the LDB1 complex at the intergenic enhancers (Figure 6). Interestingly, decreasing the level of LDB1 results in a loss of all interactions, not just those bound by LDB1. This suggests that in order to be maintained and stabilized, the chromatin hub requires several if not all the interactions (i.e., the enhancer sites and the CTCF

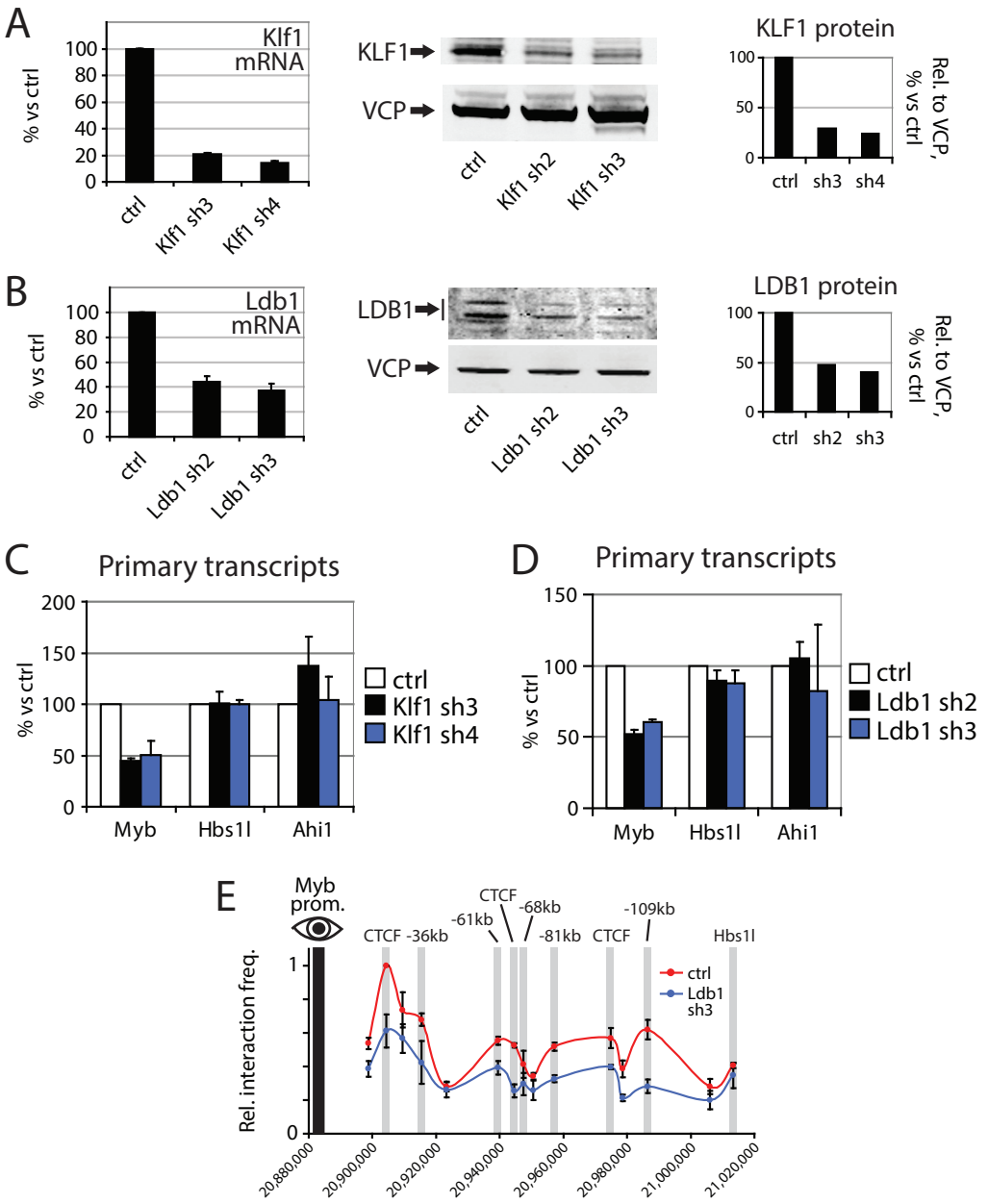


Figure 7. LDB1 and KLF1 positively regulate *Myb* expression. (A, B) Two independent shRNAs were used to decrease (A) *Klf1* and (B) *Ldb1* expression in MEL cells. Knockdown efficiency was measured at the mRNA and protein levels. Results are compared with a non-targeting scrambled shRNA. Valosin Containing Protein (VCP) served as a loading control for protein analysis. (C, D) Effect of (C) *Klf1* and (D) *Ldb1* knockdowns on *Myb*, *Hbs1l* and *Ahi1* primary transcript levels. (E) The effect of LDB1 depletion on chromatin looping was measured by 3C-qPCR using the *Myb* promoter as viewpoint. The interaction frequencies in control and LDB1-depleted samples are shown in red and blue, respectively. Data are plotted as mean±s.e.m. of at least three independent experiments.

sites). Accordingly, affecting the binding of LDB1 on some sites would induces a destabilization of the whole structure, and thus have an impact on sites not bound by the protein but normally present in the hub.

Strikingly, this observation contrasts with the general increase of binding of the LDB1 complex on induced erythroid genes during terminal differentiation (Soler et al, 2010). A mechanistic explanation for this selective loss of the LDB1 complex from the *Myb-Hbs1l* locus could be that, in the late stages of differentiation, additional TFs start competing for binding or induce a local destabilization or degradation of the complex.

Heterogeneity between the distal enhancer elements

It is not clear whether *Myb* requires the entire intergenic region for full activation, because the individual contributions of the different intergenic enhancers are unknown. They could play an additive role to ensure high local concentrations of positive transcriptional regulators and therefore high levels of transcription. Alternatively, they might be required to stabilize the chromatin hub at the *Myb* gene and first intron. Such a multi-component complex structure has already been observed for developmentally regulated genes like globins. In that case the activity of the elements appears additive, although they are individually clearly different in structure and activity. For the *Myb* locus, the elements also appear to be different in function. They show different enhancer activity *in vitro*, and differ in protein occupancy. Indeed, whereas all elements are enriched for the core components of the LDB1 complex and enhancer-associated histone modifications/proteins, the −81-kb enhancer shows a 5- to 7.5-fold higher enrichment for polII and is the only one bound by KLF1 (Figure 1), a factor essential for *Myb* transcription (Figure 7). The −81-kb element also shows a high degree of sequence conservation between mouse and human. This regulatory element is therefore likely to play a key role in the transcriptional activation of the locus. Conditional deletion of the individual enhancers will provide crucial information about their role(s) *in vivo*, in particular whether the −81-kb element represents an enhancer with a specialized function.

Transcription and elongation factors at distal regulatory elements: a model for Myb transcriptional activation during development

Our data are in agreement with previous reports highlighting the regulatory potential and the importance of the *Myb-Hbs1l* intergenic region for *Myb* transcriptional regulation (Mukai et al, 2006; Wahlberg et al, 2009). In addition, the presence of regulatory elements within the *Myb* first intron affecting transcription elongation has been reported >20 years ago, although their role is still not fully understood (Bender et al, 1987; Hugo et al, 2006). An attenuation element was mapped in the first intron, where a poly-T tract was predicted to yield a stem-loop structured nascent RNA. Based on this finding, it was speculated that the stable intronic stem-loop transcript might provide a docking site for RNA-binding proteins to overcome the transcription elongation block in a way similar to the HIV TAR stem-loop RNA (Ramsay and Gonda, 2008). Although we cannot exclude this hypothesis, our data indicate that the intronic transcription elongation region corresponds to a domain containing a highly conserved CTCF binding site (coinciding with the start of the Ser2-P polII and H3K36me3 elongation signature, Figure 3), which appears to function in combination with the upstream elements. For example, the −36 and −81 kb enhancers loaded with erythroid TFs, polII and the elongation factors CDK9 and TIF1γ loop towards the *Myb* intron 1 CTCF site (Figure 2B). As the intergenic elements actively cluster together (Figure 2B) and are also bound by transcription and elongations factors (Figures 1 and 3), they are likely to contribute to the stimulation of transcription elongation. To further support this idea, we carried out CDK9 inhibition experiments (Figure 4). As stated above, CDK9 is primarily bound to the upstream regulatory elements. Its inhibition resulted in a loss of elongating (Ser2-P) polymerase and 3′ *Myb* transcription, while the initiating (Ser5-P) polymerase and 5′ transcription were retained, without affecting looping. A plausible explanation would therefore be that CDK9 is brought to the intronic transition site by looping, as represented in our model (Figure 8). As the chromatin loops were still able to form under these conditions (Figure 4E), they may have become 'non-functional' due to an inability to provide kinase activity.

Interestingly, a role for the β-globin LCR in the transition from transcriptional initiation to elongation has been proposed (Sawado et al, 2003). Indeed, both CDK9 and TIF1γ bind the LCR (unpublished observation). It remains to be tested whether the *Myb* and globin ACHs fulfil similar tasks in the transition

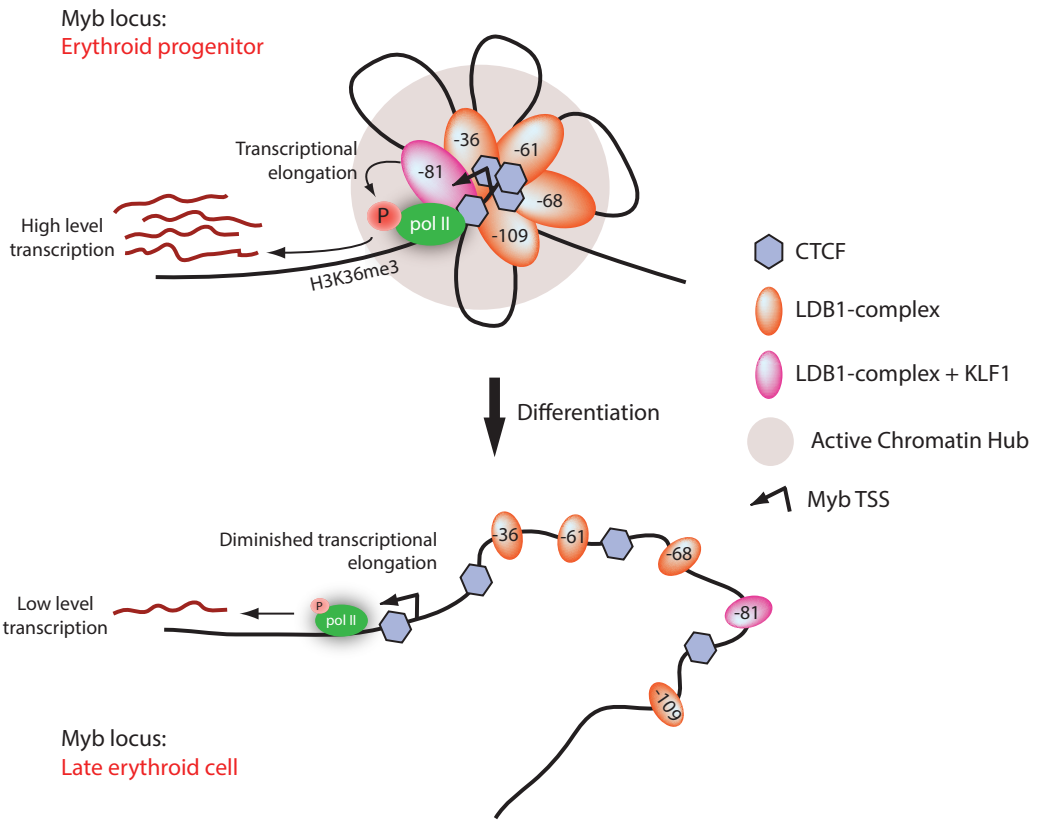


Figure 8. Model of the dynamic transcriptional regulation of *Myb* in differentiating erythroid cells. The *Myb* Active Chromatin Hub (ACH, grey sphere) is a structured nuclear compartment containing clustered cis-regulatory elements enriched for activating transcription factor complexes containing transcription elongation factors (orange and pink ovals) and CTCF (blue diamonds). The ACH provides a local high concentration of polII, transcription and elongation factors around the *Myb* gene, allowing for high-level expression in erythroid progenitors. During differentiation, intergenic transcription factor occupancy decreases (small ovals) at the cis-regulatory elements, leading to a destabilization of the ACH and a dramatic decrease of *Myb* transcription, allowing cells to terminally differentiate.

to productive elongation. In the *Myb* locus, the presence of CTCF is likely to play a key role in orchestrating the long-range interactions (Splinter et al, 2006) and its presence is required for high level *Myb* expression (Supplementary Figure S8C). The intronic CTCF site may mark a transcriptional barrier element preventing polII from progressing further into the gene body (Supplementary Figure S7B). However, when the distal enhancers are loaded with TFs, polII and elongation factors, it would serve as an anchoring site for the enhancers to form an ACH. Clustering all the factors around the *Myb* promoter and intronic productive elongation site would then override the transcriptional block in erythroid progenitors to allow *Myb* transcription at a high rate (Figure 8, upper half). The presence of a previously suggested structured nascent RNA (Thompson et al, 1997; Hugo et al, 2006; Ramsay and Gonda, 2008) could locally cause polII to slow down, thereby increasing the chance of phosphorylation by the elongation factors bound at the distal elements. Both mechanisms could thus participate in the elongation checkpoint operating at the *Myb* intronic attenuation region. During terminal differentiation, the ACH is destabilized due to a loss of intergenic TF occupancy, resulting in decreased *Myb* transcription to allow the cells to fully mature (Figure 8, lower half).

Implications for development and disease

Since fluctuations in *Myb* expression are a common feature of differentiating haematopoietic cells, it is expected that similar mechanisms will take place in different lineages, probably using (part of) the intergenic regulatory elements described here, but bound by other lineage-specific TF complexes. Recent genome-wide studies in early haematopoietic stem/progenitor cells revealed the binding of several haematopoietic TFs on some of the *Myb* intergenic enhancers (Wilson et al, 2010; Li et al, 2011). It will be interesting to track enhancer usage and ACH formation during the course of haematopoietic stem/progenitor cell differentiation to the different lineages (e.g., myeloid versus lymphoid), and to investigate how the locus structure is affected in haematopoietic diseases like leukaemia. Importantly, our data provide a framework for further comparative analysis in human erythroid cells, where *MYB-HBS1L* allelic variants strongly associate with clinically relevant red blood cell traits and high fetal globin gene expression (Thein et al, 2007; Lettre et al, 2008; Ganesh et al, 2009; Galarneau et al, 2010), a crucial feature decreasing the severity of β-thalassaemia and sickle-cell anaemia. Several intergenic enhancers have high sequence conservation between mouse and human. Considering that the intronic CTCF and transcription elongation transition sites also seem to be conserved in human erythroid cells (Supplementary Figure S5A), a careful examination of the impact of intergenic SNPs on TF binding, chromatin looping and *MYB* expression in individuals bearing these SNPs will be of primary interest. A preliminary analysis of highly associated SNPs showed that some fall close to or within the conserved intergenic sequences, suggesting that they may affect regulation of *MYB* expression. However, to date we did not find clear examples where the variants either create or destroy a GATA1/LDB1 binding sequence motif. A more systematic analysis needs to be performed in order to better understand the functional impact of SNPs in the *MYB-HBS1L* intergenic region. It is likely that the impact of the variants may only have a mild effect on *MYB* expression, which may complicate the analyses. However, with recent reports implicating c-MYB in the regulation of human fetal haemoglobin expression (Jiang et al, 2006; Sankaran et al, 2011) and the maintenance of leukaemia in mice (Zuber et al, 2011), modulation of c-MYB levels could become an attractive therapeutic approach in the treatment of β-haemoglobinopathies and leukaemia.

Materials and methods

ChIP and ChIP-Seq procedures

ChIP and ChIP-Seq procedures were performed as described (Soler et al, 2010, 2011). ChIP-Seq samples were sequenced (36 bp reads) on the Illumina GAII platform and analysed by NARWHAL (Brouwer et al, 2011). Data were visualized using a local mirror of the UCSC genome browser.

3C and 3C-Seq procedures

The 3C and 3C-Seq libraries were prepared as described previously (Simonis et al, 2006; Soler et al, 2010; Supplementary Figure S3). HindIII was used as the primary restriction endonuclease. The 3C PCR signals were normalized as described (Palstra et al, 2003), with the highest crosslinking frequency set to 1. For 3C-Seq, either NlaIII (Myb prom and −36 kb viewpoints) or DpnII (−81 kb viewpoint) were used as secondary restriction enzymes. The 3C-Seq library was sequenced (76 bp reads) on the Illumina GAII platform.

For more detailed Materials and methods, see the Supplementary data.

Accession codes

The ChIP-Seq and 3C-Seq data sets were deposited to the Sequence Read Archive (the accession numbers for the ChIP-Seq were previously published (Soler et al, 2010). 3C-Seq data can be obtained using accession number SRA048225).

Supplementary data

Supplementary Material and Methods, as well as Supplementary Figures 1-8, are available at the EMBO Journal website.

Acknowledgments

We are grateful to Sjaak Philipsen for critical reading of the manuscript and to Jean-Christophe Andrau for helpful discussions. We thank Zeliha Ozgür for technical assistance and Rutger Brouwer for bioinformatics support. This work was supported by the EU-FP7 Eutracc consortium, the Netherlands Cancer Genomics Center (CGC) and the DFG SFB/Transregio5.

Author contributions: ES and FG conceived the study; RS, RJP, FG and ES designed the experiments; RS, CAS, EdB, AvdH, MS and ES performed the experiments; DE provided critical reagents and helpful comments; BL supervised informatics analyses; BL and ST designed the 3C-Seq analysis pipeline; ST, JCB and BL performed ChIP-Seq analysis; CK, AvdS and WvIJ performed ChIP-Seq and 3C-Seq DNA library preparation and Illumina sequencing; MvdH performed Illumina sequences alignments and data export; ES, FG and RS wrote the manuscript.

References

Bai X, Kim J, Yang Z, Jurynec MJ, Akie TE, Lee J, LeBlanc J, Sessa A, Jiang H, DiBiase A, Zhou Y, Grunwald DJ, Lin S, Cantor AB, Orkin SH, Zon LI (2010) TIF1gamma controls erythroid cell fate by regulating transcription elongation. Cell 142: 133–143.

Bender TP, Kremer CS, Kraus M, Buch T, Rajewsky K (2004) Critical functions for c-Myb at three checkpoints during thymocyte development. Nat Immunol 5: 721–729.

Bender TP, Thompson CB, Kuehl WM (1987) Differential expression of c-myb mRNA in murine B lymphomas by a block to transcription elongation. Science 237: 1473–1476.

Brookes E, Pombo A (2009) Modifications of RNA polymerase II are pivotal in regulating gene expression states. EMBO Rep 10: 1213–1219.

Brouwer RWW, van den Hout MCGN, Grosveld FG, van Ijcken WFJ (2011) NARWAHL, a primary analysis pipeline for NGS data. Bioinformatics (advance online publication, 8 November 2011; doi:10.1093/bioinformatics/btr613)

Buratowski S (2009) Progression through the RNA polymerase II CTD cycle. Mol Cell 36: 541–546.

Cantor AB, Orkin SH (2001) Hematopoietic development: a balancing act. Curr Opin Genet Dev 11: 513–519.

Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W (2004) The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev 18: 2485–2490.

Emambokus N, Vegiopoulos A, Harman B, Jenkinson E, Anderson G, Frampton J (2003) Progression through key stages of haemopoiesis is dependent on distinct threshold levels of c-Myb. EMBO J 22: 4478–4488.

Fujiwara T, O'Geen H, Keles S, Blahnik K, Linnemann AK, Kang YA, Choi K, Farnham PJ, Bresnick EH (2009) Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell 36: 667–681.

Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G (2010) Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet 42: 1049–1051.

Ganesh SK, Zakai NA, van Rooij FJ, Soranzo N, Smith AV, Nalls MA, Chen MH, Kottgen A, Glazer NL, Dehghan A, Kuhnel B, Aspelund T, Yang Q, Tanaka T, Jaffe A, Bis JC, Verwoert GC, Teumer A, Fox CS, Guralnik JM et al. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet 41: 1191–1198.

Gonda TJ, Metcalf D (1984) Expression of myb, myc and fos proto-oncogenes during the differentiation of a murine myeloid leukaemia. Nature 310: 249–251.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108–112.

Hugo H, Cures A, Suraweera N, Drabsch Y, Purcell D, Mantamadiotis T, Phillips W, Dobrovic A, Zupi G, Gonda TJ, Iacopetta B, Ramsay RG (2006) Mutations in the MYB intron I regulatory sequence increase transcription in colon cancers. Genes Chromosomes Cancer 45: 1143–1154.

Jiang J, Best S, Menzel S, Silver N, Lai MI, Surdulescu GL, Spector TD, Thein SL (2006) cMYB is involved in the regulation of fetal hemoglobin production in adults. Blood 108: 1077–1083.

Jing H, Vakoc CR, Ying L, Mandat S, Wang H, Zheng X, Blobel GA (2008) Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. Mol Cell 29: 232–242.

Kassouf MT, Hughes JR, Taylor S, McGowan SJ, Soneji S, Green AL, Vyas P, Porcher C (2010) Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res 20: 1064–1083.

Lettre G, Sankaran VG, Bezerra MA, Araujo AS, Uda M, Sanna S, Cao A, Schlessinger D, Costa FF, Hirschhorn JN, Orkin SH (2008) DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc Natl Acad Sci USA 105: 11869–11874.

Li L, Jothi R, Cui K, Lee JY, Cohen T, Gorivodsky M, Tzchori I, Zhao Y, Hayes SM, Bresnick EH, Zhao K, Westphal H, Love PE (2011) Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. Nat Immunol 12: 129–136.

Li L, Lee JY, Gross J, Song SH, Dean A, Love PE (2010) A requirement for Lim domain binding protein 1 in erythropoiesis. J Exp Med 207: 2543–2550.

Lieu YK, Reddy EP (2009) Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation. Proc Natl Acad Sci USA 106: 21689–21694.

Lu J, Guo S, Ebert BL, Zhang H, Peng X, Bosco J, Pretz J, Schlanger R, Wang JY, Mak RH, Dombkowski DM, Preffer FI, Scadden DT, Golub TR (2008) MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. Dev Cell 14: 843–853.

Meier N, Krpic S, Rodriguez P, Strouboulis J, Monti M, Krijgsveld J, Gering M, Patient R, Hostert A, Grosveld F (2006) Novel binding partners of Ldb1 are required for haematopoietic development. Development 133: 4913–4923.

Mucenski ML, McLain K, Kier AB, Swerdlow SH, Schreiner CM, Miller TA, Pietryga DW, Scott WJ Jr, Potter SS (1991) A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. Cell 65: 677–689.

Mukai HY, Motohashi H, Ohneda O, Suzuki N, Nagano M, Yamamoto M (2006) Transgene insertion in proximity to the c-myb gene disrupts erythroid-megakaryocytic lineage bifurcation. Mol Cell Biol 26: 7953–7965.
Mukhopadhyay M, Teufel A, Yamashita T, Agulnick AD, Chen L, Downs KM, Schindler A, Grinberg A, Huang SP, Dorward D, Westphal H (2003) Functional ablation of the mouse Ldb1 gene results in severe patterning defects during gastrulation. Development 130: 495–505.
Nuez B, Michalovich D, Bygrave A, Ploemacher R, Grosveld F (1995) Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. Nature 375: 316–318.
Oh IH, Reddy EP (1999) The myb gene family in cell growth, differentiation and apoptosis. Oncogene 18: 3017–3033.
Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12: 283–293.
Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, de Laat W (2003) The beta-globin nuclear compartment in development and erythroid differentiation. Nat Genet 35: 190–194.
Perkins AC, Sharpe AH, Orkin SH (1995) Lethal beta-thalassaemia in mice lacking the erythroid CACCC-transcription factor EKLF. Nature 375: 318–322.
Pilon AM, Arcasoy MO, Dressman HK, Vayda SE, Maksimova YD, Sangerman JI, Gallagher PG, Bodine DM (2008) Failure of terminal erythroid differentiation in EKLF-deficient mice is associated with cell cycle perturbation and reduced expression of E2F2. Mol Cell Biol 28: 7394–7401.
Ramsay RG, Gonda TJ (2008) MYB function in normal and cancer cells. Nat Rev Cancer 8: 523–534.
Reddy CD, Reddy EP (1989) Differential binding of nuclear factors to the intron 1 sequences containing the transcriptional pause site correlates with c-myb expression. Proc Natl Acad Sci USA 86: 7326–7330.
Sandberg ML, Sutton SE, Pletcher MT, Wiltshire T, Tarantino LM, Hogenesch JB, Cooke MP (2005) c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation. Dev Cell 8: 153–166.
Sankaran VG, Menne TF, Scepanovic D, Vergilio JA, Ji P, Kim J, Thiru P, Orkin SH, Lander ES, Lodish HF (2011) MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. Proc Natl Acad Sci USA 108: 1519–1524.
Sawado T, Halow J, Bender MA, Groudine M (2003) The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. Genes Dev 17: 1009–1018.
Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38: 1348–1354.
Soler E, Andrieu-Soler C, Boer E, Bryne JC, Thongjuea S, Rijkers E, Demmers J, Ijcken W, Grosveld F (2011) A systems approach to analyze transcription factors in mammalian cells. Methods 53: 151–162.
Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Stevens M, Kockx C, van Ijcken W, Hou J, Steinhoff C, Rijkers E, Lenhard B, Grosveld F (2010) The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev 24: 277–289.
Song SH, Hou C, Dean A (2007) A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. Mol Cell 28: 810–822.
Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev 20: 2349–2354.
Tallack MR, Keys JR, Humbert PO, Perkins AC (2009) EKLF/KLF1 controls cell cycle entry via direct regulation of E2f2. J Biol Chem 284: 20966–20974.
Tallack MR, Whitington T, Yuen WS, Wainwright EN, Keys JR, Gardiner BB, Nourbakhsh E, Cloonan N, Grimmond SM, Bailey TL, Perkins AC (2010) A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res 20: 1052–1063.
Thein SL, Menzel S, Peng X, Best S, Jiang J, Close J, Silver N, Gerovasilli A, Ping C, Yamaguchi M, Wahlberg K, Ulug P, Spector TD, Garner C, Matsuda F, Farrall M, Lathrop M (2007) Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. Proc Natl Acad Sci USA 104: 11346–11351.
Thomas MD, Kremer CS, Ravichandran KS, Rajewsky K, Bender TP (2005) c-Myb is critical for B cell development and maintenance of follicular B cells. Immunity 23: 275–286.
Thompson MA, Flegg R, Westin EH, Ramsay RG (1997) Microsatellite deletions in the c-myb transcriptional attenuator region associated with over-expression in colon tumour cell lines. Oncogene 14: 1715–1723.
Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W (2002) Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell 10: 1453–1465.
Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA (2005) Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell 17: 453–462.
Vegiopoulos A, Garcia P, Emambokus N, Frampton J (2006) Coordination of erythropoiesis by the transcription factor c-Myb. Blood 107: 4703–4710.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854–858.
Wahlberg K, Jiang J, Rooks H, Jawaid K, Matsuda F, Yamaguchi M, Lathrop M, Thein SL, Best S (2009) The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. Blood 114: 1254–1262.
Watson RJ (1988) A transcriptional arrest mechanism involved in controlling constitutive levels of mouse c-myb mRNA. Oncogene 2: 267–272.
Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, Pimanda JE, de Bruijn MF, Gottgens B (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7: 532–544.
Wong P, Hattangadi SM, Cheng AW, Frampton GM, Young RA, Lodish HF (2011) Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. Blood 118: e128–e138.
Xiao C, Calado DP, Galler G, Thai TH, Patterson HC, Wang J, Rajewsky N, Bender TP, Rajewsky K (2007) MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. Cell 131: 146–159.
Yu M, Riva L, Xie H, Schindler Y, Moran TB, Cheng Y, Yu D, Hardison R, Weiss MJ, Orkin SH, Bernstein BE, Fraenkel E, Cantor AB (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol Cell 36: 682–695.
Zuber J, Rappaport AR, Luo W, Wang E, Chen C, Vaseva AV, Shi J, Weissmueller S, Fellmann C, Taylor MJ, Weissenboeck M, Graeber TG, Kogan SC, Vakoc CR, Lowe SW (2011) An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. Genes Dev 25: 1628–1640.

Chapter 6

HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range *MYB* enhancers

Ralph Stadhouders[1]*, Suleyman Aktuna[2]*, Supat Thongjuea[3,4], Ali Aghajanirefah[1], Farzin Pourfarzad[1], Wilfred van IJcken[5], Boris Lenhard[3,6], Helen Rooks[2], Steve Best[2], Stephan Menzel[2], Frank Grosveld[1,7], Swee Lay Thein[2,8]† & Eric Soler[1,7,9]†

[1]Department of Cell Biology, Erasmus MC, Rotterdam, Netherlands.
[2]King's College London, Department of Molecular Haematology, London, United Kingdom.
[3]Computational Biology Unit, Bergen Center for Computational Science, Bergen, Norway.
[4]MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, United Kingdom.
[5]Centre for Biomics, Erasmus Medical Centre, Rotterdam, Netherlands.
[6]Department of Molecular Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, and MRC Clinical Sciences Centre, London, United Kingdom.
[7]Cancer Genomics Center, Erasmus Medical Center, Rotterdam, Netherlands.
[8]King's College Hospital Foundation Trust, London, United Kingdom.
[9]INSERM UMR967, CEA/DSV/iRCM, Fontenay-aux-Roses, France.

***These authors contributed equally.**
†Corresponding authors.

Published in:
Journal of Clinical Investigation
2014 ; 124:1699-710

Abstract

Genetic studies have identified common variants within the intergenic region (*HBS1L-MYB*) between GTP-binding elongation factor *HBS1L* and myeloblastosis oncogene *MYB* on chromosome 6q that are associated with elevated fetal hemoglobin (HbF) levels and alterations of other clinically important human erythroid traits. It is unclear how these noncoding sequence variants affect multiple erythrocyte characteristics. Here, we determined that several *HBS1L-MYB* intergenic variants affect regulatory elements that are occupied by key erythroid transcription factors within this region. These elements interact with *MYB*, a critical regulator of erythroid development and HbF levels. We found that several *HBS1L-MYB* intergenic variants reduce transcription factor binding, affecting long-range interactions with *MYB* and *MYB* expression levels. These data provide a functional explanation for the genetic association of *HBS1L-MYB* intergenic polymorphisms with human erythroid traits and HbF levels. Our results further designate *MYB* as a target for therapeutic induction of HbF to ameliorate sickle cell and β-thalassemia disease severity.

Introduction

Approximately half of our blood volume is made up of erythrocytes, providing the oxygen and carbon dioxide transport necessary for cellular respiration throughout the body. Erythroid parameters (e.g., red blood cell count [RBC], mean cell volume [MCV], and mean cell hemoglobin [MCH] content) are routinely used for the diagnosis and monitoring of a wide range of disorders as well as overall human health. Significant variation in these parameters, which is highly heritable, occurs among humans (1,2). Genome-wide association studies (GWAS) and other studies have investigated the genetic basis of variation in erythroid and other hematological traits within different ethnic populations. As observed in the majority of association studies, some genome-wide, sequence variants modulating human traits are predominantly located in noncoding regions of the genome (3), complicating the functional interpretation of their effects. A set of common intergenic SNPs at chromosome 6q23 has been consistently identified as highly associated with clinically important human erythroid traits (4–13) (Table1). Prominent among these traits is the persistence of fetal hemoglobin (Hb) in adults (HbF, measured as %HbF of total Hb or as proportion of red blood cells carrying HbF [%F cells] (4,14,15)). General diagnostic erythroid parameters such as RBC, MCV, MCH, and others (5,7,8,10,13) have also been found to be highly associated with the presence of the 6q23 variants. Traits with weaker, but significant association are packed blood cell volume (PCV, also referred to as hematocrit) (7,10,13), total Hb (13), HbA2 (12), and even nonerythroid traits (i.e., monocyte and platelet counts) (5,10). The genetic regulation of HbF levels is of particular therapeutic interest, as increased HbF levels significantly ameliorate disease severity of the 2 main β-hemoglobinopathies - β-thalassemias and sickle cell disease (16,17) - which represent some of the most common human genetic disorders (18). Erythroid-trait associated SNPs (Table 1) reside within a 126-kb intergenic region between the *HBS1L* and *MYB* genes (Figure 1A). As originally reported in studies investigating the genetic basis of variation in HbF levels (4,15), a small number of these SNPs were shown to display an especially strong association; these observations were largely confirmed for the other erythroid phenotypes investigated (7,8,10,13). These SNPs are closely linked with each other and span a region of about 24 kb (originally termed *HBS1L-MYB* intergenic polymorphism block 2 [HMIP-2]) (4,7,11). Association of these HMIP-2 SNPs with the erythroid traits has been replicated and validated in populations from diverse ethnic backgrounds (6–8,10). Despite extensive genetic evidence, a clear mechanistic basis for the association between the intergenic SNPs and erythroid biology has remained elusive, although the 2 flanking genes (*HBS1L* and *MYB*) are candidate target genes (4,19–22).

Table 1

Human erythroid phenotypes associated with *HBS1L-MYB* intergenic variants

Erythroid phenotype	References
Hb	13
MCH	5, 7, 10, 13
MCHC	7, 10, 13
MCV	5, 7, 8, 10, 13
PCV/Hct	7, 10, 13
RBC	5, 7, 8, 10, 13
HbF	4, 6, 9, 11, 14, 15
HbA_2	12

Only variants with $P < 10^{-8}$ (99 in total) were selected for further study. MCH, mean cell hemoglobin; MCHC, Mean cell hemoglobin concentration; Hct, hematocrit.

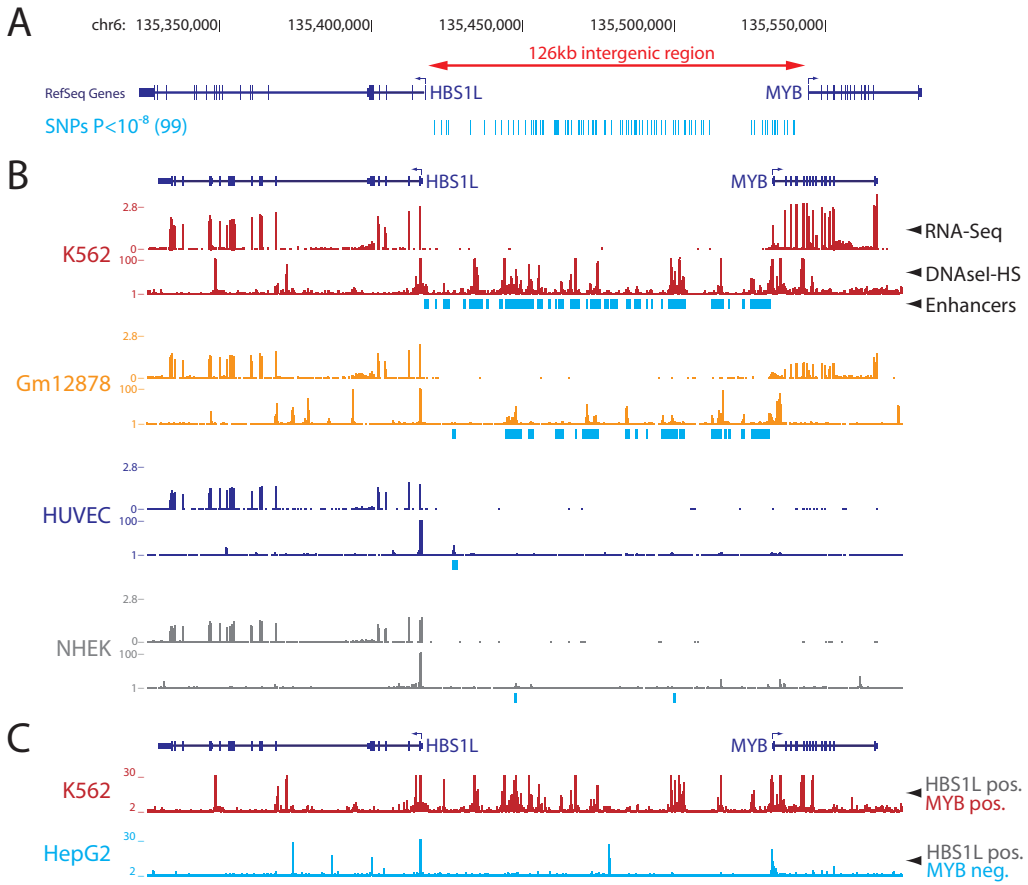


Figure 1. The erythroid/hematopoietic-specific regulatory signature of the *HBS1L-MYB* intergenic region associated with HbF levels and other human erythroid traits. (A) Intergenic SNPs associated ($P < 10^{-8}$) with different erythroid phenotypes (listed in Table 1) as reported by published GWAS (Table 1) are plotted below the *HBS1L-MYB* locus. (B) Locus-wide expression, DNaseI hypersensitivity, and enhancer chromatin signature data for 4 different cell types representing erythrocytes (K562), lymphocytes (Gm12878), endothelial cells (HUVEC) and keratinocytes (NHEK). The y axis represents sequence tag density. (C) Locus-wide digital genomic footprinting data shown for an erythroid cell line (K562) expressing both *MYB* and *HBS1L* (HBS1L pos/MYB pos) and for a liver cell line (HepG2) expressing only *HBS1L* (HBS1L pos/MYB neg). The y axis represents sequence tag density. Genome-wide data sets were obtained from the ENCODE consortium and accessed through the UCSC Genome Browser (http://genome.ucsc.edu/). DNaseI-HS, DNaseI hypersensitivity.

Whereas the function of *HBS1L* in red blood cell development is uncharacterized, the *MYB* gene (encoding the c-MYB transcription factor [TF]) is a key regulator of hematopoiesis and erythropoiesis (23,24). c-MYB plays an essential role in controlling the erythroid cellular proliferation/differentiation balance (25) and regulates HbF levels through an undefined mechanism (19,20). The functional importance of the intergenic region was first observed when transgene insertion within the murine *Hbs1l-Myb* intergenic region almost completely abolished *Myb* transcription and resulted in severe anemia (22). A recently reported follow-up investigation mapped the location of transgene insertion to the HMIP-2 orthologous region and showed elevated levels of embryonic globin genes in splenic erythroid cells of these transgenic mice (21), confirming the importance of the intergenic region for globin gene regulation in the mouse. We previously identified several distal regulatory elements in the mouse *Hbs1l-Myb* intergenic region that regulate *Myb* transcription by physically interacting with the *Myb* promoter and first intron in erythroid progenitors via

chromatin looping (26,27). In humans, microarray-based experiments have demonstrated the presence of erythroid-specific transcription and active histone modifications in this region (28). We therefore set out to characterize the regulatory potential of the human *HBS1L-MYB* intergenic region in detail and to investigate the functional impact of the erythroid phenotype–associated variants.

Results

Regulatory activity at the human HBS1L-MYB intergenic region strictly correlates with MYB expression levels.

Genome-wide data sets generated by the ENCODE consortium (29) were inspected to explore gene expression and intergenic regulatory potential within the *HBS1L-MYB* region for a number of cell lines representing a variety of tissues. This showed that high-level *MYB* expression was restricted to hematopoietic cells (erythroid K562 and lymphoid GM12878 cells), while *HBS1L* was expressed at similar levels in all cell types (Figure 1B and Supplemental Figure 1; supplemental material available online with this article), confirming previous observations (28). Next, intergenic regulatory activity was assessed using a combination of genome-wide histone modification, DNaseI hypersensitivity, and genomic footprinting data sets (30–32). A strong positive correlation between *MYB* expression levels and intergenic regulatory activity emerged (Figure 1, B and C, and Supplemental Figure 1). In erythroid K562 cells, which express the highest levels of *MYB*, the intergenic interval contains numerous enhancer chromatin signatures. Lymphoid

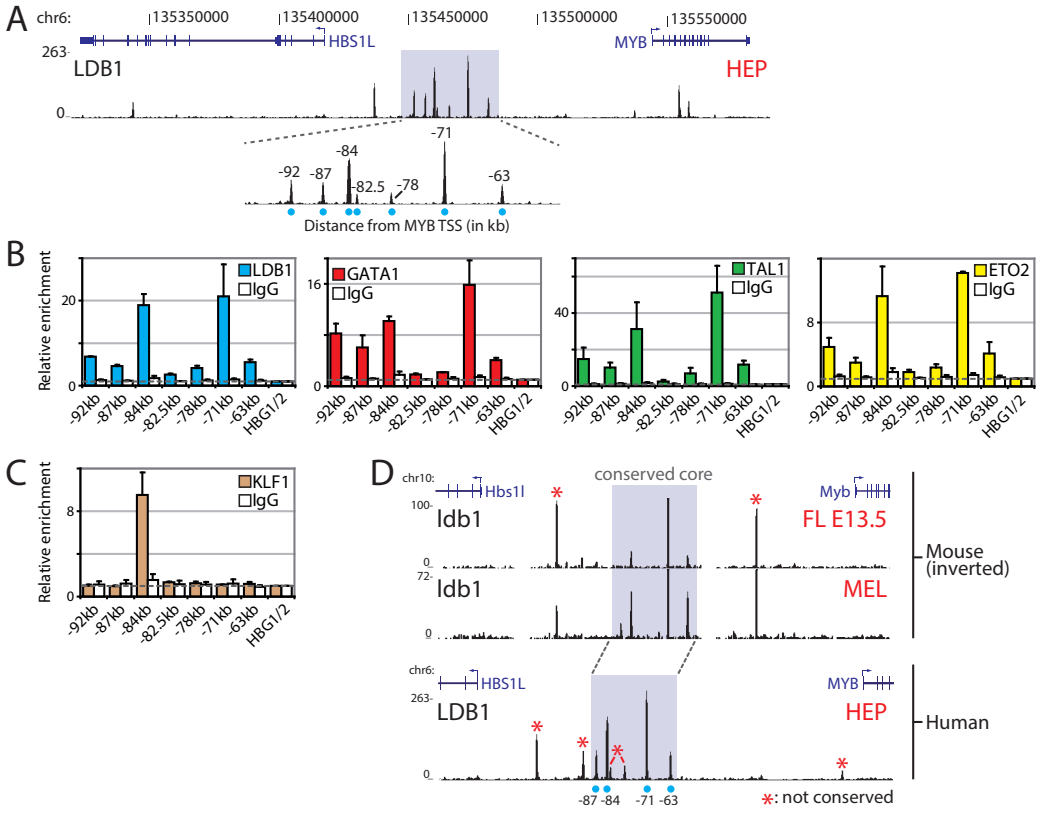


Figure 2. The *HBS1L-MYB* intergenic region contains regulatory elements bound by erythroid TFs. (A) LDB1 ChIP-Seq data from primary HEPs. LDB1 peaks were marked by their distance to the *MYB* TSS. (B and C) ChIP-qPCR data (HEPs) showing enrichment (n = 3) for LDB1 complex members (B) and KLF1 (C) at the intergenic binding sites. IgG serum was used as control (IgG); the *HBG1/2* promoter for normalization. (D) Comparison of mouse and human LDB1 ChIP-Seq data from erythroid progenitors. Binding sites not conserved are marked (*). The region containing the 4 conserved sites (conserved core) is highlighted in purple. Error bars display SEM. FL E13.5, 13.5 dpc fetal liver erythroid progenitors.

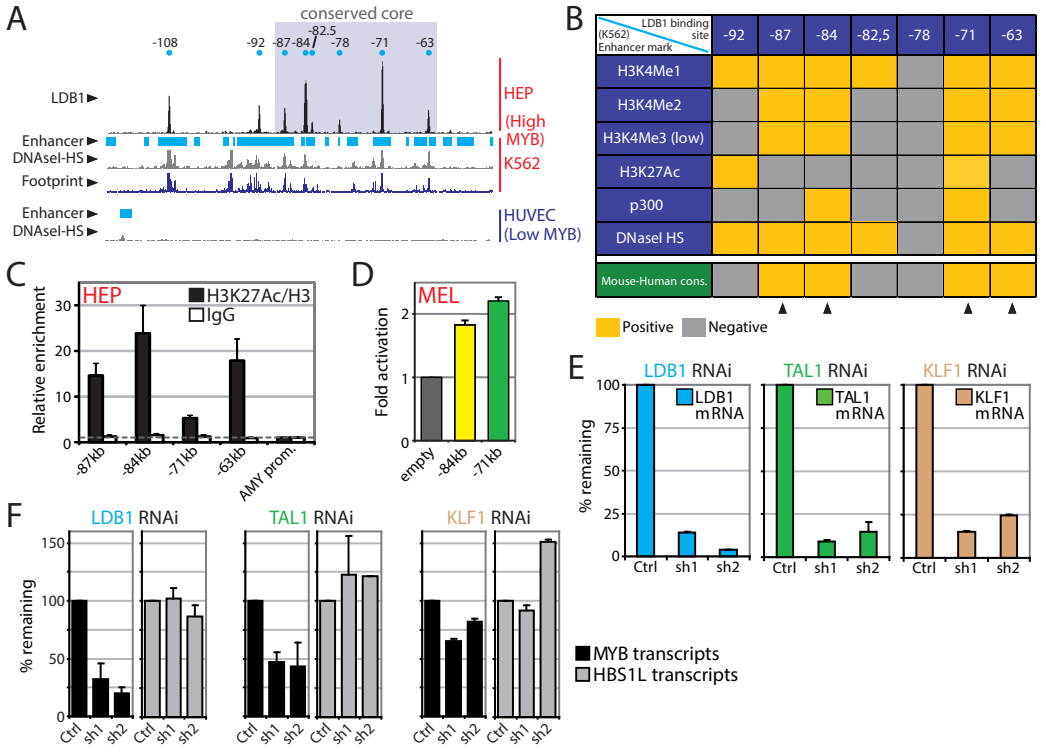


Figure 3. Erythroid TFs bind intergenic enhancer regions and are required for *MYB* expression. (A) Alignment of LDB1-binding sites (HEPs) to enhancer chromatin signature, DNaseI-HS, and footprinting data from erythroid (K562) and endothelial (HUVEC) cell lines. (B) Table summarizing the comparison between LDB1 sites (HEPs) and enhancer marks (K562). Arrowheads denote conserved sites with highly enriched enhancer signatures. (C) H3K27 acetylation as measured by ChIP-qPCR in HEPs for indicated LDB1-binding sites (n = 2). Enrichments were corrected for total H3 levels and normalized to the *AMY2A* promoter (AMY prom.). (D) Luciferase reporter assays in MEL cells measuring (n = 3) enhancer activity of the –84-kb and –71-kb elements. Promoter activity without enhancer (empty) was set to 1. (E and F) Gene expression analysis (n = 3) on K562 cells depleted for the indicated TFs by RNAi. A scrambled shRNA was used as control (Ctrl). Error bars display SEM.

Gm12878 cells expressing lower levels of *MYB* display fewer areas of regulatory activity. Finally, cell types not expressing *MYB* (i.e., HUVEC, NHEK, HepG2) display heterochromatinized or polycomb-repressed intergenic regions with an absence of DNaseI-hypersensitivity while still expressing *HBS1L* at high levels (Figure 1, B and C, and Supplemental Figure 1). These observations suggest that the *HBS1L-MYB* intergenic region is likely to contain *MYB*-specific regulatory elements.

Erythroid TF complexes occupy regulatory sequences in the HBS1L-MYB intergenic region and are required for MYB expression.

To identify regulatory elements controlling *MYB* expression more precisely, we profiled chromatin occupancy of the key erythroid LDB1 TF complex (33) in primary human erythroid progenitors (HEPs) using ChIP coupled to high-throughput sequencing (ChIP-Seq) and quantitative PCR (ChIP-qPCR). We detected an intergenic cluster containing 7 binding sites for the LDB1 complex, characterized by strong binding and co-occupancy of core complex proteins LDB1, GATA1, TAL1, and ETO2 (Figure 2, A and B, marked by their distance from the *MYB* transcriptional start site (TSS). Furthermore, we found one of these sites to be co-occupied by the erythroid-specific TF KLF1 (Figure 2C), a protein that was found to bind the murine

intergenic region (26,34). These TFs are critical regulators of erythroid development (33,35,36), are positive regulators of murine *Myb* expression (26), and have been implicated in establishing long-range promoter-enhancer communication (37–40). The emerging TF-binding profile is reminiscent of the one observed in mouse erythroid cells (26). When LDB1 ChIP-Seq profiles from mouse and HEPs were compared, a core region of 4 highly conserved binding sites emerged, which included the single LDB1/KLF1 co-occupied site 84 kb upstream of the *MYB* TSS (Figure 2D). Interestingly, as previously observed in mouse erythroid cells (26), these 4 conserved core sites (at positions –87, –84, –71 and –63) displayed strong enhancer signatures (41) in K562 and HEPs (Figure 3, A–C, Supplemental Figure 2). Furthermore, several of these putative regulatory elements showed enhancer activity in luciferase reporter assays (Figure 3D). These data suggest that the *HBS1L-MYB* intergenic interval contains enhancer elements bound by erythroid TFs. Depletion of LDB1, TAL1 and KLF1 in K562 cells using RNA interference (RNAi) resulted in a specific downregulation of *MYB* expression while leaving *HBS1L* levels unaffected (Figure 3, E and F), demonstrating that the erythroid TFs occupying the intergenic enhancers are required for *MYB* expression.

Intergenic TF-bound regulatory elements spatially cluster around the MYB gene in primary erythroid cells.

To test whether the intergenic regulatory elements indeed act as long-range enhancers to regulate *MYB* in primary human cells, we analyzed the *in vivo* 3D chromatin structure of the locus using chromosome conformation capture (3C) coupled to high-throughput sequencing (3C-Seq) (42). We also profiled CTCF occupancy within the locus; a protein known to be important for chromatin looping (43,44) that has recently been implicated in regulating *Myb* expression in mouse erythroid cells (26). Several strong chromatin coassociations between the *MYB* promoter and intergenic sequences were detected, almost all of which correlated with TF-binding events (Figure 4A). Importantly, the highest interaction density was observed within the conserved core region, further strengthening the importance of the TF-bound regulatory elements within this region. Performing 3C-Seq using the –84 LDB1 complex/KLF1–binding site as a viewpoint produced a similar pattern of long-range chromatin interactions within the intergenic region and around the *MYB* promoter (Figure 4A). 3C-qPCR analysis on HEP and K562 cells confirmed the nuclear proximity between *MYB*, the –84 regulatory element, and a CTCF site in between (Figure 4B and Supplemental Figure 3), which was not observed in cells expressing very low levels of *MYB* (HeLa, Figure 4, B and C) Thus, *in vivo*, the intergenic regulatory elements cluster in the nuclear space and are involved in long-range interactions with the active *MYB* gene.

Common variants modulating human erythroid traits colocalize with TF-bound intergenic regulatory elements.

Next, we set out to compare the locations of the TF-bound regulatory sequences with those of the SNPs reported to be associated with erythroid phenotypes. This trait-associated variation involves more than 100 SNPs and small deletions spanning the entire interval between *HBS1L* and *MYB* (Figure 5). The locus was first identified as associated with HbF persistence (4). It was subsequently shown (5,7,8,13) that an analogous pattern of association exists with routine diagnostic hematological parameters, especially MCV, MCH, and RBC, but also other erythroid and hematological nonerythroid parameters. A distinct small subset of these variants is set apart by their particularly strong association with these traits and with each other (linkage disequilibrium [LD]) in individuals of European and Asian descent. This LD block of SNPs (termed HMIP-2; ref. 4) is distributed over a physical area of 24 kb (Figure 5). From published GWAS (Table 1), we identified 17 common HMIP-2 variants (15 SNPs and a 3-bp deletion/SNP combination, detailed in Methods) that showed an exceptionally strong genetic association across the erythroid traits. These 17 variants, or a subset of them, are most likely functionally involved in modulating erythroid biology. We subsequently investigated the physical and functional relationship of these candidate variants to key sequences of TF-binding and regulatory activity within the *HBS1L-MYB* intergenic interval.

Strikingly, the sequence area spanned by our candidate variants (analogous to the HMIP-2 block) is largely identical to the conserved core region containing the TF-bound regulatory elements (Figure 5). Of our 17 candidate variants, 5 were located within sequences showing both enhancer signatures and protein-binding features (Figures 2, 3, and 5). Four of these 5 variants are positioned directly under LDB1 complex ChIP-Seq peaks: 2 are located within the –84 LDB1 complex/KLF1–binding site (rs66650371, a 3-bp deletion and rs7775698, a SNP located inside its nondeleted allele) and 2 within the highly enriched –71 LDB1

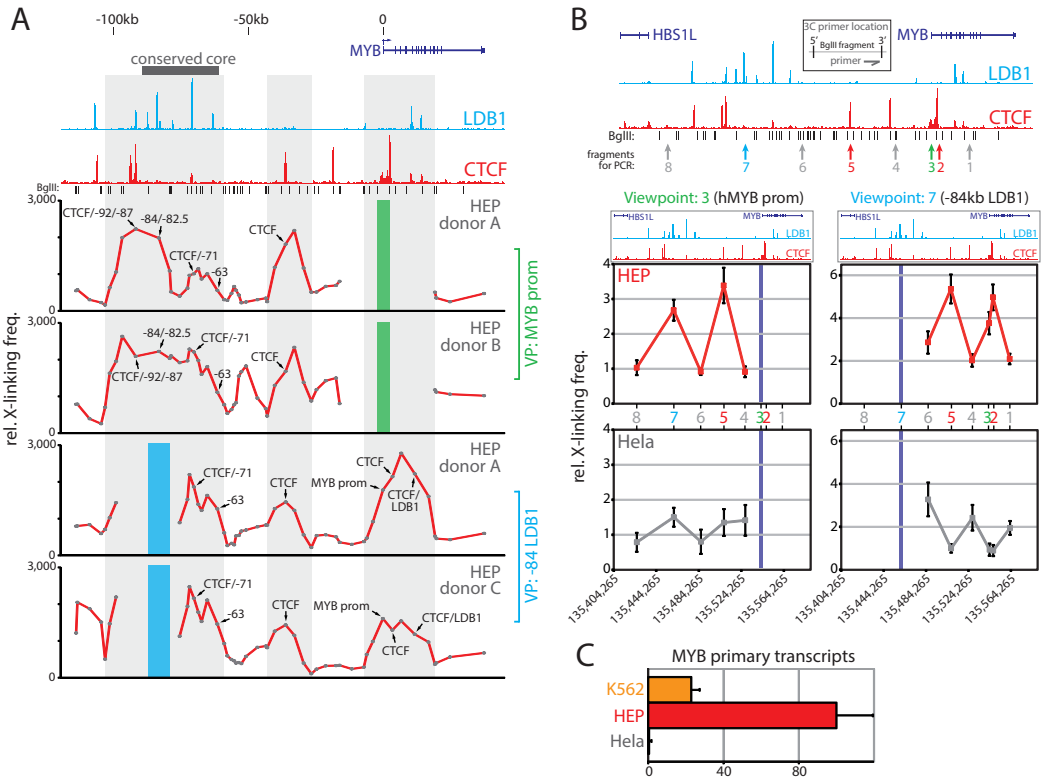


Figure 4. 3C analysis of the *HBS1L-MYB* locus reveals long-range interactions between intergenic elements and the *MYB* gene. (A) 3C-Seq analysis performed on primary HEPs from 3 different donors using the *MYB* promoter (green bar) or the –84 regulatory element (blue bar) as a viewpoint (VP). LDB1 and CTCF ChIP-Seq results from primary HEPs and gene locations are shown at the top. Gray shading highlights regions of coinciding protein binding and chromatin looping. The y axis represents relative crosslinking frequencies per BglII fragment as measured by sequence tag density. (B) 3C-qPCR experiments on primary HEPs (red, n = 5) and HeLa cells (gray, n = 3) using the same viewpoints as in A. The locus is plotted on top, with the different 3C restriction fragments (BglII) used for PCR indicated. A schematic depicting the location of the primers on the chosen restriction fragments is shown. Interaction frequencies between 2 fragments within the ERCC3 locus were used for normalization. (C) Gene expression analysis (n = 3) of *MYB* transcript levels in the different cell types used for the 3C analysis. *ACTB* levels were used for normalization. Error bars display SEM.

complex binding site (SNPs rs6920211 and rs9494142) (Figure 5). Both these conserved TF-binding sites displayed typical active enhancer signatures (Figures 2 and 3) and showed high-interaction frequencies with the *MYB* gene in 3C assays (Figures 4 and 5). The 2 overlapping variants at the center of the –84 LDB1 complex–binding site (rs66650371/rs7775698) are located in the immediate vicinity of a TAL1 and GATA1 motif, as noted before (11). In individuals of European descent, these 2 polymorphisms are in complete LD (4) and therefore the association cannot be distinguished. Observations in individuals of African descent showed that of the 2 variants, the 3-bp deletion is the actual associated one (ref. 11 and discussed below). Additionally, one of the SNPs in the –71 binding site (rs9494142) is located directly adjacent to a GATA1 motif. These observations suggest that the variants falling in these regions may affect long-range *MYB* regulation and through this mechanism exert their influence on human erythroid blood parameters.

In individuals of African descent, the link between the HMIP-2 variants is less rigid, and the block breaks down into 2 independently associated groups of variants (6). Interestingly, the "upstream" group is located at and immediately next to the –84 LDB1 complex–binding site (including rs9399137 and the rs66650371 3-bp deletion, while the overlapping rs7775698 SNP is not associated with erythroid traits). The "downstream" HMIP-2 association signal in African-descended populations was found to be strongest in the

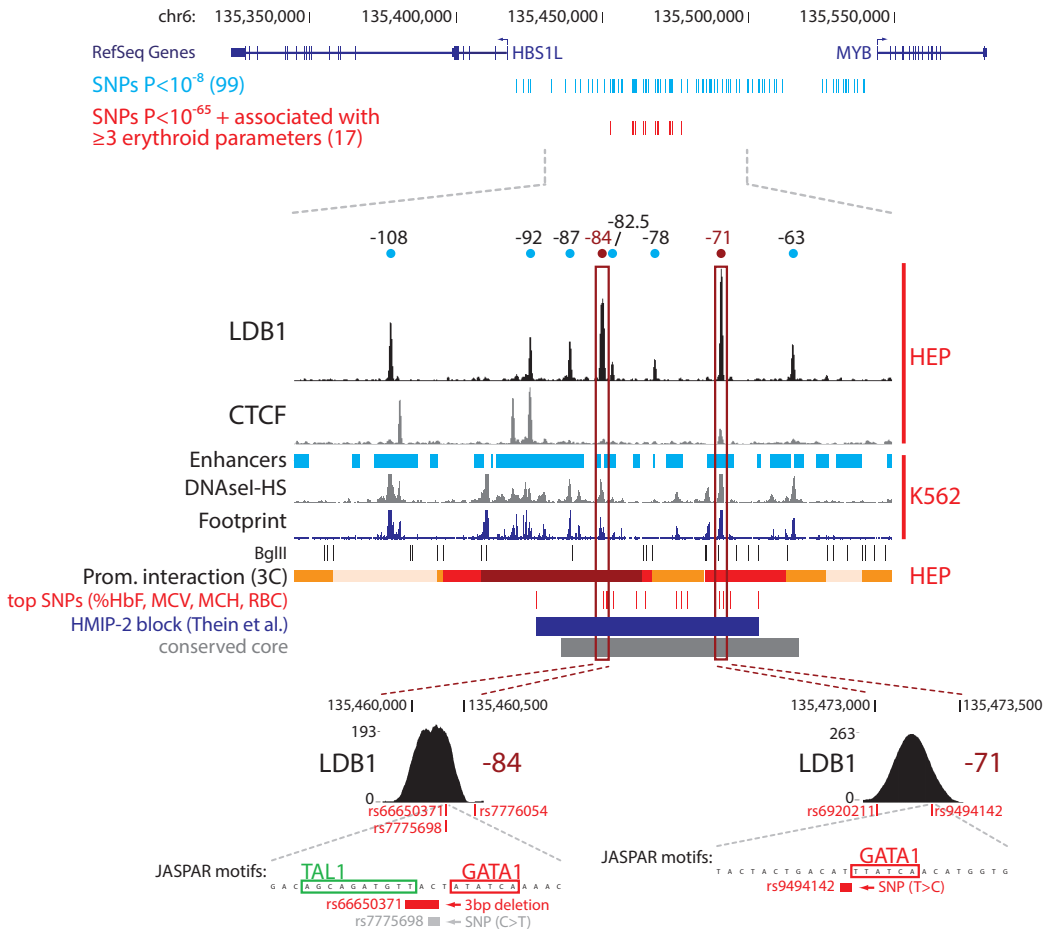


Figure 5. Intergenic polymorphisms associated with HbF and other erythroid parameters localize to the intergenic regulatory elements. All published intergenic SNPs associated with human erythroid traits ($P < 10^{-8}$; blue) and the most highly associated ($P < 10^{-65}$ and 3 or more major erythroid parameters [%HbF, MCV, MCH and RBC]; red) variants are shown directly under *MYB* and *HBS1L* gene locations. Below, a zoom-in picture of the LDB1 binding-site cluster and its regulatory signature is further compared with the location of the conserved core (gray), HMIP-2 block (dark blue) and the 17 highly associated candidate SNPs. Chromatin looping with the *MYB* promoter (Figure 4A) is depicted on a white (no interaction) to red (strong interaction) color gradient. Two additional zoom-in pictures display the locations of the SNPs relative to the TF-binding motifs (identified by JASPAR) within the –84 and –71 sites. Within the –84 element, rs66650371 is the actual associated variant (in red; see Results for details).

middle of the conserved region of regulatory elements (rs4895441 and rs9402686; refs. 6,9), but extended across the region to include the –71 LDB1 complex–binding site with SNPs rs6920211 and rs9494142 (S. Menzel et al., unpublished observations).

rs66650371 affects TF binding, enhancer activity, and promoter-enhancer communication in erythroid cells.

To begin probing the functional impact of one of the most prominent variants, the rs66650371 3-bp deletion (Figure 5 and ref. 11), we designed allele-specific assays (see Methods and Supplemental Figure 4) using K562 cells, which are heterozygous for this –84 variant (Supplemental Figure 4A and ref. 11), but not informative for rs9494142 (data not shown). First, using allele-specific ChIP, we observed diminished (25%–

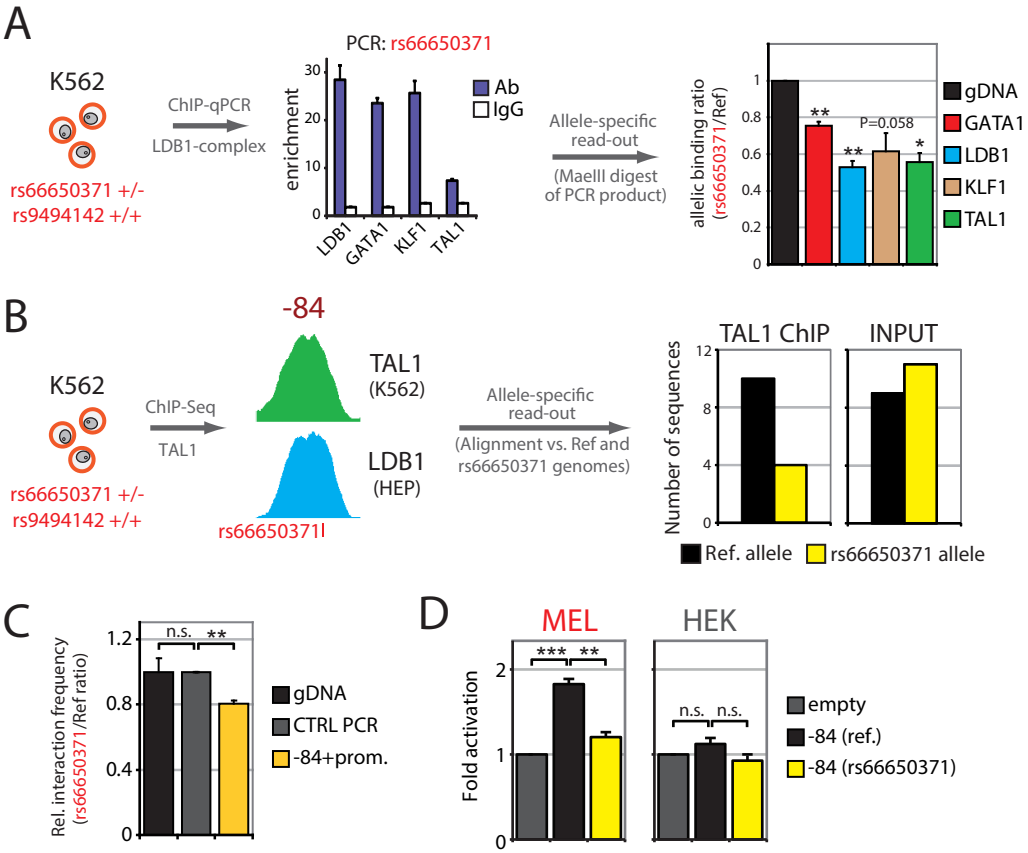


Figure 6. rs66650371 affects protein binding, chromatin looping, and enhancer activity within the erythroid *HBS1L-MYB* locus. (A) Allele-specific ChIP experiments for the rs66650371 alleles in K562 cells heterozygous for this variant. Occupancy of rs66650371 (within the –84 element) by LDB1, GATA1, TAL1, and KLF1 was measured by ChIP-qPCR (n = 2, normalized against AMY2A promoter values), followed by an allele-specific read-out using MaeIII digestion (n = 2, see Methods and Supplemental Figure 4). Allelic abundance was expressed as a rs66650371 (minor)/reference (major) ratio, which was set to 1 for genomic DNA (gDNA). A ratio of less than 1 is the result of a relative lower abundance of the rs66650371 minor allele in the ChIP samples. (B) TAL1 ChIP-Seq was performed in K562 cells, and sequence reads were mapped against the reference and rs66650371 (containing the minor 3-bp deletion allele) genomes. K562 input genomic DNA was PCR amplified (amplicon spanning rs66650371) and cloned into a plasmid; colonies were sequenced (n = 20). (C) Allele-specific quantification (n = 3) of chromatin looping between the –84 element and the *MYB* promoter in K562 cells. A long-range PCR approach was combined with an MaeIII digestion-based read-out for quantification (see Methods). (D) Luciferase reporter assays measuring enhancer activity of the reference (ref.) and rs66650371 minor –84 enhancer alleles in erythroid (MEL) and nonerythroid (HEK) cells. Error bars display SEM. Statistical significance was determined using Student's t test. *P < 0.05; **P < 0.01; ***P < 0.001.

50%) binding of LDB1, GATA1, TAL1, and KLF1 to the rs66650371 allele carrying the deletion (as compared with the nondeleted reference allele, Figure 6A), showing that rs66650371 affects local TF binding. Allele-specific mapping of K562 TAL1 ChIP-Seq reads further confirmed the detrimental effect of this 3-bp deletion on TF binding (Figure 6B). Second, using an allele-specific 3C analysis (see Methods and Supplemental Figure 4), we showed reduced interactions between the rs66650371-deleted –84 allele and *MYB* compared with the nondeleted –84 allele (Figure 6C). Finally, we measured the impact of the rs66650371 deletion on –84 enhancer activity in erythroid (murine erythroleukemia [MEL]) and nonerythroid (human embryonic kidney [HEK]) cells using luciferase reporter assays. In MEL cells, a significant reduction in promoter activation

was observed when the rs66650371 minor allele was present in the –84 enhancer element (Figure 6D). In contrast, the –84 region did not show any enhancer activity in HEK cells, regardless of which rs66650371 allele was present (Figure 6D). Together, these results suggest that the minor allele of a highly associated intergenic variant negatively affects enhancer function and *MYB* regulation in erythroid cells.

Trait-associated intergenic variants affect TF binding, chromatin looping, and MYB expression in humans.

To validate and further expand our observations made in erythroid cell lines, we obtained primary erythroid cells from high HbF individuals homozygous for all minor alleles of the phenotype-associated HMIP-2 block (SNP/SNP, containing the –84-kb and –71-kb intergenic variants in the conserved core), and normal HbF individuals homozygous for the absence of the phenotype-associated HMIP-2 variants (WT/WT). Cells cultured *ex vivo* from SNP/SNP individuals showed consistently lower *MYB* levels throughout phase II of the culture as compared with WT/WT control cells (37% lower *MYB* on average; Figure 7A). To further strengthen the observed negative correlation between the presence of the enhancer variants and *MYB* expression, we measured *MYB* expression in HEPs from a larger cohort of healthy individuals with different genotypes (4 SNP/SNP, 9 WT/SNP, and 8 WT/WT; Figure 7B). Linear regression analysis revealed a highly significant correlation between the presence of the variants and reduced *MYB* levels ($P = 0.005$, allelic effect size = –0.113). Moreover, we observed accelerated differentiation kinetics in late-stage SNP/SNP cultures as well as an increased percentage of CD14+ monocytes (Supplemental Figure 5). This is in agreement with the phenotype observed in HEPs depleted for *MYB* by RNAi (20). ChIP experiments carried out in primary erythroid progenitors harvested at day 7 showed reduced binding of GATA1 and KLF1 at the –84 and –71 regulatory elements (containing the associated variants) in SNP/SNP compared with WT/WT individuals (Figure 7C). Similar results were obtained using erythroid progenitors harvested at later stages of differentiation (i.e., day 11; Supplemental Figure 6 and data not shown). Because of the reduced cell numbers (data not shown), reduced intergenic TF enrichments (Supplemental Figure 6A), and accelerated differentiation of SNP/SNP cultures (Supplemental Figure 5), we decided to perform further experiments on cells harvested at day 7. Allele-specific ChIP assays using SNaPshot analysis (45) showed reduced GATA1 binding to the minor rs9494142 C allele in erythroid cells cultured from healthy heterozygous donors (SNP/WT; Figure 7D), confirming the ChIP results on erythroid chromatin from HEPs of SNP/SNP and WT/WT individuals. 3C-qPCR assays on cultured SNP/SNP and WT/WT cells demonstrated diminished looping between the –84 element and the *MYB* promoter in SNP/SNP individuals (Figure 7E). Finally, we determined whether the allele-specific effects observed at the regulatory elements resulted in an allelic imbalance of *MYB* transcripts. HEPs from several healthy unrelated individuals heterozygous for the –84 and –71 intergenic variants were used as test samples, while cells from individuals homozygous (WT/WT and SNP/SNP) for the variants were used as controls. We utilized the intronic rs210796 *MYB* variant (heterozygous in all test and control individuals) to assess allele-specific *MYB* expression levels. Transcript levels in HEPs heterozygous for the phenotype-associated variants indeed showed an allelic imbalance that was not observed in homozygous control cells, which showed a 1:1 allelic ratio (Figure 7F). A correlation between allelic expression imbalance and the presence of the intergenic variants was not detected for *HBS1L* (data not shown), further confirming the specific effect of the intergenic variants on *MYB* regulation. Taken together, these data show that *HBS1L-MYB* intergenic variants affect *MYB* expression by reducing TF binding to its regulatory elements and disrupting long-range enhancer gene communication.

Discussion

Features of red blood cells, such as their number, size, and Hb content, are subtly different among healthy human individuals. Mapping of the underlying genetic variability has identified candidate genes and loci affecting iron metabolism, cytoskeleton function, globin regulation, and other critical processes controlling erythropoiesis and red cell function. However, direct mechanistic interpretation of the effects of the identified variants is often obscured by their nongenic localization, implying that the majority of the associated genetic variation affects noncoding regulatory sequences (3). Detailed investigations of these loci are thus required to fully understand the genetic basis of human trait variation and disease risk (46), as exemplified by the in-depth studies of GWAS-identified SNPs at the *MYC* (47,48) and *OCA2* (49) loci. Several GWAS (4–13) have identified a cluster of common variants in the interval between *HBS1L* and *MYB* that modulate a broad spectrum of hematological traits, in particular erythroid phenotypes, suggesting that this

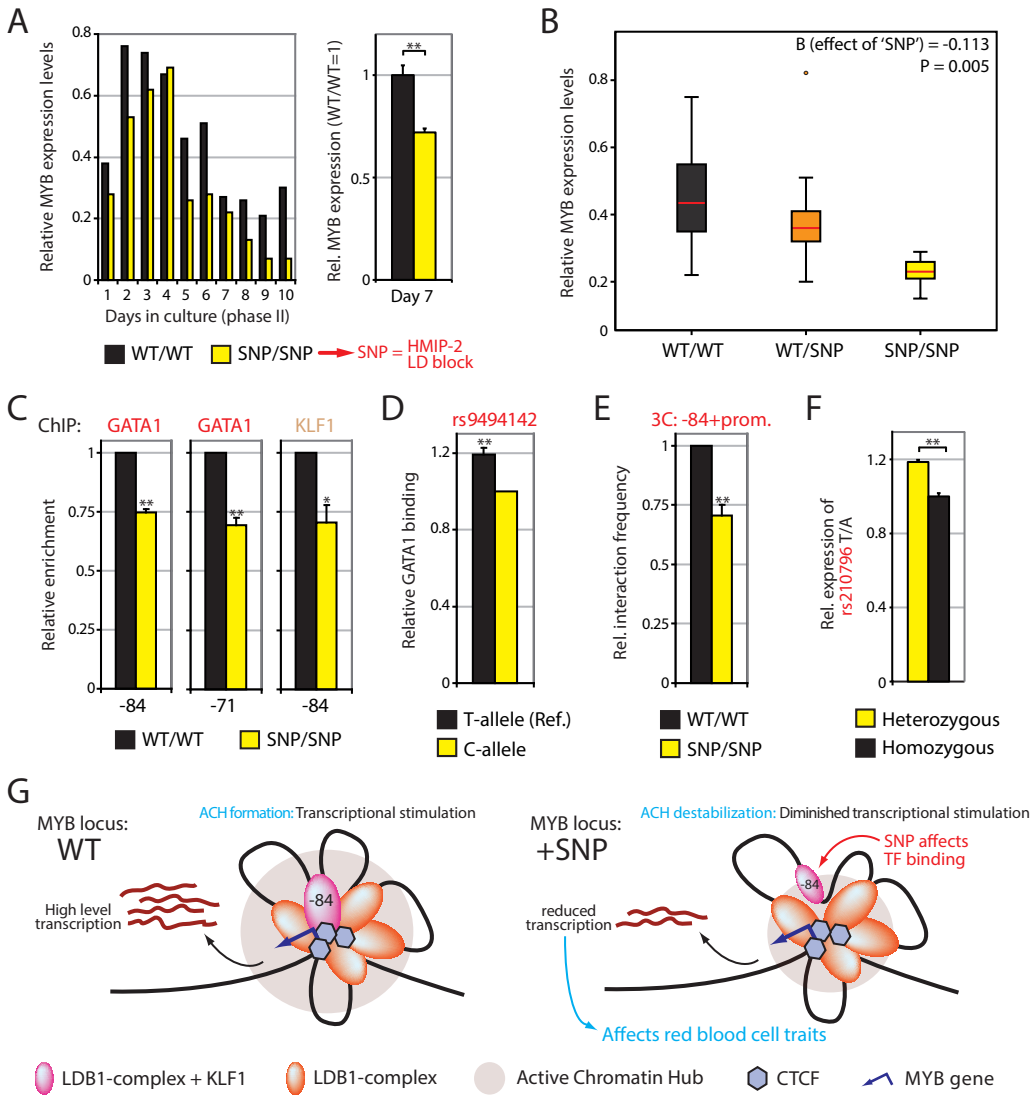


Figure 7. Intergenic variants affect TF binding, chromatin looping, and *MYB* expression in primary HEPs. (A) HEPs from individuals homozygous for the minor allele of the phenotype-associated variants (HMIP-2 LD block variants; SNP/SNP) and WT control individuals (WT/WT) were cultured and assayed for *MYB* expression at indicated days (left: representative experiment, right: n = 4). (B) Correlation between intergenic genotype and *MYB* expression was determined using HEPs from 21 individuals (WT/WT, WT/SNP, and SNP/SNP intergenic genotypes; see Methods). Circle represents single data point considered to be an outlier. (C) ChIP-qPCR (n = 3) for GATA1/KLF1 using SNP/SNP and WT/WT HEPs. Enrichments were normalized to IgG and α-globin HS40 values (WT/WT set to 1). (D) Allele-specific measurement of GATA1 binding to rs9494142 (T/C) alleles using SNaPshot on heterozygous individuals (n = 4). rs9494142 C is the phenotype-associated minor allele. (C-allele set to 1). (E) Interaction frequencies between the –84 element and *MYB* promoter were measured (n = 5) using 3C-qPCR in SNP/SNP and WT/WT HEPs. (F) Allele-specific expression measured by SNaPshot in HEPs from individuals heterozygous (n = 5) or homozygous (n = 5) for the intergenic SNPs; rs210796 SNP (T/A) was used for quantification. (G) Proposed model explaining the effect of trait-associated intergenic SNPs on *MYB* regulation. Transcription factor-bound regulatory elements cluster around *MYB* to form an ACH, stimulating transcription (left). Intergenic SNPs reduce TF binding and chromatin looping, partially destabilizing the ACH and reducing *MYB* transcription (right). Lower *MYB* levels subsequently affect red cell traits. Error bars display SEM. Statistical significance was determined using linear regression analysis or Student's t test. *P < 0.05; **P < 0.01.

locus may have a key role in the regulation of erythropoiesis. However, molecular insight into how these intergenic polymorphisms could affect erythroid parameters remains elusive.

Here we have characterized the regulatory potential of the *HBS1L-MYB* intergenic region in detail and identified a cluster of erythroid-specific enhancers controlling the expression of *MYB* (Figures 1–4), a critical regulator of erythropoiesis (23,25). Common variants affecting human erythroid traits were found to cluster close to or within the enhancers (Figure 5), where they disrupt enhancer activity through the attenuation of TF binding and enhancer-promoter looping, resulting in reduced *MYB* expression levels (Figures 6 and 7). These experiments provide what we believe is the first causal link among the intergenic variants, *MYB* regulation, and their influence on erythroid traits.

Regulatory control of the *MYB* gene in erythroid cells has thus far remained incompletely defined, although it involves regulation via its proximal promoter region (50,51) and microRNAs (20,52,53). Our experiments show that *MYB* is additionally controlled distally by enhancer elements more than 80 kb upstream of its promoter, illustrating the high degree of regulatory complexity that governs *MYB* expression. It has been postulated that enhancers cluster in the nuclear space to form active chromatin hubs (ACH) to stimulate target-gene transcription (54), a process likely to involve the concerted action of TFs. Our observations of *in vivo* clustering of enhancers around *MYB* suggest the presence of a *MYB* ACH (Figure 7G), similar to that observed in murine erythroid progenitors (26). Intergenic polymorphisms, through their detrimental effect on TF recruitment to the enhancers, could partially destabilize the *MYB* ACH, in turn resulting in decreased transcriptional output and a subsequent modulation of erythroid traits (Figure 7G).

The most significantly associated variants (Figure 5) cluster within a discrete 24-kb region that appears to function as an erythroid-specific long-range *MYB* enhancer. In this core regulatory region, 5 of the polymorphisms are located within 2 regulatory elements 84 and 71 kb upstream of the *MYB* TSS. There they alter nucleotides adjacent to or within E-Box/GATA TF binding motifs used to recruit the LDB1 complex (35,55) and affect the spacing between these motifs. Spacing between TF-binding motifs within enhancer sequences has been reported to be a constraint for optimal binding (56,57), and reduction of E-box/GATA motif spacing by the rs66650371 3-bp deletion in the –84 element could underlie the diminished TF binding and enhancer activity of the deleted rs66650371 allele (Figures 6 and 7). In addition, sequences flanking core binding motifs are known to be important for optimal TF binding (57). For example, a stretch of A or T residues adjacent to core motifs was observed as a specificity determinant (57). The rs9494142 minor allele (C) disrupts a stretch of 3 A/T residues adjacent to the "TATC" core GATA1 motif, providing a possible explanation for the reduced observed GATA1 binding to the rs9494142 minor allele (Figure 7, C–D). Alternatively, the variants might affect TF binding indirectly, for example, through local changes in chromatin structure (58) or by creating a new TF-binding site that might affect LDB1 complex–binding through competition (59). Even though the individual effects of the enhancer variants on TF binding and chromatin looping were modest, it is likely that several of the most significantly associated enhancer variants (which are in strong LD, i.e., rs66650371 and rs9494142) act in concert to cause the observed significant reduction in *MYB* expression levels (Figure 7B). Indeed, a recent study (60) showed that such an "additive enhancer variant mechanism" takes place at several other loci identified in GWAS.

Exactly how c-MYB controls HbF levels and the many other erythroid traits is not yet fully understood. A clear anti-correlation between *MYB* and HbF levels has emerged (19, 20), which was further confirmed by the reduced *MYB* expression we observed in erythroid cells from high HbF individuals (Figure 7A). Studies investigating the effects of lower *MYB* levels in mouse and human erythroid cells reported that cell-cycle progression was slower and accelerated differentiation kinetics were observed in later stages of erythroid development (19,20,61). In accordance with these results, ChIP-Seq experiments (29) detected c-Myb binding to key cell-cycle regulators (i.e., *Bcl2*, *Cdk6*, *Myc*) in murine erythroid progenitors (Supplemental Figure 7B). Furthermore, several of these genes were found to be misregulated in an analysis of published *MYB* loss-of-function studies (20, 21, 61) in HEPs (Supplemental Figure 7C). Accelerated differentiation in an environment of lower *MYB* levels could favor premature cell-cycle termination during the proliferation cycles of adult erythropoiesis, producing more erythroid cells that synthesize predominantly HbF ("F-cells") before the switch to adult Hb synthesis occurs (Supplemental Figure 7D) (62). In this context, lower *MYB* levels will lead to lower RBC (resulting from the reduced number of proliferation cycles) and higher MCV as the erythrocytes are younger red cells (Supplemental Figure 7D) (5,63); indeed, these traits are genetically associated with the minor alleles of the intergenic SNPs (5).

Alternatively, recent studies suggest that the c-MYB TF plays an important role in the emerging TF

network governing γ-globin expression, in which the BCL11A and KLF1 proteins play key repressive roles (17, 64–67). Remarkably, we noticed that c-Myb in murine erythroid progenitors occupied the β-globin locus and many of the established γ-globin repressor genes, including *Bcl11a* and *Klf1* (Supplemental Figure 7A). Analysis of previous c-MYB loss-of-function studies (20,21,61) indeed showed that several of the c-Myb–bound γ-globin repressor genes (i.e., *Bcl11a*, *Klf1*) are downregulated upon *MYB* depletion (Supplemental Figure 7C). These observations suggest that c-MYB directly activates key γ-globin repressor genes and thus fulfils an important role within the established molecular HbF repression mechanisms (Supplemental Figure 7D).

Direct targeting of TFs that regulate γ-globin expression to induce HbF production in adults has remained challenging, as conventional TFs not highly signal dependent (such as BCL11A or c-MYB) have been very difficult drug targets (68). However, the ongoing revolution in genome engineering methods (e.g., custom-made zinc-finger or TALE-mediated targeting; ref. 69) has made it possible to specifically target genomic sites of interest. Two recent studies (70,71) have provided examples of how to exploit genome-editing technology to modulate gene expression by interfering with enhancer function. Such strategies could also be applied to the erythroid-specific *MYB* enhancers described in our current work. Targeted repression (or perhaps even deletion) of the –84 and/or –71 *MYB* enhancers could reduce *MYB* expression specifically in erythrocytes (analogous to the effect of the high-HbF–associated variants), resulting in elevated HbF levels. Although *MYB* is essential for proper erythroid development, moderately reduced *MYB* levels seem to be well tolerated by the erythroid system (19,24,72).

As elevated HbF levels ameliorate the severity of β-thalassemia and sickle cell anemia, induction of HbF in adults has been a major focus of research in the past decades (16,17). Our work provides the essential mechanistic basis and enhancer characterization that are necessary for the potential future development of therapeutic strategies aimed at inducing HbF by reducing *MYB* levels via its intergenic regulatory elements.

Methods

Subjects and analyses of blood samples.

A total of 50 healthy unrelated adults of diverse ethnic backgrounds were recruited as well as selected members of the Asian-Indian kindred (73). HbF levels were measured using high-performance liquid chromatography (BioRad Variant; BioRad) and F cells as previously described (19), using blood in EDTA. Genomic DNA was isolated from peripheral blood of these individuals and genotyped for the relevant *HBS1L-MYB* intergenic variants and for the intronic rs210796 SNP on chromosome 6q23. Individuals with the appropriate intergenic and *MYB* intron 4 genotypes were selected for culture studies. HEPs were cultured from 21 individuals with certain combinations of the trait-associated *HBS1L-MYB* intergenic variants (8 homozygous for the reference alleles WT/WT, 9 heterozygous WT/SNP, and 4 homozygous for the minor alleles SNP/SNP), as appropriate for allele-specific ChIP and allele-specific expression studies.

Cell culture.

HEPs were cultured from buffy coats or whole blood in EDTA (as previously described; refs. 19,74) using a 2-phase culture system. K562, MEL, HEK, and HeLa cells were maintained in DMEM supplemented with 10% fetal calf serum and penicillin/streptomycin. Cells were counted with an electronic cell counter (CASY-1; Schärfe System).

ENCODE and expression microarray data mining.

A detailed description of ENCODE project and expression microarray data mining can be found in the Supplemental Methods.

Intergenic SNP selection and TF motif prediction.

HBS1L-MYB intergenic common DNA variants associated with erythroid traits (Figure 1) were identified from published data (Table 1). A more detailed analysis of SNP selection and TF motif prediction can be found in the Supplemental Methods.

ChIP and ChIP-Seq.

ChIP experiments were carried out according to procedures described before (35). Antibodies used have been described before (26). KLF1 antibody was provided by Sjaak Philipsen (Department of Cell Biology, Erasmus Medical Centre). High-throughput sequencing of ChIP DNA libraries was performed on the Illumina GAII or HiSeq2000 platforms and analyzed using the NARWHAL (75) pipeline. Data were visualized using a local mirror of the UCSC genome browser (hg18).

Luciferase reporter assays, RNAi, and gene-expression analysis.

Details on reporter assays, RNAi, and expression analysis can be found in the Supplemental Methods. Primer sequences can be found in Supplemental Table 1.

3C and 3C-Seq.

3C and 3C-Seq experiments were essentially carried out as described (26,42). For all experiments, BglII was used as the primary restriction enzyme. See Supplemental Methods for additional information on data normalization. For 3C-Seq library preparations, we used NlaIII as a secondary restriction enzyme. Initial 3C-Seq data processing was performed as described elsewhere (42). Detailed analysis and visualization was carried out using r3Cseq software (76). Primer sequences can be found in Supplemental Table 1.

Allele-specific ChIP, ChIP-Seq, 3C, and SNaPshot analysis.

Allele-specific ChIP, ChIP-Seq, 3C, and SNaPshot strategies are further described in the Supplemental Methods. Primer sequences can be found in Supplemental Table 1.

FACS analysis.

FACS analysis was performed as previously described (19).

Accession codes.

ChIP-Seq and 3C-Seq data sets were deposited in the GEO repository (GSE52637).

Statistics.

Statistical significance was determined using an unpaired, 2-tailed Student's t test unless stated otherwise. Linear regression analysis was performed using SPSS Statistics software (IBM), including an ANOVA test for statistical significance. *MYB* expression measurements used for regression analysis were performed in 2 batches, and the systematic differences between them were corrected for in the statistical analysis (and in Figure 7B). $P < 0.05$ was considered significant.

Study approval.

Investigations using human blood samples conformed to the principles outlined in the Helsinki Declaration of the World Medical Association. Written informed consent was received from participants prior to inclusion in the study. This study was approved by the Local Ethical Committee (LREC no 10/H0808/035) of King's College Hospital, London.

Supplementary Material

Supplementary Material and Methods, as well as Supplementary Figures 1-7 and Supplementary Table 1, are available at the JCI website.

Acknowledgments

We thank members of the Grosveld, Thein, and Soler labs for helpful discussions and Biomics department personnel for excellent technical assistance. We are grateful to Sjaak Philipsen (Department of Cell Biology, Erasmus Medical Centre) for providing the KLF1 antibody. We are also grateful to Nicholas J. Bray and Matthew Hill (Department of Neurosciences, Institute of Psychiatry, King's College London) for help and advice with snapshot, and to Emil Van den Akker (Department of Hematopoiesis, Sanquin Research, Amsterdam, The Netherlands) and Jackie Sloan-Stanley (Weatherall Institute of Molecular Medicine) for help and advice with the erythroid cultures. This work was supported by the EU-FP7 EuTRACC consortium (to F. Grosveld), the Royal Netherlands Academy of Arts and Sciences (KNAW) (to R. Stadhouders), the Dutch Cancer Genomics Center (to E. Soler and F. Grosveld), the Netherlands Genomics Initiative (to E. Soler and F. Grosveld), the Norwegian Research Council (to B. Lenhard), the Bergen Research Foundation (to B. Lenhard), the French Alternative Energies and Atomic Energy Commission (CEA) and the Atip-Avenir Program (to E. Soler), and the Medical Council Research, United Kingdom (to S.L. Thein).

References

1. Evans DM, Frazer IH, Martin NG. Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. 1999;2(4):250–257. doi: 10.1375/twin.2.4.250.
2. Garner C, et al. Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood. 2000;95(1):342–346.
3. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190–1195. doi: 10.1126/science.1222794.
4. Thein SL, et al. Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. Proc Natl Acad Sci U S A. 2007;104(27):11346–11351. doi: 10.1073/pnas.0611393104.
5. Menzel S, et al. The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. Blood. 2007;110(10):3624–3626. doi: 10.1182/blood-2007-05-093419.
6. Lettre G, et al. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. Proc Natl Acad Sci U S A. 2008;105(33):11869–11874. doi: 10.1073/pnas.0804799105.
7. Ganesh SK, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet. 2009;41(11):1191–1198. doi: 10.1038/ng.466.
8. Soranzo N, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet. 2009;41(11):1182–1190. doi: 10.1038/ng.467.
9. Galarneau G, et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet. 2010;42(12):1049–1051. doi: 10.1038/ng.707.
10. Kamatani Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet. 2010;42(3):210–215. doi: 10.1038/ng.531.
11. Farrell JJ, et al. A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. Blood. 2011;117(18):4935–4945. doi: 10.1182/blood-2010-11-317081.
12. Menzel S, et al. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. Br J Haematol. 2013;160(1):101–105. doi: 10.1111/bjh.12084.
13. van der Harst P, et al. Seventy-five genetic loci influencing the human red blood cell. Nature. 2012;492(7429):369–375. doi: 10.1038/nature11677.
14. Menzel S, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nat Genet. 2007;39(10):1197–1199. doi: 10.1038/ng2108.
15. Uda M, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc Natl Acad Sci U S A. 2008;105(5):1620–1625. doi: 10.1073/pnas.0711566105.
16. Thein SL, et al. Control of fetal hemoglobin: new insights emerging from genomics and clinical implications. Hum Mol Genet. 2009;18(R2):R216–223. doi: 10.1093/hmg/ddp401.
17. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. Cold Spring Harb Perspect Med. 2013;3(1):a011643.
18. Williams TN, Weatherall DJ. World distribution, population genetics, and health burden of the hemoglobinopathies. Cold Spring Harb Perspect Med. 2012;2(9):a011692.
19. Jiang J, et al. cMYB is involved in the regulation of fetal hemoglobin production in adults. Blood. 2006;108(3):1077–1083. doi: 10.1182/blood-2006-01-008912.
20. Sankaran VG, et al. MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13. Proc Natl Acad Sci U S A. 2011;108(4):1519–1524. doi: 10.1073/pnas.1018384108.
21. Suzuki M, et al. Disruption of the Hbs1l-Myb locus causes hereditary persistence of fetal hemoglobin in a mouse model. Mol Cell Biol. 2013;33(8):1687–1695. doi: 10.1128/MCB.01617-12.
22. Mukai HY, et al. Transgene insertion in proximity to the c-myb gene disrupts erythroid-megakaryocytic lineage bifurcation. Mol Cell Biol. 2006;26(21):7953–7965. doi: 10.1128/MCB.00718-06.
23. Ramsay RG, Gonda TJ. MYB function in normal and cancer cells. Nat Rev Cancer. 2008;8(7):523–534. doi: 10.1038/nrc2439.
24. Mucenski ML, et al. A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. Cell. 1991;65(4):677–689. doi: 10.1016/0092-8674(91)90099-K.
25. Vegiopoulos A, et al. Coordination of erythropoiesis by the transcription factor c-Myb. Blood. 2006;107(12):4703–4710. doi: 10.1182/blood-2005-07-2968.
26. Stadhouders R, et al. Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. EMBO J. 2012;31(4):986–999. doi: 10.1038/emboj.2011.450.
27. Stadhouders R, et al. Transcription regulation by distal enhancers: who's in the loop? Transcription. 2012;3(4):181–186. doi: 10.4161/trns.20720.

28. Wahlberg K, et al. The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. Blood. 2009;114(6):1254–1262. doi: 10.1182/blood-2009-03-210146.
29. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. doi: 10.1038/nature11247.
30. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–49. doi: 10.1038/nature09906.
31. Thurman RE, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75–82. doi: 10.1038/nature11232.
32. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489(7414):83–90. doi: 10.1038/nature11212.
33. Cantor AB, Orkin SH. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. Oncogene. 2002;21(21):3368–3376. doi: 10.1038/sj.onc.1205326.
34. Tallack MR, et al. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res. 2010;20(8):1052–1063. doi: 10.1101/gr.106575.110.
35. Soler E, et al. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev. 2010;24(3):277–289. doi: 10.1101/gad.551810.
36. Tsiftsoglou AS, Vizirianakis IS, Strouboulis J. Erythropoiesis: model systems, molecular regulators, and developmental programs. IUBMB Life. 2009;61(8):800–830. doi: 10.1002/iub.226.
37. Song SH, Hou C, Dean A. A positive role for NLI/Ldb1 in long-range beta-globin locus control region function. Mol Cell. 2007;28(5):810–822. doi: 10.1016/j.molcel.2007.09.025.
38. Drissen R, et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes Dev. 2004;18(20):2485–2490. doi: 10.1101/gad.317004.
39. Vakoc CR, et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol Cell. 2005;17(3):453–462. doi: 10.1016/j.molcel.2004.12.028.
40. Deng W, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012;149(6):1233–1244. doi: 10.1016/j.cell.2012.03.051.
41. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144(3):327–339. doi: 10.1016/j.cell.2011.01.024.
42. Stadhouders R, et al. Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc. 2013;8(3):509–524. doi: 10.1038/nprot.2013.018.
43. Splinter E, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 2006;20(17):2349–2354. doi: 10.1101/gad.399506.
44. Ribeiro de Almeida C, et al. The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. Immunity. 2011;35(4):501–513. doi: 10.1016/j.immuni.2011.07.014.
45. Norton N, et al. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. Hum Genet. 2002;110(5):471–478. doi: 10.1007/s00439-002-0706-6.
46. Freedman ML, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43(6):513–518. doi: 10.1038/ng.840.
47. Tuupanen S, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009;41(8):885–890. doi: 10.1038/ng.406.
48. Pomerantz MM, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009;41(8):882–884. doi: 10.1038/ng.403.
49. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 2012;22(3):446–455. doi: 10.1101/gr.128652.111.
50. Saleque S, et al. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. Mol Cell. 2007;27(4):562–572. doi: 10.1016/j.molcel.2007.06.039.
51. Sullivan J, et al. Identification of the major positive regulators of c-myb expression in hematopoietic cells of different lineages. J Biol Chem. 1997;272(3):1943–1949. doi: 10.1074/jbc.272.3.1943.
52. Lu J, et al. MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. Dev Cell. 2008;14(6):843–853. doi: 10.1016/j.devcel.2008.03.012.
53. Zhao H, et al. The c-myb proto-oncogene and microRNA-15a comprise an active autoregulatory feedback loop in human hematopoietic cells. Blood. 2009;113(3):505–516. doi: 10.1182/blood-2008-01-136218.
54. de Laat W, Grosveld F. Spatial organization of gene expression: the active chromatin hub. Chromosome Res. 2003;11(5):447–459. doi: 10.1023/A:1024922626726.
55. Fujiwara T, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell. 2009;36(4):667–681. doi: 10.1016/j.molcel.2009.11.001.
56. Spitz F, Furlong EE. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–626. doi: 10.1038/nrg3207.
57. Jolma A, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152(1–2):327–339. doi: 10.1016/j.cell.2012.12.009.
58. Valouev A, et al. Determinants of nucleosome organization in primary human cells. Nature. 2011;474(7352):516–520. doi: 10.1038/nature10002.
59. Lower KM, et al. Analysis of sequence variation underlying tissue-specific transcription factor binding and gene expression. Human Mutat. 2013;34(8):1140–1148. doi: 10.1002/humu.22343.
60. Corradin O, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014;24(1):1–13. doi: 10.1101/gr.164079.113.
61. Bianchi E, et al. c-Myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression. Blood. 2010;116(22):e99–110. doi: 10.1182/blood-2009-08-238311.
62. Stamatoyannopoulos G. Control of globin gene expression during development and erythroid differentiation. Exp Hematol. 2005;33(3):259–271. doi: 10.1016/j.exphem.2004.11.007.
63. Sankaran VG, et al. Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev. 2012;26(18):2075–2087. doi: 10.1101/gad.197020.112.

64. Borg J, et al. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. Nat Genet. 2010;42(9):801–805. doi: 10.1038/ng.630.
65. Sankaran VG, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science. 2008;322(5909):1839–1842. doi: 10.1126/science.1165409.
66. Tallack MR, Perkins AC. Three fingers on the switch: Kruppel-like factor 1 regulation of gamma-globin to beta-globin gene switching. Curr Opin Hematol. 2013;20(3):193–200. doi: 10.1097/MOH.0b013e32835f59ba.
67. Zhou D, et al. KLF1 regulates BCL11A expression and gamma- to beta-globin gene switching. Nat Genet. 2010;42(9):742–744. doi: 10.1038/ng.637.
68. Koehler AN. A complex task? Direct modulation of transcription factors with small molecules. Curr Opin Chem Biol. 2010;14(3):331–340. doi: 10.1016/j.cbpa.2010.03.022.
69. Gaj T, Gersbach CA, Barbas CF., 3rd ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 2013;31(7):397–405. doi: 10.1016/j.tibtech.2013.04.004.
70. Bauer DE, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science. 2013;342(6155):253–257. doi: 10.1126/science.1242088.
71. Mendenhall EM, et al. Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol. 2013;31(12):1133–1136.
72. Zuber J, et al. An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. Genes Dev. 2011;25(15):1628–1640. doi: 10.1101/gad.17269211.
73. Craig JE, et al. Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. Nature Genetics. 1996;12(1):58–64. doi: 10.1038/ng0196-58.
74. van den Akker E, et al. The majority of the in vitro erythroid expansion potential resides in CD34(-) cells, outweighing the contribution of CD34(+) cells and significantly increasing the erythroblast yield from peripheral blood samples. Haematologica. 2010;95(9):1594–1598. doi: 10.3324/haematol.2009.019828.
75. Brouwer RW, et al. NARWHAL, a primary analysis pipeline for NGS data. Bioinformatics. 2012;28(2):284–285. doi: 10.1093/bioinformatics/btr613.
76. Thongjuea S, et al. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res. 2013;41(13):e132. doi: 10.1093/nar/gkt373.

Chapter 7

The DNA-binding protein CTCF limits proximal V_κ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus

Claudia Ribeiro de Almeida[1], Ralph Stadhouders[2], Marjolein J.W. de Bruijn[1], Ingrid M. Bergen[1], Supat Thongjuea[5], Boris Lenhard[5,6], Wilfred van IJcken[3], Frank Grosveld[2,4], Niels Galjart[2], Eric Soler[2,4] & Rudi W. Hendriks[1]†

[1]Department of Pulmonary Medicine, Erasmus MC, Rotterdam, The Netherlands.
[2]Department of Cell Biology and Genetics, Erasmus MC, Rotterdam, The Netherlands.
[3]Center for Biomics, Erasmus MC, Rotterdam, The Netherlands.
[4]The Cancer Genomics Center, Erasmus MC, Rotterdam, The Netherlands.
[5]Computational Biology Unit-Bergen Center for Computational Science and Sars Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway.
[6]Department of Biology, University of Bergen, Bergen, Norway.

†Corresponding author.

Published in:
Immunity
2011; 35:501-13

Summary

Regulation of immunoglobulin (Ig) V(D)J gene rearrangement is dependent on higher-order chromatin organization. Here, we studied the *in vivo* function of the DNA-binding zinc-finger protein CTCF, which regulates interactions between enhancers and promoters. By conditional deletion of the *Ctcf* gene in the B cell lineage, we demonstrate that loss of CTCF allowed Ig heavy chain recombination, but pre-B cell proliferation and differentiation was severely impaired. In the absence of CTCF, the Igκ light chain locus showed increased proximal and reduced distal V_κ usage. This was associated with enhanced proximal V_κ and reduced J_κ germline transcription. Chromosome conformation capture experiments demonstrated that CTCF limits interactions of the Igκ enhancers with the proximal V_κ gene region and prevents inappropriate interactions between these strong enhancers and elements outside the Igκ locus. Thus, although Ig gene recombination can occur in the absence of CTCF, it is a critical factor determining V_κ segment choice for recombination.

Highlights

- CTCF regulates germline transcription over V kappa and J kappa gene segments
- CTCF critically influences the choice of V kappa segments for recombination
- CTCF restricts Ig enhancer activity within the kappa locus
- CTCF-regulated long-range interactions are not essential for VDJ recombination per se

Introduction

Antigen receptor diversity of lymphocytes is achieved through recombinase activating gene (RAG)-mediated DNA recombination of V (variable), D (diversity), and J (joining) gene segments at the immunoglobulin (Ig) and T cell receptor (Tcr) loci in B and T lymphocytes, respectively (Jung and Alt, 2004 and Schlissel, 2003). The process of V(D)J recombination is regulated at three different levels: lineage specificity, temporal order within a lineage, and allelic exclusion.

B cells develop in the bone marrow (BM) through an orchestrated network of transcription factors and signaling pathways (Nutt and Kee, 2007). Ig heavy-chain (Igh) V(D)J recombination starts at the pre-pro-B cell stage with D_H-to-J_H rearrangement, which precedes V_H-to-DJ_H rearrangement in committed pro-B cells. Productive Igh rearrangement leads to Igμ H chain expression on the cell surface together with the surrogate light chain (SLC) components λ5 and VpreB as the precursor-B cell receptor (pre-BCR). Signals from the pre-BCR and the interleukin-7 receptor (IL-7R) drive proliferation of large pre-B cells (Hendriks and Middendorp, 2004 and Herzog et al., 2009). Upon cessation of proliferation, pre-B cells undergo cellular differentiation and transit to the small pre-B cell stage where Ig κ or λ light-chain (Igl) V_L-to-J_L recombination is initiated (Herzog et al., 2009 and Schlissel, 2003). Productive Igl rearrangement results in surface BCR expression and progression to immature B cells, which are checked for autoreactivity before they leave the BM.

Mechanisms regulating V(D)J recombination are complex and rely on developmental-stage specific changes in locus accessibility to the RAG-recombinase (Jhunjhunwala et al., 2009, Jung and Alt, 2004 and Schlissel, 2003). They include subnuclear relocation, histone modifications, DNA demethylation, germline transcription, antisense intergenic transcription, and locus contraction mediated by looping of individual chromatin domains. Accessibility is controlled by cis-regulatory elements within the Ig loci, such as promoters, matrix attachment regions, silencers, and enhancers, where binding of cell type-specific transcription factors like Pax5, E2A, Ikaros, IRF4, or OBF-1 account for lineage- and developmental-stage specificity of V(D)J recombination (Jhunjhunwala et al., 2009, Jung and Alt, 2004 and Schlissel, 2003). Pax5 and Ikaros were shown to be involved in Igh locus contraction and in their absence only the D_H proximal V_H genes recombine (Fuxa et al., 2004 and Reynaud et al., 2008). Deletion of the transcription factor YY1 also prevented Igh locus contraction, resulting in severely reduced distal V_H rearrangement (Liu et al., 2007). These studies indicate that lineage-specific and ubiquitously expressed transcription factors cooperate to establish higher-order chromatin structures that facilitate long-range interactions required for V_H-DJ_H or V_κ-J_κ rearrangement (Jhunjhunwala et al., 2008 and Jhunjhunwala et al., 2009).

One DNA-binding factor implicated in long-range interactions is the CCCTC-binding factor (CTCF), a ubiquitously expressed and highly conserved 11 zinc finger protein (Phillips and Corces, 2009). CTCF often controls specific interactions by preventing inappropriate communication between neighboring regulatory elements and/or independent chromatin domains, in a developmentally regulated fashion. Gene insulation mediated by CTCF may occur through the formation of chromatin loop domains, as shown for the

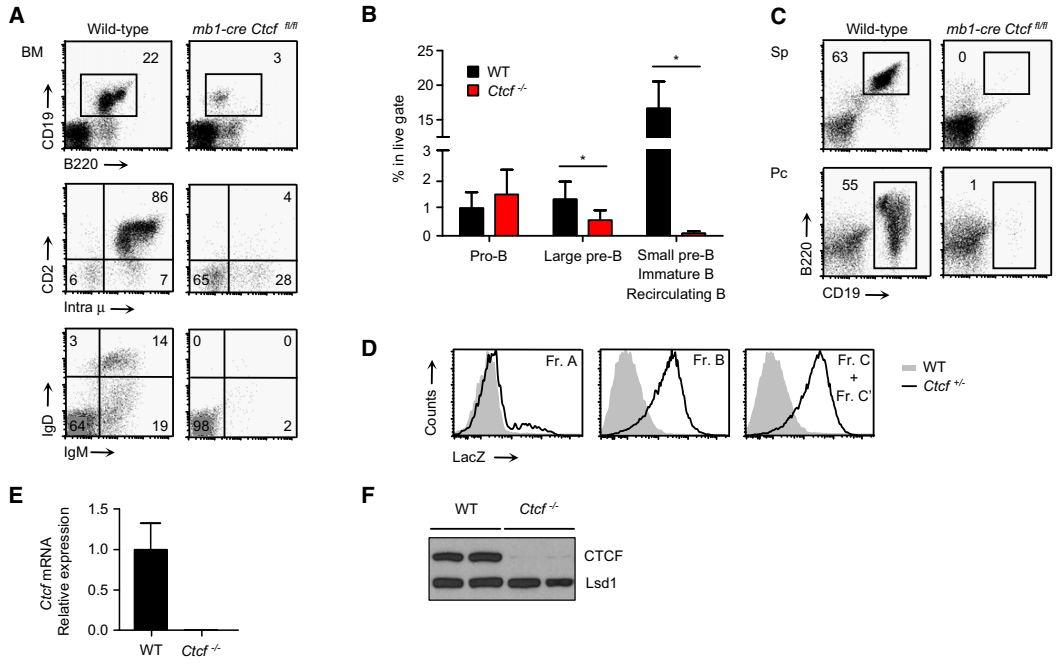


Figure 1. Impaired B Cell Development in *Mb1*-Cre *Ctcf*$^{fl/fl}$ Mice. (A) Flow cytometric analysis of WT and *mb1*-cre *Ctcf*$^{fl/fl}$ total BM cells for expression of B220 and CD19 (top). B220+CD19+ B cell fractions were gated and analyzed for intracellular Igμ-CD2 (middle) or IgM-IgD (bottom). Data are representative of 14–28 mice per genotype. Numbers in dot plots indicate the percentages of cells in each gate. (B) Proportions of cells in live gate were calculated for pro-B (CD2− intracellular Igμ−), large pre-B (CD2− intracellular Igμ+) and total small pre-B, immature B, and recirculating mature B cell fractions (CD2+ intracellular Igμ+) in BM of WT and *mb1*-cre *Ctcf*$^{fl/fl}$ mice (mean and standard deviation [SD], *p<0.001). (C) CD19/B220 flow cytometry profiles of WT and *mb1*-cre *Ctcf*$^{fl/fl}$ spleen (Sp) and peritoneal cavity (PC) lymphoid fractions. Dot plots are representative of 6–7 mice per genotype. (D) Flow cytometric lacZ expression analysis of BM cells of the indicated mice. Results are shown as histogram overlays within three subsets: fraction A (Lin−B220+CD19−HSA−/^{low}CD43+), fraction B (CD19+CD43+BP-1−HSA+), and fraction C+C′ (CD19+CD43+BP-1+HSA+/^{high}). (E and F) Purified B220+CD19+CD2− pro-B or large pre-B cells from BM of the indicated mice were analyzed by quantitative RT-PCR (E) or immunoblotting (F). *Ctcf* expression levels were normalized to the levels of *Gapdh*, whereby the values in WT cells were set to one (mean and SD, for 3 pools of 6–7 mice per genotype). Nuclear protein lysates were analyzed by immunoblotting for CTCF and Lsd1 as a protein loading control (2 pools of 6–7 mice per genotype). See also Figure S1.

imprinted H19-Igf2 locus, the mouse β-globin locus, and at boundaries of domains escaping inactivation on the inactive X chromosome (Splinter et al., 2006; reviewed in Phillips and Corces, 2009). In T helper 1 (Th1) cells, CTCF cooperates with the Th1 cell lineage-specific transcription factor T-bet for proper interferon-γ (IFN-γ) expression via regulation of chromatin looping (Sekimata et al., 2009). We recently showed that CTCF is also a critical regulator of cytokine genes at the Th2 cytokine locus (Ribeiro de Almeida et al., 2009).

A putative role for CTCF in Ig loci long-range interactions was highlighted by the recent mapping of CTCF-binding sites across the Ig loci (Degner et al., 2009, Ebert et al., 2011 and Lucas et al., 2011). Two CTCF-binding sites in the Igh V_H-D_H intergenic region were implicated in the control of lineage-specific and ordered V(D)J recombination by separating the V_H and D_H regions into distinct chromatin domains (Featherstone et al., 2010 and Giallourakis et al., 2010). CTCF sites flank recombination signal sequences (RSSs) for many Igh proximal V_H segments (Lucas et al., 2011). In the distal V_H region, CTCF and E2A were shown to interact with Pax5-activated intergenic repeat (PAIR) elements, which direct antisense transcription (Ebert et al., 2011). Recently, knockdown of CTCF resulted in a modest reduction in Igh locus contraction and increased antisense transcription throughout the D_H region and in distal V_H segments near PAIR elements

(Degner et al., 2011).

Collectively, these data prompted us to investigate the *in vivo* function of CTCF in B cell development. Here we show that conditional *Ctcf* deletion in the B cell lineage still allowed for the generation of cytoplasmic Igμ expressing pre-B cells, although they were severely hampered in proliferation and cellular differentiation. For the Igκ locus, we found preferential recombination and increased germline transcription of proximal V_κ gene segments. Chromosome conformation capture assays coupled to high-throughput sequencing (3C-Seq) (Soler et al., 2010) revealed that CTCF limits interactions of κ enhancers with proximal V_κ genes and prevents inappropriate interactions between these strong enhancers and elements outside the Igκ locus.

Results

Deletion of CTCF Ablates Early B Cell Development

To determine the function of CTCF in B cell development, we crossed *Ctcf* floxed mice (*Ctcf*$^{fl/fl}$) (Heath et al., 2008) with *mb1*-cre mice expressing Cre recombinase specifically in the B cell lineage (Hobeika et al., 2006). Flow cytometric analyses showed severely decreased proportions of B220+CD19+ B lineage cells in the BM of *mb1*-cre *Ctcf*$^{fl/fl}$ mice, when compared with wild-type (WT) controls (Figure 1A). Residual B-lineage cells were mainly intracellular Igμ– pro-B cells, although a detectable fraction expressed intracellular Igμ, indicating productive Igh rearrangement (Figures 1A and 1B). An almost complete block of early B cell development in *mb1*-cre *Ctcf*$^{fl/fl}$ BM was evidenced by the lack of CD2+ small pre-B, immature B, and recirculating mature B cells (Figures 1A and 1B). B220+CD19+ cells were virtually absent in *mb1*-cre *Ctcf*$^{fl/fl}$ spleen or peritoneal cavity (Figure 1C). Crosses of *mb1*-cre *Ctcf*$^{fl/fl}$ mice with mice carrying the antiapoptotic Eμ-Bcl2 transgene (Strasser et al., 1991) demonstrated that the developmental block of CTCF-deficient pre-B cells could not be explained by defective survival only (see Figure S1 available online).

We used the LacZ reporter in the targeted *Ctcf* allele (Heath et al., 2008) to evaluate the efficiency of deletion (Figure S1). Consistent with previously reported highly efficient gene deletion at the earliest stages of B cell development with *mb1*-cre mice (Hobeika et al., 2006 and Liu et al., 2007), *Ctcf* deletion occurred already at the pre-pro-B cell stage (fraction A) (Hardy et al., 1991) and was almost complete from the pro-B cell stage (fraction B) onward (Figure 1D). Accordingly, CTCF mRNA and protein were essentially undetectable in B220+CD19+ B lineage fractions purified from *mb1*-cre *Ctcf*$^{fl/fl}$ BM (Figures 1E and 1F).

These findings show that in *mb1*-cre *Ctcf*$^{fl/fl}$ mice CTCF expression is efficiently ablated in early stages of B cell development and that CTCF is essential beyond the pre-B cell stage.

Defective Pre-B Cell Proliferation and Differentiation in Mb1-Cre Ctcf$^{fl/fl}$ *Mice*

The pre-BCR acts as a checkpoint that monitors functional Igh rearrangement and induces, together with IL-7R signaling, clonal expansion and survival of Igμ+ large pre-B cells. Upon cessation of proliferation, pre-BCR signals are additionally required for developmental progression of large into small pre-B cells (Hendriks and Middendorp, 2004 and Herzog et al., 2009).

Flow cytometric analysis of Igμ+ pre-B cells from *mb1*-cre *Ctcf*$^{fl/fl}$ mice showed reduced cell size, indicating defective proliferation (Figure 2A). Additionally, *mb1*-cre *Ctcf*$^{fl/fl}$ pro-B and large pre-B cells both showed decreased expression of IL-7R α chain by flow cytometry (Figure 2A) and RT-PCR (Figure S2C). To directly examine the cell cycle status of B-lineage cells in *mb1*-cre *Ctcf*$^{fl/fl}$ mice, we measured DNA content by propidium iodide staining. Although the proportions of cycling pro-B cells were similar, we observed a considerable reduction in the proportions of cycling large pre-B cells in *mb1*-cre *Ctcf*$^{fl/fl}$ mice, when compared with WT mice (~26% versus ~47%, respectively, Figure 2B). Subsequently, we determined the proliferation capacity *in vivo* by pulsing with a single dose of the thymidine analog 5-bromodeoxyuridine (BrdU), which is selectively incorporated into the DNA of cycling cells (Middendorp et al., 2002). Flow cytometric analysis revealed comparable proportions of BrdU+ pro-B cells, but a ~50% reduction in the fractions of BrdU+ large pre-B cells in *mb1*-cre *Ctcf*$^{fl/fl}$ mice, when compared with WT controls (Figure 2C).

Consistent with a strong pre-B cell arrest, the majority of B220+CD19+ cells in *mb1*-cre *Ctcf*$^{fl/fl}$ BM expressed the early pro-B cell-specific markers c-Kit, CD43, and the SLC component λ5 and failed to upregulate CD2, CD25, and MHC class II expression (Figure S2A). Quantitative RT-PCR analyses of purified B220+CD19+CD2– cells from WT and *mb1*-cre *Ctcf*$^{fl/fl}$ BM showed proper specification and commitment to

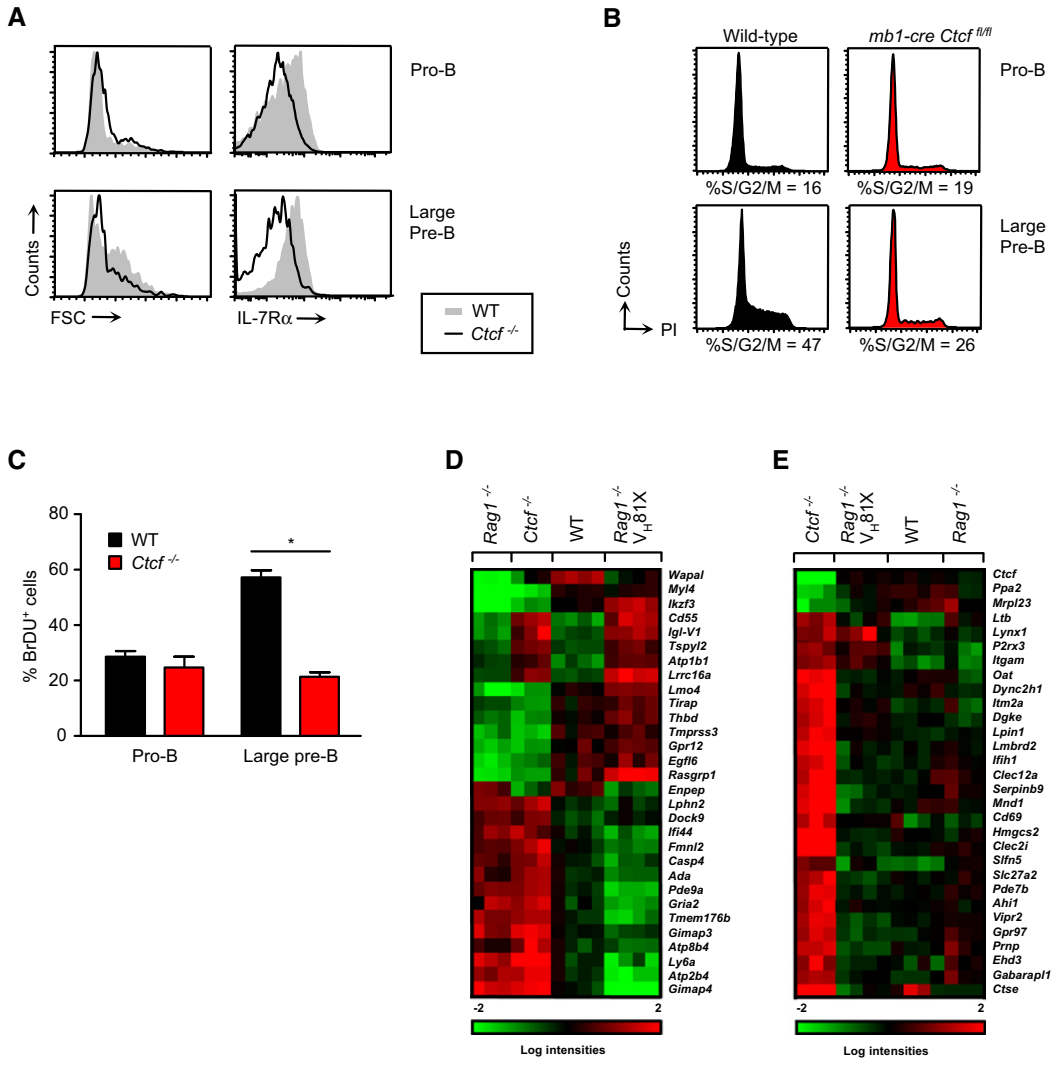


Figure 2. Defective Pre-B Cell Proliferation and Differentiation in *Mb1*-Cre *Ctcf*$^{fl/fl}$ Mice. (A) Flow cytometric analysis of WT and *mb1*-cre *Ctcf*$^{fl/fl}$ B220+CD19+ pro-B cells (CD2– intracellular Igμ–) and large pre-B cells (CD2– intracellular Igμ+) for cell size (forward side scatter [FSC]) and IL-7Rα. Results are displayed as histogram overlays (4–6 mice per genotype). (B) Propidium iodide (PI) cell cycle analysis of B220+CD19+ pro-B cells (CD2– intracellular Igμ–) and large pre-B cells (CD2– intracellular Igμ+) purified from WT and *mb1*-cre *Ctcf*$^{fl/fl}$ BM. Percentages of cells in cycle (S/G2/M; >2N DNA content) are shown (representative of two mice per genotype). (C) *In vivo* proliferation analysis of WT and *mb1*-cre *Ctcf*$^{fl/fl}$ B220+CD19+ pro-B cells (CD2– intracellular Igμ–) and large pre-B cells (CD2– intracellular Igμ+). Mice were i.p. injected with a single dose of BrdU and after 4 hr, the percentages of BrdU+ cells were determined by flow cytometry (mean values and SD for 3–4 mice per genotype; *$p < 0.001$). (D and E) DNA microarray analysis of total mRNA from purified B220+CD19+CD2– pro-B/large pre-B cell fractions in WT and *mb1*-cre *Ctcf*$^{fl/fl}$ mice. Genes differentially expressed between the two genotypes were subdivided into two groups: genes in which expression did (D) or did not differ (E) between control *Rag1*$^{-/-}$ pro-B cell and V_H81X *Rag1*$^{-/-}$ pre-B cell fractions. Heatmaps for the 30 genes with highest fold change in expression between WT and *mb1*-cre *Ctcf*$^{fl/fl}$ B cell progenitors are shown (for complete gene lists, see Tables S1 and S2). On the bottom is the logarithmic quantitative scale for gene expression (3–4 pools of 3–7 mice per genotype). See also Figure S2.

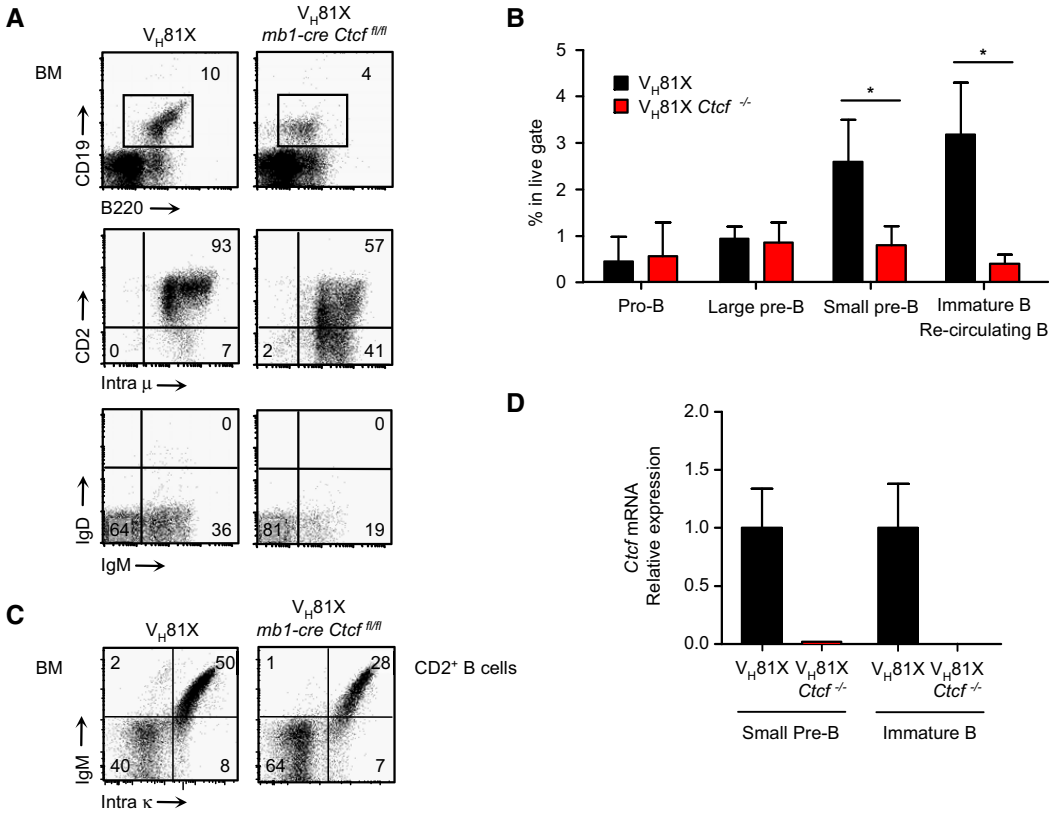


Figure 3. B Cell Development in V_H81X Transgenic *Mb1*-Cre *Ctcf*$^{fl/fl}$ Mice. (A) Flow cytometric analysis of V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM cells for the expression of B220/CD19 (top). Total B220+CD19+ B cell fractions were gated and analyzed for expression of intracellular Igμ/CD2 (middle) or IgM/IgD (bottom). Data are representative of 7–14 mice per genotype. Numbers in dot plots indicate the percentages of cells in each gate. (B) Proportions of cells in live gate were calculated for pro-B (CD2– intracellular Igμ–), large pre-B (CD2– intracellular Igμ+), small pre-B (CD2+IgM–) and immature B/recirculating mature B cell (CD2+IgM+) fractions in the BM of V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mice (mean values and SD, *p<0.001). (C) Flow cytometric analysis of gated B220+CD19+CD2+ cells from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM for the expression of IgM/intracellular Igκ. Data are representative of 13–14 mice per genotype. (D) Total RNA isolated from purified B220+CD19+CD2+ IgM– (small pre-B cells) and IgM+ (immature/recirculating mature B cells) populations from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM was analyzed by quantitative RT-PCR for *Ctcf* expression. *Ctcf* expression levels were normalized to the levels of *Gapdh* mRNA, whereby the values in V_H81X cells were set to one (mean values and SD, for three pools of six to seven mice per genotype). See also Figure S3.

the B cell fate, given that CTCF-deficient cells still expressed the early B lineage genes *Tcfe2a*, *Ebf1*, *Pax5*, *Il7ra*, *Cd79a*, and *Cd79b*, albeit often at slightly reduced levels (Figure S2A). For genes that are normally up or downregulated upon pre-BCR signaling (*Igll1*, *Vpreb1*, *Irf4*, *Ikzf1*) expression levels in CTCF-deficient B220+CD19+CD2– cells were between those in *Rag1*$^{-/-}$ and WT fractions (Figure S2A), indicating that pre-BCR signaling was not completely abrogated in the absence of CTCF.

Genome-wide expression profiling of purified and CTCF-deficient CD2– B220+CD19+ fractions (containing pro-B and pre-B cells) revealed that 174 genes were differently expressed. Approximately 50% of these genes demonstrated a pro-B cell signature for CTCF-deficient B cell precursors, given that these genes were also differentially expressed between *Rag1*$^{-/-}$ pro-B cells and V_H81X Igh transgenic *Rag1*$^{-/-}$ pre-B cells (Figure 2D; Table S1). Other genes were upregulated in the absence of CTCF, irrespective of B cell differentiation stage (Figure 2E; Table S2). In summary, our findings show that CTCF-deficient cytoplasmic Igμ+

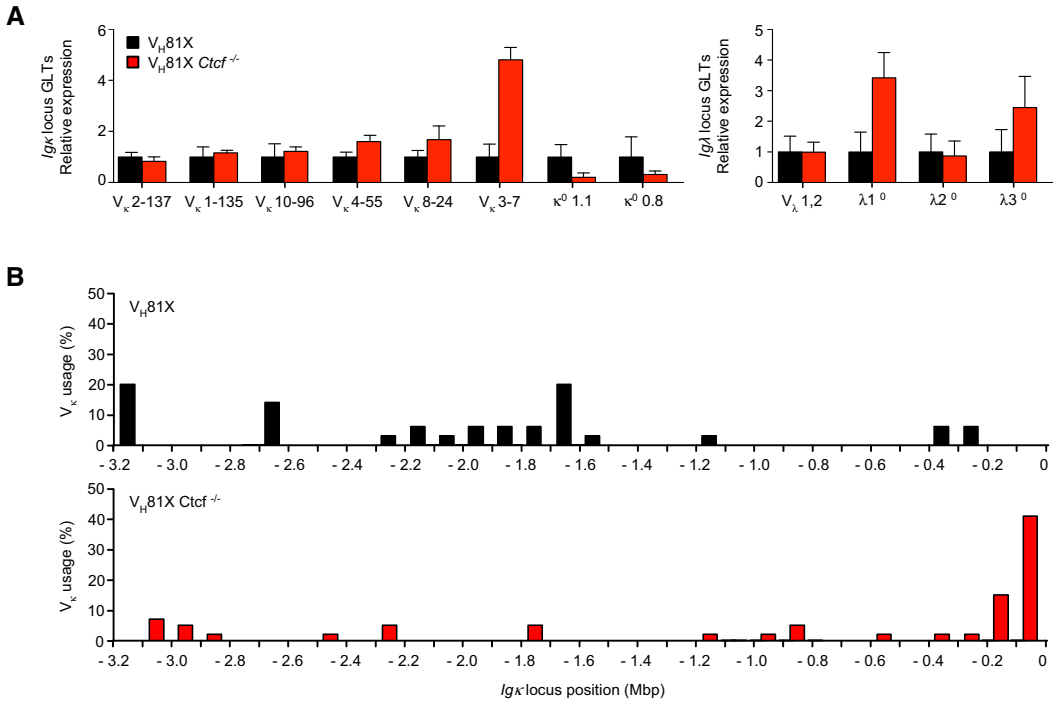


Figure 4. Proximal V_κ Usage in V_H81X Transgenic *Mb1*-Cre *Ctcf*$^{fl/fl}$ (pre-)B Cells. (A) Quantitative RT-PCR analysis for Igκ and Igλ locus germline transcription in purified B220+CD19+CD2+IgM– small pre-B cell populations from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM. Expression levels of different germline transcripts (GLTs) were normalized to the levels of *Gapdh*, whereby the values in V_H81X small pre-B cells were set to one. Data represent V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mean values and SD for three independent pools of three to five mice. (B) DNA sequencing analysis of V_κ gene segment usage of non-productive alleles from B220+CD19+CD2+IgM– small pre-B cell and B220+CD19+CD2+IgM+ (im)mature B cell populations purified from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM. Genomic DNA was isolated and used for PCR amplification of V_κ-J_κ2 recombination products, which were further cloned and analyzed by DNA sequencing. Data represent relative frequency of V_κ usage per 0.2 Mbp intervals in the Igκ locus. A schematic representation of the Igκ locus (top) shows the location of V_κ genes in which germline transcription was analyzed in (A). Data are from three independent pools of three to five mice per genotype (number of sequences analyzed: 35 sequences for V_H81X B cells; 41 sequences for V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ B cells). See also Figure S4.

pre-B cells manifested defective proliferation and a severe block of cellular differentiation.

Igh V(D)J Rearrangement Occurs in CTCF-Deficient Pro-B Cells

The presence of intracellular Igμ+ pre-B cells in *mb1*-cre *Ctcf*$^{fl/fl}$ BM indicated successful Igh gene rearrangement in the absence of CTCF. In addition, RT-PCR analysis revealed normal levels of V_HJ558, V_H7183, and Iμ germline transcripts in B220+CD19+CD2– fractions from *mb1*-cre *Ctcf*$^{fl/fl}$ mice, suggesting unaltered Igh locus accessibility (Schlissel, 2003) (Figures S2D and S2E). We used quantitative RT-PCR to compare V_H family usage (Fuxa et al., 2004) in purified B220+CD19+CD2– fractions from WT and *mb1*-cre *Ctcf*$^{fl/fl}$ mice. We included μMT mice harboring a targeted deletion of the Igμ membrane exon (Kitamura et al., 1991) as a control because they parallel CTCF-deficient mice in that pre-BCR-induced proliferative expansion and concomitant selection for particular V_H segments in the context of the pre-BCR (ten Boekel et al., 1997) is absent. We found that in the absence of CTCF proximal (V_H7183) as well as distal (V_HJ558) gene segments were used, whereby their relative usage was close to that of μMT mice (Figure S2F). Thus, Igh gene recombination, even to distal V_H gene segments, occurred in CTCF-deficient pro-B cells.

A Productively Rearranged Igh Transgene Allows CTCF-Deficient Cells to Develop beyond the Pre-B Cell Stage

The nearly complete developmental block observed in *mb1*-cre *Ctcf*$^{fl/fl}$ mice precluded the analysis of CTCF function past the pre-B cell stage. However, upon introduction of the functionally pre-rearranged Igμ transgene V_H81X (Martin et al., 1997), B cell differentiation was partially rescued: we found substantial populations of CD2+ and surface IgM+ B-lineage cells in V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM (Figure 3A). Although the V_H81X transgene did not appear to rescue the proliferation defect of CTCF-deficient large pre-B cells (Figures S3A–S3C), the proportions of large pre-B cells *in vivo* were not significantly different between V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM (Figure 3B). Furthermore, we observed a partial correction of the expression profiles of the developmentally regulated markers c-Kit, CD43, CD2, CD25, MHC class II, and λ5 (Figure S3D). V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM manifested decreased proportions of CD2+ small pre-B cells, immature and recirculating mature IgM+ B cells, when compared with V_H81X controls (Figures 3A and 3B). However, intracellular Igκ L chain expression in surface IgM– small pre-B cells and IgM+ immature B cells was quite similar in the two groups of mice (Figure 3C).

The reduced size of the small pre-B and immature B cell population in V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM prompted us to investigate the kinetics of the developmental progression of pre-B cells *in vivo* by BrdU injection. We found only a minor developmental delay of ~1.5 hr in CTCF-deficient V_H81X Igκ+ immature B cells (~12 hr), compared with V_H81X controls (~10.5 hr, Figures S3E and S3F). B220^{low}CD19+ B cells were detected in the spleen at low numbers but were absent in peritoneal cavity of V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mice (Figure S3G). As assessed by quantitative RT-PCR, CTCF mRNA was strongly reduced in B220+CD19+ subsets purified from V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM (Figure 3D). Thus, expression of the V_H81X transgene allowed significant differentiation of CTCF-deficient cells beyond the large pre-B cell stage.

Increased Proximal V_κ Gene and Reduced J_κ Germline Transcription in the Absence of CTCF

In small pre-B cells, successful Igl V-to-J recombination requires that these gene segments are accessible to the recombination machinery, which is reflected by germline transcription (Schlissel and Baltimore, 1989) (Schlissel, 2003). We determined Igκ and Igλ locus germline transcription in sorted CD2+ small pre-B cell fractions by quantitative RT-PCR (Düber et al., 2003 and Inlay et al., 2006). We found remarkably increased germline transcription of the proximal V_κ gene segment V_κ3-7 in *Ctcf*-deficient V_H81X small pre-B cells, compared with V_H81X controls (Figure 4A). Germline transcription of more distal V_κ gene segments was only slightly increased or not significantly different. In contrast, germline transcripts initiating from promoters located upstream of J_κ ($\kappa^0$0.8, $\kappa^0$1.1) (Grawunder et al., 1995) were substantially reduced in the absence of CTCF (Figure 4A). Germline transcription from the J_λ1 and J_λ3 clusters ($\lambda1^0$ and $\lambda3^0$, respectively (Engel et al., 1999) were increased, whereas V_λ1,2 (Düber et al., 2003) and $\lambda2^0$ germline transcripts were not affected (Figure 4A). In conclusion, loss of CTCF resulted in increased germline transcription of the V_κ proximal region and reduced κ0 germline transcription over the J_κ region in pre-B cells.

Proximal V_κ Usage in V_H81X Transgenic CTCF-Deficient B Cells

Next, we analyzed V_κ gene usage by DNA sequencing of V_κ-J_κ recombination products from purified pre-B and B cells. To exclude repertoire effects of BCR-mediated selection, we first focused on non-productively rearranged alleles from control (n = 35) and CTCF-deficient (n = 41) V_H81X (pre-)B cells. We calculated V_κ gene usage for 100 kb intervals within the Igκ locus. In control V_H81X (pre-)B cells, V_κ usage was diverse and for > 80% directed to the middle and distal regions of the Igκ locus (Figure 4B). Remarkably, in CTCF-deficient V_H81X (pre-)B cells, >50% of all V_κ segments used were located in the most proximal ~200 kb region, exclusively containing members of the V_κ3 family. Importantly, this region was not used in control V_H81X (pre-)B cells (Figure 4B).

It is conceivable that CTCF indirectly controls Igκ locus recombination. CTCF might affect survival and thereby receptor editing, the process of ongoing IgL chain recombination that serves to replace autoreactive BCR specificities. However, when we excluded cells that had performed receptor editing by analyzing V_κ gene usage in J_κ1 recombination products only, we still observed increased V_κ3 usage in CTCF-deficient cells (24 out of 31, as compared with 0/14 in WT cells). We also found that in the absence of CTCF both

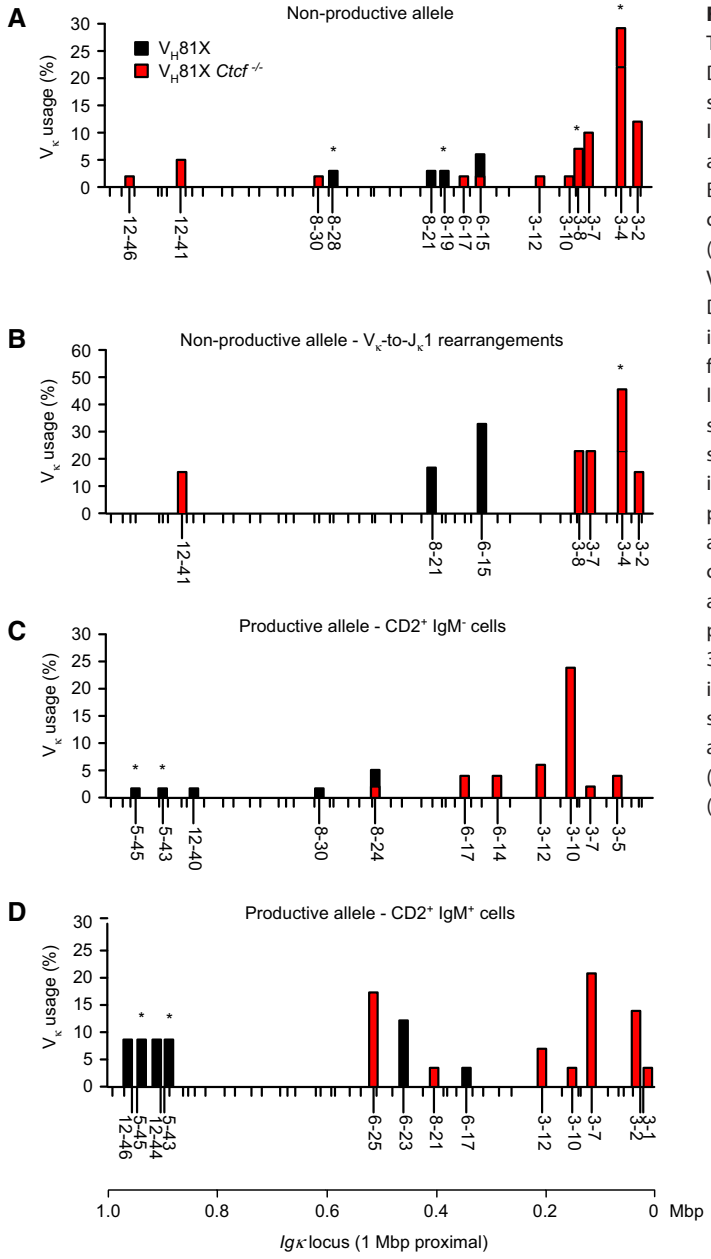


Figure 5. Proximal $V_κ$ Repertoire in V_H81X Transgenic *Mb1*-cre *Ctcf*$^{fl/fl}$ (pre-)B Cells. DNA sequencing analysis of $V_κ$ gene segments used in total non-productive Igκ alleles (A), non-productive $V_κ$-to-$J_κ$1 alleles (B), productive Igκ alleles from B220+CD19+CD2+IgM– small pre-B cells (C), and B220+CD19+CD2+IgM+ (im)mature B cells (D) purified from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ BM. Data represent relative frequency of individual $V_κ$ gene segments used for the proximal 1 Mbp region of the Igκ locus. In each panel, the x-axis shows the location of various $V_κ$ gene segments. Collective data are from three independent pools of three to five mice per genotype (number of sequences analyzed for V_H81X and for V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mice are for non-productive alleles: 35 and 41, respectively; for productive alleles in IgM– small pre-B: 32 and 30; and for productive alleles in IgM+ B cells: 36 and 30. Some DNA sequences cannot exclusively be assigned to a single $V_κ$ gene segment (see 8-19/8-28, 3-4/ 3-8 and 5-43/5-45 (marked with an asterisk).

inversional and deletional Igκ recombination events occurred (Figure S4A). Thus, reduced receptor editing or specific defects in either inversional or deletional recombination cannot explain the increased $V_κ$3 usage in the absence of CTCF. It remained possible that CTCF indirectly controls Igκ locus recombination through regulation of other transcription factors, but we found that loss of CTCF did not significantly affect the expression of nuclear proteins implicated in Igκ locus recombination (Schlissel, 2003), including *Tcfe2a*, *Id3*, *Ikzf3*, *Irf4*, *Irf8*, *Pou2af1*, *Ccdn3*, *Rag1*, and *Rag2* (Figure S4B).

Taken together, these findings show that in the absence of CTCF $V_κ$-$J_κ$ recombination activity is preferentially targeted to $V_κ$3, consistent with the observed increased germline transcription over the $V_κ$ proximal region.

BCR-Mediated Selection Still Occurs in CTCF-Deficient Immature B Cells

Detailed analysis of the Igκ locus 1.0 Mbp proximal region showed that in the absence of CTCF $V_κ$ usage in non-productive rearrangements was highly restricted to the $V_κ3$ family, whereby $V_κ3$-4 was dominant (Figure 5A). Analysis of $V_κ$-to-$J_κ1$ rearrangements yielded a similar distribution, showing that the observed differences between control and CTCF-deficient alleles were not dependent on receptor editing events (Figure 5B). In addition, productive alleles in CTCF-deficient cells manifested preferential $V_κ3$ usage (32/60 alelles, versus 0/60 in WT controls). Hereby $V_κ3$-10 and $V_κ3$-7 segments were predominantly used in surface IgM– small pre-B cells and surface IgM+ B cells, respectively (Figures 5C and 5D). The finding that relative frequencies of individual $V_κ3$ family members differed considerably between productive and unproductive Igκ alleles indicates that BCR-mediated selection still occurred in the absence of CTCF. Moreover, the observed increased usage of $V_κ3$ in the absence of CTCF affected all $V_κ3$ family members.

$V_κ$ Usage Is Correlated with CTCF-Binding Sites in the Igκ Locus

Next, we identified CTCF-binding sites in the Igκ locus in cultured primary pre-B cells using chromatin immunoprecipitation coupled to high-throughput sequencing (ChIP-Seq) (Figure 6A). We identified predominant CTCF binding at the 5′ and 3′ boundaries of Igκ locus, as well as at the SIS (silencer in intervening sequence) recombination silencer element residing in the $V_κ$-$J_κ$ region (Liu et al., 2006), in agreement with reported findings (Degner et al., 2009). The SIS element has been shown to negatively regulate rearrangement and to specify targeting to centromeric heterochromatin (Liu et al., 2006). In contrast to the previously reported low density of CTCF occupancy in the Igκ locus (Degner et al., 2009), we found ~60 CTCF-binding sites, which were not evenly distributed throughout the Igκ locus. We identified five regions with a high density of CTCF sites (regions H1–H5, Figure 6A), and four regions of 150–250 kb with low CTCF occupancy (regions L1–L4, Figure 6A). These four regions, including the proximal region containing $V_κ3$-family segments (L4; pos. 0 to −250 kb), contained 29 $V_κ$ gene segments, which were rarely used in WT mice (only 1/35 nonproductive alleles). In contrast, regions H2 and H3 contain $V_κ$ gene segments that were frequently used in WT mice (Figure 6B).

Because in the Igh locus proximal V_H segments are frequently associated with nearby CTCF sites (Lucas et al., 2011), we examined CTCF occupancy and $V_κ$ segment localization. We found that 26 of the ~60 CTCF sites were located near (<5 kb) a $V_κ$ gene segment (Figure 6A). $V_κ$ gene segments with nearby CTCF sites were mainly present in the proximal half of the Igκ locus and were more often used in CTCF WT (9/35) than in CTCF KO (1/41) non-productive rearrangements.

In summary, we identified ~60 CTCF-binding sites in the Igκ locus in pre-B cells. Four regions of 150–250 kb, including the most proximal $V_κ$ region, lacked CTCF occupancy and $V_κ$ segments in these regions were rarely used in WT Igκ alleles.

Loss of CTCF Affects Interactions between the SIS Element and the Proximal $V_κ$ Region

Because of the presence of a predominant CTCF peak at the SIS element, we decided to determine genome-wide interactions mediated by the SIS region in the presence and absence of CTCF, using chromosome conformation capture coupled to high-throughput sequencing (3C-seq) (Soler et al., 2010). To ensure analysis of long-range interactions in non-rearranged Igκ loci, we crossed V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mice on the *Rag1*$^{-/-}$ background. Quantitative RT-PCR analysis of $V_κ$ germline transcription in sorted B220+CD19+ pre-B cell fractions confirmed that also on the *Rag1*$^{-/-}$ background loss of CTCF was associated with increased germline transcription of the proximal $V_κ$ gene segment $V_κ3$-7, compared with CTCF-expressing V_H81X *Rag1*$^{-/-}$ pre-B cells (Figure S6). We purified B220+CD19+ pre-B cell populations from control and CTCF-deficient V_H81X *Rag1*$^{-/-}$ BM and performed 3C-Seq experiments, whereby we included E13.5 fetal liver cells as controls. To facilitate direct comparison with $V_κ$ gene usage, we calculated the number of interactions for 100 kb intervals within the Igκ locus. Figure 6B shows the long-range interactions identified within a 9 Mb region encompassing the Igκ locus and its flanking regions.

In V_H81X *Rag1*$^{-/-}$ small pre-B cells, the SIS element showed interactions throughout the Igκ locus, also with regions located at considerable distance (~-3.2 Mbp). Major interactions were found with regions

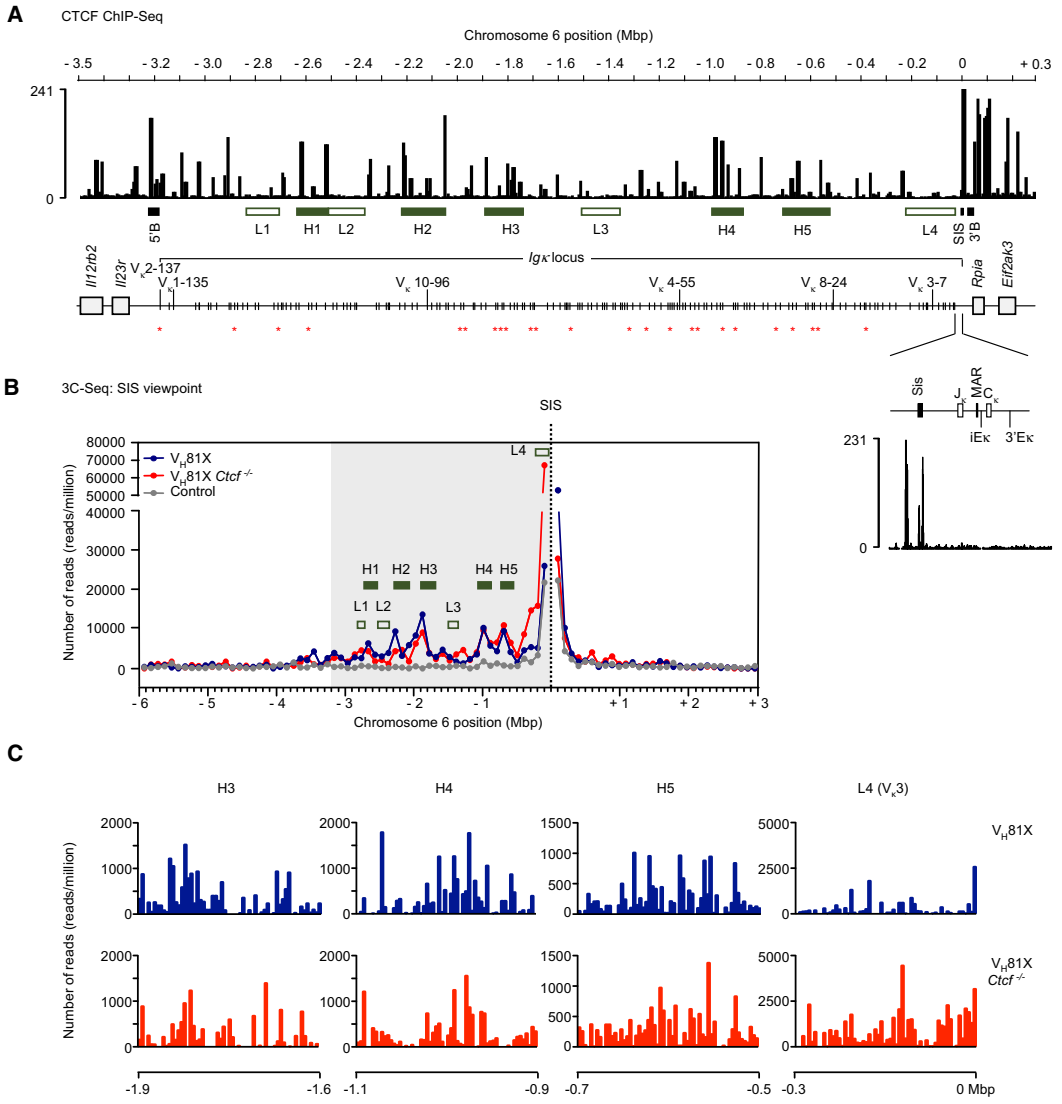


Figure 6. The Igκ Locus: Regions with High and Low CTCF Occupancy and the Effects of CTCF Deficiency on Long-Range Chromatin Interactions. (A) Schematic representation of the Igκ locus and CTCF-occupancy in pre-B cells, as determined by ChIP-Seq. Strong CTCF sites at the 3' and 5'boundaries of the locus (5'B and 3'B), as well as within the locus (H1-H5) are indicated. L1-L4 represent regions with low CTCF-occupancy. (B) 3C-Seq analysis of long-range interactions with the SIS region in B220+CD19+ small pre-B cell populations purified from V_H81X *Rag1*$^{-/-}$ and V_H81X *Rag1*$^{-/-}$ *mb1*-cre *Ctcf*$^{fl/fl}$ BM and total fetal liver cells. Cross-linked and BglII-digested DNA fragments were ligated and subsequently digested by NlaIII followed by re-ligation. Viewpoint-specific primers on the fragment of interest (containing the SIS viewpoint) were used for generating 3C-Seq libraries (see Experimental Procedures for details). The graph shows the number of reads per million in 0.1 Mbp intervals in the 3.2 Mb Igκ locus (shaded area) plus 3.0 Mbp upstream and 2.8 Mbp downstream genomic regions. The dashed line represents the viewpoint. 3C-Seq counts obtained for fragments adjacent to the viewpoints were excluded from the analysis. (C) Number of reads per million obtained for each of the 60 BglII fragments that cover the regions indicated in (B). In each graph, the x axis shows the location in relation to the viewpoint (in Mb). See also Figure S5.

H1–H5, for which we observed high CTCF occupancy. On the other hand, signals were low for the regions L1–L4, where CTCF occupancy was low. In contrast, in non-rearranging control fetal liver cells genomic contacts were reduced to background levels throughout the Igκ locus (Figure 6B).

Importantly, when we analyzed CTCF-deficient V_H81X transgenic *Rag1*$^{-/-}$ pre-B cells, we found that the interactions between the SIS region and most of the Igκ locus were not different from those identified in CTCF-expressing pre-B cells. Only at the proximal $V_κ$ region (0–300 kb, containing the L4 region) did the SIS region manifest significantly increased interactions (Figure 6B), consistent with increased $V_κ$ usage in the absence of CTCF. Detailed views of $V_κ$ regions with high levels of interaction (H3 and H4) and of the proximal L4 region are shown in Figure 6C.

In summary, these findings demonstrate long-range interactions between the SIS-region and the entire Igκ locus, whereby regions with high numbers of identified contacts correlate with the localization of prominent CTCF-binding sites. Nevertheless, the long-range interactions in pre-B cells between the SIS and $V_κ$ region were not notably affected by loss of CTCF; only in the proximal $V_κ$ region did the loss of CTCF result in significantly increased interactions.

CTCF Restricts the Activity of the Enhancer Elements in the Igκ Locus

CTCF is thought to function in spatial organization of chromatin topology via loop formation, whereby the positioning of CTCF-binding sites with respect to genes and regulatory elements dictates the types of CTCF-based chromatin loop structures formed (Phillips and Corces, 2009). As a result, CTCF-mediated contacts may either confer enhancer blocking or may enable promoter-enhancer interactions. In rearranged and actively transcribed Igκ alleles, long-range interactions between active $V_κ$ gene promoters and the intronic and 3′ enhancers (iEκ and 3′Eκ) are essential for $V_κ$-$J_κ$ recombination (Liu and Garrard, 2005). Therefore, we next investigated whether CTCF-mediated looping controlled $V_κ$-$J_κ$ recombination by regulating the spatial proximity of $V_κ$ gene segments relative to the Igκ enhancer elements.

In 3C-seq experiments using the iEκ and 3′Eκ enhancers as viewpoints, the identified interactions in the Igκ locus paralleled those found for the SIS region, with clear peaks at the H1–H5 regions (Figures 7A and 7B). Although many long-range interactions were preserved in the absence of CTCF, we observed an altered distribution of enhancer contacts: increased interactions with the most proximal part of the $V_κ$ region (H4 and H5 and particularly L4; detailed analyses in Figures 7C and 7D) and decreased distal interactions. In particular, 3′Eκ contacts near the 5′ boundary region, containing the two most distal $V_κ$ genes ($V_κ$2-137 and $V_κ$1-135) often used in control (~20%, Figure 4B) but not in CTCF-deficient (pre-)B cells, were reduced (Figures 7B and 7D).

Loss of CTCF also changed interaction of the κ enhancers with regions outside the Igκ locus. Beyond the 5′ boundary of the Igκ locus, loss of CTCF was associated with increased iEκ interactions at −4.2 Mbp (region O1) as well as with reduced interactions at −3.7 Mbp (region O2) (Figures 7A and 7C). A large region of ~0.8 Mb, located downstream the 3′ boundary of the Igκ locus (region O3) showed high amounts of interaction with the κ enhancers in CTCF-deficient pre-B cells (Figures 7A, 7B and 7D).

Finally, to exclude the possibility that the observed interactions in our 3C-Seq experiments were affected by the absence of focal Rag1 binding to the $J_κ$ region (Ji et al., 2010), we also performed 3C-Seq experiments for the iEκ and 3′Eκ enhancer viewpoints in sorted pre-B cells from V_H81X and V_H81X *mb1*-cre *Ctcf*$^{fl/fl}$ mice that were on a *Rag1*-proficient background. In these analyses, we also found that in the absence of CTCF contacts between the enhancers and the proximal $V_κ$ region or regions outside the Igκ locus were increased (Figure S6).

Taken together, these 3C-seq analyses revealed that interaction between the SIS, iEκ, and 3′Eκ elements and the Igκ V region did not require CTCF per se. The absence of CTCF significantly increased the interactions between the enhancer elements and the proximal $V_κ$ region and regions outside the Igκ locus.

Discussion

In this study, we used conditional *Ctcf* gene targeting to investigate the function of CTCF during B cell development *in vivo*. We found that CTCF was critical at the pre-B cell developmental checkpoint, probably as a regulator of genes involved in proliferation and cellular differentiation, but V(D)J recombination still occurred when the *Ctcf* gene was deleted. We studied the Igκ locus in detail and found that loss of CTCF resulted in (1) increased proximal $V_κ$ and reduced $J_κ$ germline transcription, (2) increased recombination to

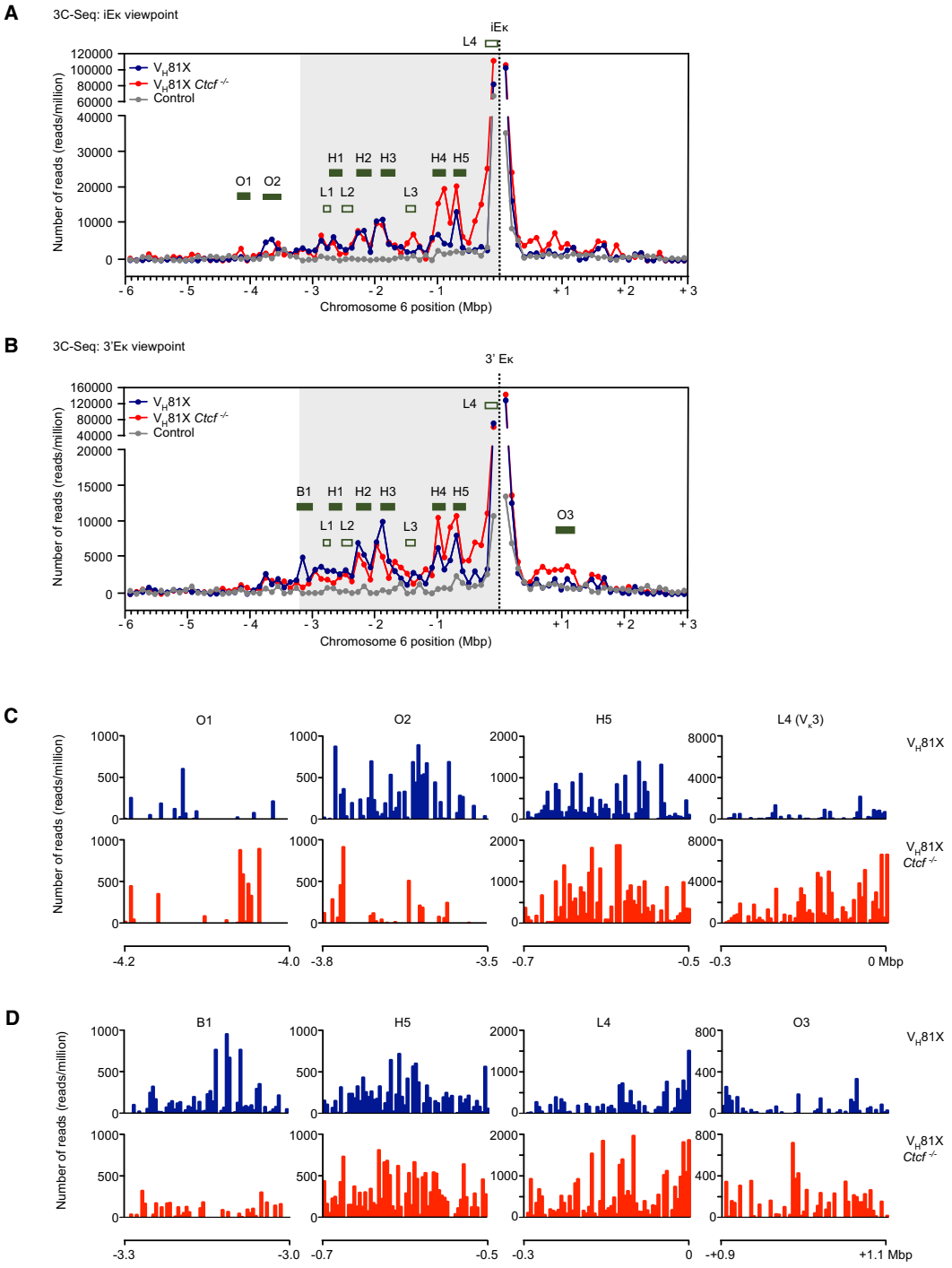


Figure 7. *(Legend at the bottom of the next page)*

proximal $V_κ$ genes, (3) decreased usage of more distal $V_κ$ genes, particularly of the two most distal $V_κ$ genes, and (4) increased interactions of the iEκ and 3'Eκ enhancers with the proximal $V_κ$ region and with elements

outside the Igκ locus. Our 3C-seq experiments revealed that long-range interactions between the SIS silencer, iEκ or 3'Eκ elements and the Igκ V region did not require CTCF per se. Rather, we conclude that CTCF is required for the specificity of interactions between these regulatory elements and V_κ segments, thereby restricting Igκ enhancer activity and controlling V_κ gene segment choice.

We demonstrated that in CTCF-deficient pre-B cells proximal V_κ genes are preferentially used for recombination, consistent with increased germline transcription and increased iEκ and/or 3'Eκ enhancer interactions with the V_κ proximal region. In contrast, germline $\kappa^0$0.8/1.1 transcripts, initiated from promoters located upstream of J_κ, were considerably reduced. On the basis of CTCF occupancy in the Igκ locus and the observed long-range interactions in the presence and absence of CTCF, we propose a model (Figure S6). In this model, CTCF activity regulates Igκ locus recombination by orchestrating functional communications between V_κ gene segments and enhancers, while limiting the actions of Igκ locus specific cis- and trans-acting factors elsewhere in the genome. Strong CTCF-binding sites flanking the Igκ locus and in the SIS element in the V_κ-J_κ intergenic region (Degner et al., 2009) would partition off the locus into three main chromatin loop domains, separating the J_κ-C_κ cluster containing the iEκ-3'Eκ enhancers, the proximal part of the V_κ region (with only V_κ3 family segments) and the remaining central and distal parts of the V_κ region. Hence, CTCF would prevent inappropriate communication between the Igκ enhancers and the promoters of proximal V_κ3 gene segments or genomic regions outside the Igκ locus. On the basis of our 3C-Seq experiments showing that in the absence of CTCF most of the long-range interactions within the Igκ locus were conserved, we propose that loss of CTCF has limited effects on the global architecture of the Igκ locus, except for the V_κ3 region and the very distal V region containing V_κ2-137 and V_κ1-135 (Figure S6). CTCF-dependent chromatin looping of the J_κ-C_κ cluster and the iEκ/3'Eκ enhancers would also restrict V_κ3 transcription and ensure proper J_κ germline transcription. Consistent with this, deletion of the SIS element results in reduced levels of $\kappa^0$0.8/1.1 germline transcripts (Liu et al., 2006). Very recently, it was shown that mice possessing a 3.7 kb targeted deletion of the SIS element, associated with reduced occupancy of Ikaros and CTCF, display enhanced proximal V_κ usage (Xiang et al., 2011). Increased proximal V_κ usage resulting from loss of CTCF can therefore be attributed to the CTCF sites present in the SIS element.

In our model, dynamic scanning of the V_κ region for recombination would depend on further regulatory sub-loops bringing V_κ gene segments into close spatial proximity with the J_κ-C_κ cluster containing the iEκ-3'Eκ enhancers. Interestingly, recent data demonstrate that the highly active chromatin region encompassing the J_κ segments exhibits focal RAG protein binding (Ji et al., 2010). In this region, V_κ gene segments compete for capture by RAG proteins and therefore it is referred to as a recombination center (Ji et al., 2010). Our finding that long-range interactions between SIS or enhancer elements were most prominent in parts of the Igκ locus with strong CTCF-binding sites suggest a function of CTCF in the control of the positioning of V_κ segments relative to the J_κ-C_κ-enhancer region. However, we demonstrated that iEκ-3'Eκ-mediated interactions and recombination to distal V_κ regions can occur in the absence of CTCF, strongly suggesting that CTCF is largely dispensable for Igκ locus sub-loop formation. It is conceivable that CTCF cooperates in a redundant fashion with lineage-specific factors bound to the κ enhancers (such as E2A, Pax5 or IRF4) for cell-specific regulation of chromatin looping, as previously demonstrated in Th1 cells (Sekimata et al., 2009). CTCF could alternatively function to direct local histone modifications at the Igκ locus (Splinter et al., 2006), and thereby target RAG protein binding. In this context, CTCF may function to epigenetically mark the Igκ locus in specific regions at early stages, after which it is no longer essential for the actual recombination process. Nevertheless, in this model loss of CTCF increases contacts between the J_κ-C_κ-enhancer region and V_κ3 elements, leading to altered recombination events and inappropriate proximal V_κ usage (Figure S6).

Our data demonstrating that CTCF is not essential for enhancer contacts with the majority of V_κ

Figure 7. CTCF Restricts Intronic and 3'κ Enhancer Interactions to the Igκ Light Chain Locus. (A and B) 3C-Seq analysis of long-range interactions with the iEκ (A) or 3'Eκ (B) region in B220+CD19+ small pre-B cell populations purified from V_H81X *Rag1*$^{-/-}$ and V_H81X *Rag1*$^{-/-}$ *mb1*-cre *Ctcf*$^{fl/fl}$ BM and total fetal liver cells. See legend to Figure 6 and Experimental Procedures for details. The graph shows the number of reads per million in 0.1 Mbp intervals in the 3.2 Mb Igκ locus (shaded area) plus 3.0 Mbp upstream and 2.8 Mbp downstream genomic regions. The dashed line represents the viewpoint. 3C-Seq counts obtained for fragments adjacent to the viewpoints were not considered. (C and D) Number of reads per million obtained for each of the 60 BglII fragments that cover the regions indicated in (A) and (B). In each graph, the x-axis shows the location in relation to the viewpoint (in Mb). See also Figure S6.

gene segments is consistent with our previous findings suggesting that CTCF is not absolutely required for TCR-α or TCR-β chain rearrangement in thymocytes (Heath et al., 2008). However, for the Igh chain locus a role for CTCF as a regulator of V(D)J rearrangement through the establishment of higher-order chromatin structures has been hypothesized (Degner et al., 2011, Lucas et al., 2011 and Ebert et al., 2011). This would be supported by several lines of evidence. First, shRNA-mediated CTCF knockdown increases D_H antisense transcription and decreases Igh locus compaction and interactions between D_H and the 3'regulatory region (Degner et al., 2011). Second, CTCF binds to PAIR elements, which are characterized by Pax5-dependent active chromatin, in the distal V_H cluster in pro-B cells (Ebert et al., 2011) and to many proximal V_H segments, remarkably within 200 bp of their RSS sequences (Lucas et al., 2011). Third, CTCF-binding DNase I-hypersensitive sites within the V_H-D_H intergenic region influence antisense transcription and lineage-specific V(D)J recombination (Featherstone et al., 2010 and Giallourakis et al., 2010). However, our findings would not support a model in which CTCF is essential for V(D)J rearrangement, like previously shown for Pax5, Ikaros, or YY1 (Fuxa et al., 2004, Liu et al., 2007 and Reynaud et al., 2008). Previous studies have established that conformational changes in Igh locus topology that localize V_H regions within close proximity of DJ_H elements occur in committed pro-B cells (Jhunjhunwala et al., 2008). Upon virtually complete deletion of CTCF protein at the pro-B cell stage in the *mb1*-cre *Ctcf*$^{fl/fl}$ mice, pre-B cells with productive Igh rearrangements were still generated. Nevertheless, our finding that the introduction of a pre-rearranged Ig H chain transgene partially rescued differentiation of CTCF-deficient pro-B cells would suggest that loss of CTCF reduces the efficiency of Igh chain recombination. Although both proximal and distal V_H segments can be used in the absence of CTCF, it still remains possible that CTCF is required for efficient recombination to particular V_H segments. Furthermore, we cannot formally exclude that even very low levels of CTCF proteins are sufficient to occupy critical sites in the Igh locus at the time of initiation of V(D)J recombination. It is also conceivable that CTCF determines the establishment of a higher-order chromatin structure at early stages in B lymphocyte specification, before the pro-B cell stage, when *mb1*-cre mediated *Ctcf* gene deletion is not complete. This would be supported by the recent finding that CTCF and E2A already bind to the PAIR elements in *Pax5*$^{-/-}$ pro-B cells prior to Igh locus contraction (Ebert et al., 2011). Because CTCF has the ability to influence histone modifications (Splinter et al., 2006), further experiments are required to investigate whether CTCF functions to epigenetically mark the Igh locus and thereby control Igh looping in lymphoid progenitors.

In summary, our study identifies a role for CTCF in directing functional communications between Igκ enhancers and $V_κ$ promoters in the Igκ locus, thereby regulating $V_κ$ gene segment repertoire and restricting the activity of the strong iEκ-3'Eκ enhancers to the Igκ locus.

Experimental Procedures

Mice

Ctcf floxed mice (*Ctcf*$^{fl/fl}$; C57BL/6) (Heath et al., 2008) and *Rag1*$^{-/-}$ (Mombaerts et al., 1992), Eμ-Bcl2 (Strasser et al., 1991), μMT (Kitamura et al., 1991), and V_H81X (Martin et al., 1997) mice have been described previously. Mice were bred and maintained in the Erasmus MC animal care facility under specific pathogen-free conditions and were used at 6–13 weeks of age. Experimental procedures were reviewed and approved by the Erasmus University committee of animal experiments.

Flow Cytometry

Preparation of single-cell suspensions, monoclonal antibodies (mAbs) incubations, fluorescein di-β-D-galactopyranoside loading, *in vivo* BrdU-labeling, and cell cycle analysis have been previously described (Heath et al., 2008, Middendorp et al., 2002 and Ribeiro de Almeida et al., 2009). See Supplemental Experimental Procedures for monoclonal antibodies.

Quantitative RT-PCR Analysis

Total RNA was extracted with the GeneElute mammalian total RNA miniprep system (Sigma-Aldrich) and reverse-transcribed with SuperScript II reverse transcriptase (Invitrogen). For cDNA amplification, Maxima Probe/ROX or SYBR Green/ROX qPCR MasterMix (Fermentas) were used. Primers were designed with the ProbeFinder software (Roche Applied Science) and probes were from the Universal ProbeLibrary (Roche

Applied Science) or designed manually (*Gapdh*) and purchased from Eurogentec (See Supplemental Experimental Procedures). Triplicate reactions were done for each cDNA sample. Gene expression was analyzed with an ABI Prism 7300 Sequence Detector and ABI Prism Sequence Detection Software version 1.4 (Applied Biosystems). Cycle-threshold levels were calculated for each gene and normalized to values obtained for the endogenous reference gene *Gapdh*. For assessment of the purity of the amplified products, standard agarose gel electrophoresis or melting curve analysis were performed.

Immunoblot Analysis

Nuclear extracts were prepared as previously described (Ribeiro de Almeida et al., 2009). In brief, cytoplasmic proteins were extracted on ice for 10 min (in 10 mM/HEPES-KOH [pH = 7.9], 1.5 mM/MgCl2, 10 mM/KCl, 0.5 mM/DTT, 0.2 mM/PMSF). High-salt extraction of nuclear proteins was carried out on ice for 5 min (in 20 mM/HEPES-KOH [pH = 7.9], 25%/glycerol, 420 mM/NaCl, 1.5 mM/MgCl2, 0.2 mM/EDTA, 0.5 mM/DTT, 0.2 mM/PMSF), followed by centrifugation for removal of cellular debris. Blots were probed with anti-CTCF (1:1500, Upstate), anti-Lsd1 (1:2000, Abcam) as an internal loading control, and the secondary antibody swine anti-rabbit-HRP (1:3000, Dako). Primary Ab incubation was done overnight at 4°C in TBS containing 3% nonfat dry milk and 0.05% Tween-20. Signal detection was performed with ECL (Amersham Biosciences).

DNA Microarray Analysis

Biotin-labeled cRNA was hybridized to the Mouse Gene 1.0 ST Array (Affymetrix). Data were analyzed with the BRB-ArrayTools version 3.7.0 software (National Cancer Institute) with Affymetrix CEL files obtained from GCOS (Affymetrix). The RMA approach was used for normalization. Thresholds for selecting important genes were set at a relative difference > 1.75. Changes in gene expression patterns for *mb1*-cre *Ctcf*$^{fl/fl}$ versus WT pro-/pre-B cells were evaluated with Student's t test (with random variance model) and were considered significant with $p<0.001$. Quantitative RT-PCR was performed on selected genes identified by the microarray analysis to verify their expression levels.

V_κ Gene Sequencing, ChIP Sequencing, and 3C Sequencing

See Supplemental Experimental Procedures in the Supplemental Information for details of V_κ gene usage analysis, ChIP-sequencing, and 3C-sequencing.

Statistical Analysis

To analyze statistical significance, we used a two-tailed Student's t test. p values<0.05 were considered significant.

Acknowledgments

We thank E. Hobeika and M. Reth (Max Planck Institute for Immunobiology, Freiburg, Germany) for kindly providing *mb1*-cre mice, Z. Özgür, C.E.M. Kockx (Biomics, Erasmus MC), and M. Pescatori (Bioinformatics, Erasmus MC) for assistance on Affymetrix microarray analysis and Illumina sequencing, T. Langerak (Immunology) for facilitating DNA sequencing, and the Erasmus MC animal care facility staff. This work was partly supported by Fundação para a Ciência e a Tecnologia (C. R. A.), Dutch Cancer Foundation (KWF, R.W.H.), Royal Netherlands Academy of Arts and Sciences (KNAW, R.S.), the Center of Biomedical Genetics and the EuTRACC Consortium (F.G., E.S. and R. S.).

Supplemental Information

Supplemental Experimental Procedures, Supplementary Figures S1-S6 and Supplementary Tables S1 and S2 are available at the Immunity website.

References

S.C. Degner, T.P. Wong, G. Jankevicius, A.J. Feeney
Cutting edge: Developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development

J. Immunol., 182 (2009), pp. 44–48

S.C. Degner, J. Verma-Gaur, T.P. Wong, C. Bossen, G.M. Iverson, A. Torkamani, C. Vettermann, Y.C. Lin, Z. Ju, D. Schulz et al.
CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells
Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 9566–9571

S. Düber, H. Engel, A. Rolink, K. Kretschmer, S. Weiss
Germline transcripts of immunoglobulin light chain variable regions are structurally diverse and differentially expressed
Mol. Immunol., 40 (2003), pp. 509–516

A. Ebert, S. McManus, H. Tagoh, J. Medvedovic, G. Salvagiotto, M. Novatchkova, I. Tamir, A. Sommer, M. Jaritz, M. Busslinger
The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells
Immunity, 34 (2011), pp. 175–187

H. Engel, A. Rolink, S. Weiss
B cells are programmed to activate kappa and lambda for rearrangement at consecutive developmental stages
Eur. J. Immunol., 29 (1999), pp. 2167–2176

K. Featherstone, A.L. Wood, A.J. Bowen, A.E. Corcoran
The mouse immunoglobulin heavy chain V-D intergenic sequence contains insulators that may regulate ordered V(D)J recombination
J. Biol. Chem., 285 (2010), pp. 9327–9338

M. Fuxa, J. Skok, A. Souabni, G. Salvagiotto, E. Roldan, M. Busslinger
Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene
Genes Dev., 18 (2004), pp. 411–422

C.C. Giallourakis, A. Franklin, C. Guo, H.L. Cheng, H.S. Yoon, M. Gallagher, T. Perlot, M. Andzelm, A.J. Murphy, L.E. Macdonald et al.
Elements between the IgH variable (V) and diversity (D) clusters influence antisense transcription and lineage-specific V(D)J recombination
Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 22207–22212

U. Grawunder, A. Rolink, F. Melchers
Induction of sterile transcription from the kappa L chain gene locus in V(D)J recombinase-deficient progenitor B cells
Int. Immunol., 7 (1995), pp. 1915–1925

R.R. Hardy, C.E. Carmack, S.A. Shinton, J.D. Kemp, K. Hayakawa
Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow
J. Exp. Med., 173 (1991), pp. 1213–1225

H. Heath, C. Ribeiro de Almeida, F. Sleutels, G. Dingjan, S. van de Nobelen, I. Jonkers, K.W. Ling, J. Gribnau, R. Renkawitz, F. Grosveld et al.
CTCF regulates cell cycle progression of alphabeta T cells in the thymus
EMBO J., 27 (2008), pp. 2839–2850

R.W. Hendriks, S. Middendorp
The pre-BCR checkpoint as a cell-autonomous proliferation switch
Trends Immunol., 25 (2004), pp. 249–256

S. Herzog, M. Reth, H. Jumaa
Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling
Nat. Rev. Immunol., 9 (2009), pp. 195–205

E. Hobeika, S. Thiemann, B. Storch, H. Jumaa, P.J. Nielsen, R. Pelanda, M. Reth
Testing gene function early in the B cell lineage in mb1-cre mice
Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 13789–13794

M.A. Inlay, T. Lin, H.H. Gao, Y. Xu
Critical roles of the immunoglobulin intronic enhancers in maintaining the sequential rearrangement of IgH and Igk loci
J. Exp. Med., 203 (2006), pp. 1721–1732

S. Jhunjhunwala, M.C. van Zelm, M.M. Peak, S. Cutchin, R. Riblet, J.J. van Dongen, F.G. Grosveld, T.A. Knoch, C. Murre
The 3D structure of the immunoglobulin heavy-chain locus: Implications for long-range genomic interactions
Cell, 133 (2008), pp. 265–279

S. Jhunjhunwala, M.C. van Zelm, M.M. Peak, C. Murre
Chromatin architecture and the generation of antigen receptor diversity
Cell, 138 (2009), pp. 435–448

Y. Ji, W. Resch, E. Corbett, A. Yamane, R. Casellas, D.G. Schatz
The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci
Cell, 141 (2010), pp. 419–431

D. Jung, F.W. Alt
Unraveling V(D)J recombination; insights into gene regulation
Cell, 116 (2004), pp. 299–311

D. Kitamura, J. Roes, R. Kühn, K. Rajewsky
A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin mu chain gene
Nature, 350 (1991), pp. 423–426

Z. Liu, W.T. Garrard
Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3′ boundary sequence spanning 46 kilobases
Mol. Cell. Biol., 25 (2005), pp. 3220–3231

Z. Liu, P. Widlak, Y. Zou, F. Xiao, M. Oh, S. Li, M.Y. Chang, J.W. Shay, W.T. Garrard
A recombination silencer that specifies heterochromatin positioning and ikaros association in the immunoglobulin kappa locus
Immunity, 24 (2006), pp. 405–415

H. Liu, M. Schmidt-Supprian, Y. Shi, E. Hobeika, N. Barteneva, H. Jumaa, R. Pelanda, M. Reth, J. Skok, K. Rajewsky, Y. Shi
Yin Yang 1 is a critical regulator of B-cell development
Genes Dev., 21 (2007), pp. 1179–1189

J.S. Lucas, C. Bossen, C. Murre
Transcription and recombination factories: Common features?
Curr. Opin. Cell Biol., 23 (2011), pp. 318–324

F. Martin, X. Chen, J.F. Kearney
Development of VH81X transgene-bearing B cells in fetus and adult: Sites for expansion and deletion in conventional and CD5/B1 cells
Int. Immunol., 9 (1997), pp. 493–505

S. Middendorp, G.M. Dingjan, R.W. Hendriks
Impaired precursor B cell differentiation in Bruton's tyrosine kinase-deficient mice
J. Immunol., 168 (2002), pp. 2695–2703

P. Mombaerts, J. Iacomini, R.S. Johnson, K. Herrup, S. Tonegawa, V.E. Papaioannou
RAG-1-deficient mice have no mature B and T lymphocytes
Cell, 68 (1992), pp. 869–877

S.L. Nutt, B.L. Kee
The transcriptional regulation of B cell lineage commitment
Immunity, 26 (2007), pp. 715–725

J.E. Phillips, V.G. Corces
CTCF: Master weaver of the genome
Cell, 137 (2009), pp. 1194–1211

D. Reynaud, I.A. Demarco, K.L. Reddy, H. Schjerven, E. Bertolino, Z. Chen, S.T. Smale, S. Winandy, H. Singh
Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros
Nat. Immunol., 9 (2008), pp. 927–936

C. Ribeiro de Almeida, H. Heath, S. Krpic, G.M. Dingjan, J.P. van Hamburg, I. Bergen, S. van de Nobelen, F. Sleutels, F. Grosveld, N. Galjart, R.W. Hendriks
Critical role for the transcription regulator CCCTC-binding factor in the control of Th2 cytokine expression
J. Immunol., 182 (2009), pp. 999–1010

M.S. Schlissel
Regulating antigen-receptor gene assembly
Nat. Rev. Immunol., 3 (2003), pp. 890–899

M.S. Schlissel, D. Baltimore
Activation of immunoglobulin kappa gene rearrangement correlates with induction of germline kappa gene transcription
Cell, 58 (1989), pp. 1001–1007

M. Sekimata, M. Pérez-Melgosa, S.A. Miller, A.S. Weinmann, P.J. Sabo, R. Sandstrom, M.O. Dorschner, J.A. Stamatoyannopoulos, C.B. Wilson
CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus
Immunity, 31 (2009), pp. 551–564

E. Soler, C. Andrieu-Soler, E. de Boer, J.C. Bryne, S. Thongjuea, R. Stadhouders, R.J. Palstra, M. Stevens, C. Kockx, W. van Ijcken et al.
The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation
Genes Dev., 24 (2010), pp. 277–289

E. Splinter, H. Heath, J. Kooren, R.J. Palstra, P. Klous, F. Grosveld, N. Galjart, W. de Laat
CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus
Genes Dev., 20 (2006), pp. 2349–2354

A. Strasser, S. Whittingham, D.L. Vaux, M.L. Bath, J.M. Adams, S. Cory, A.W. Harris
Enforced BCL2 expression in B-lymphoid cells prolongs antibody responses and elicits autoimmune disease
Proc. Natl. Acad. Sci. USA, 88 (1991), pp. 8661–8665

E. ten Boekel, F. Melchers, A.G. Rolink
Changes in the V(H) gene repertoire of developing precursor B lymphocytes in mouse bone marrow mediated by the pre-B cell receptor
Immunity, 7 (1997), pp. 357–368

Y. Xiang, X. Zhou, S.L. Hewitt, J.A. Skok, W.T. Garrard
A multifunctional element in the mouse Igκ locus that specifies repertoire and Ig loci subnuclear location
J. Immunol., 186 (2011), pp. 5356–5366

Chapter 8

Pre-B cell receptor signaling induces immunoglobulin κ locus accessibility by functional redistribution of enhancer-mediated chromatin interactions

Ralph Stadhouders[1], Marjolein J. W. de Bruijn[2], Magdalena B. Rother[3], Saravanan Yuvaraj[2], Claudia Ribeiro de Almeida[2], Petros Kolovos[1], Menno C. Van Zelm[3], Wilfred van IJcken[4], Frank Grosveld[1,5], Eric Soler[1,5,6] & Rudi W. Hendriks[2]†

[1]Department of Cell Biology, Erasmus MC Rotterdam, The Netherlands.
[2]Department of Pulmonary Medicine, Erasmus MC Rotterdam, The Netherlands.
[3]Department of Immunology, Erasmus MC Rotterdam, The Netherlands.
[4]Center for Biomics, Erasmus MC Rotterdam, The Netherlands.
[5]The Cancer Genomics Center, Erasmus MC Rotterdam, The Netherlands.
[6]INSERM UMR967 and French Alternative Energies and Atomic Energy Commission (CEA), Fontenay-aux-Roses, France.

†Corresponding author.

Published in:
PLOS Biology
2014; 12:e1001791

Abstract

During B cell development, the precursor B cell receptor (pre-BCR) checkpoint is thought to increase immunoglobulin κ light chain (Igκ) locus accessibility to the V(D)J recombinase. Accordingly, pre-B cells lacking the pre-BCR signaling molecules Btk or Slp65 showed reduced germline V_κ transcription. To investigate whether pre-BCR signaling modulates V_κ accessibility through enhancer-mediated Igκ locus topology, we performed chromosome conformation capture and sequencing analyses. These revealed that already in pro-B cells the κ enhancers robustly interact with the ~3.2 Mb V_κ region and its flanking sequences. Analyses in wild-type, Btk, and Slp65 single- and double-deficient pre-B cells demonstrated that pre-BCR signaling reduces interactions of both enhancers with Igκ locus flanking sequences and increases interactions of the 3′κ enhancer with V_κ genes. Remarkably, pre-BCR signaling does not significantly affect interactions between the intronic enhancer and V_κ genes, which are already robust in pro-B cells. Both enhancers interact most frequently with highly used V_κ genes, which are often marked by transcription factor E2a. We conclude that the κ enhancers interact with the V_κ region already in pro-B cells and that pre-BCR signaling induces accessibility through a functional redistribution of long-range chromatin interactions within the V_κ region, whereby the two enhancers play distinct roles.

Author Summary

B lymphocyte development involves the generation of a functional antigen receptor, comprising two heavy chains and two light chains arranged in a characteristic "Y" shape. To do this, the receptor genes must first be assembled by ordered genomic recombination events, starting with the immunoglobulin heavy chain (IgH) gene segments. On successful rearrangement, the resulting IgH μ protein is presented on the cell surface as part of a preliminary version of the B cell receptor—the "pre-BCR." Pre-BCR signaling then redirects recombination activity to the immunoglobulin κ light chain gene. The activity of two regulatory κ enhancer elements is known to be crucial for opening up the gene, but it remains largely unknown how the hundred or so Variable (V) segments in the κ locus gain access to the recombination system. Here, we studied a panel of pre-B cells from mice lacking specific signaling molecules, reflecting absent, partial, or complete pre-BCR signaling. We identify gene regulatory changes that are dependent on pre-BCR signaling and occur via long-range chromatin interactions between the κ enhancers and the V segments. Surprisingly the light chain gene initially contracts, but the interactions then become more functionally redistributed when pre-BCR signaling occurs. Interestingly, we find that the two enhancers play distinct roles in the process of coordinating chromatin interactions towards the V segments. Our study combines chromatin conformation techniques with data on transcription factor binding to gain unique insights into the functional role of chromatin dynamics.

Introduction

B lymphocyte development is characterized by stepwise recombination of immunoglobulin (Ig), variable (V), diversity (D), and joining (J) genes, whereby in pro-B cells the Ig heavy (H) chain locus rearranges before the Igκ or Igλ light (L) chain loci [1],[2]. Productive IgH chain rearrangement is monitored by deposition of the IgH μ chain protein on the cell surface, together with the pre-existing surrogate light chain (SLC) proteins λ5 and VpreB, as the pre-B cell receptor (pre-BCR) complex [3]. Pre-BCR expression serves as a checkpoint that monitors for functional IgH chain rearrangement, triggers proliferative expansion, and induces developmental progression of large cycling into small resting Ig μ+ pre-B cells in which the recombination machinery is reactivated for rearrangement of the Igκ or Igλ L chain loci [3],[4].

During the V(D)J recombination process, the spatial organization of large antigen receptor loci is actively remodelled [5]. Overall locus contraction is achieved through long-range chromatin interactions between proximal and distal regions within these loci. This process brings distal V genes in close proximity to (D)J regions, to which Rag (recombination activating gene) protein binding occurs [6] and the nearby regulatory elements that are required for topological organization and recombination [5],[7],[8]. The recombination-associated changes in locus topology thereby provide equal opportunities for individual V genes to be recombined to a (D)J segment. Accessibility and recombination of antigen receptor loci are controlled by many DNA-binding factors that interact with local cis-regulatory elements, such as promoters, enhancers, or silencers [7]–[9]. The long-range chromatin interactions involved in this

process are thought to be crucial for the regulation of V(D)J recombination and orchestrate changes in subnuclear relocation, germline transcription, histone acetylation and/or methylation, DNA demethylation, and compaction of antigen receptor loci [5],[10].

The mouse Igκ locus harbors 101 functional V_κ genes and four functional J_κ elements and is spread over >3 Mb of genomic DNA [11]. Mechanisms regulating the site-specific DNA recombination reactions that create a diverse Igκ repertoire are complex and involve local differences in the accessibility of the V_κ and J_κ genes to the recombinase proteins [12]. Developmental-stage-specific changes in gene accessibility are reflected by germline transcription, which precedes or accompanies gene recombination [13]. In the Igκ locus, germline transcription is initiated from promoters located upstream of J_κ (referred to as κ^0 transcripts) and from V_κ promoters [14]. Deletion of the intronic enhancer (iEκ), located between J_κ and C_κ, or the downstream 3'κ enhancer (3'Eκ), both containing binding sites for the E2a and Irf4/Irf8 transcription factors (TFs), diminishes Igκ locus germline transcription and recombination [15]–[19]. On the other hand, the Sis (silencer in intervening sequence) element in the V_κ–J_κ region negatively regulates Igκ rearrangement [20]. This Sis element was shown to target Igκ alleles to centromeric heterochromatin and to associate with the Ikaros repressor protein that also colocalizes with centromeric heterochromatin. Sis contains a strong binding site for the zinc-finger transcription regulator CTCC-binding factor (Ctcf) [21],[22]. Interestingly, deletion of the Sis element or conditional deletion of the *Ctcf* gene in the B cell lineage both resulted in reduced κ0 germline transcription and enhanced proximal V_κ usage [21],[23]. Very recently, a novel Ctcf binding element located directly upstream of the Sis region was shown to be essential for locus contraction and recombination to distal V_κ genes [23]. In addition, the Igκ repertoire is controlled by the polycomb group protein YY1 [24].

Induction of Igκ rearrangements requires the expression of the Rag1 and Rag2 proteins, the attenuation of the cell cycle, and transcriptional activation of the Igκ locus, all of which are thought to be crucially dependent on pre-BCR signaling [4],[25]. At first, pre-BCR signals synergize with interleukin-7 receptor (IL-7R) signals to drive proliferative expansion of IgH μ+ large pre-B cells [4]. In these cells, transcription of the *Rag* genes is low and the Rag2 protein is unstable due to cell-cycle-dependent degradation [26]. Subsequently, signaling through the pre-BCR downstream adapter Slp65 (SH2-domain-containing leukocyte protein of 65 kDa, also known as Blnk or Bash) switches cell fate from proliferation to differentiation [4]. Importantly, Slp65 (i) induces the TF Aiolos, which down-regulates λ5 expression [27]; (ii) binds Jak3 and thereby interferes with IL-7R signaling [28]; and (iii) reduces inhibitory phosphorylation of Foxo TFs [29]. All these changes result in attenuation of the cell cycle and thus Rag protein stabilization. Moreover, *Rag* gene transcription is induced by Foxo proteins [30].

Although rearrangement and expression of the Igκ locus can occur independently of IgH μ chain expression [31],[32], several lines of evidence indicate that pre-BCR signaling is actively involved in inducing Igκ and Igλ locus accessibility and gene rearrangement. First, surface IgH μ chain expression correlates with germline transcription in the Igκ locus [33]. Second, in the absence of Slp65, κ0 germline transcription is reduced [34]. Third, mice deficient for Bruton's tyrosine kinase (Btk), which is a pre-BCR downstream signaling molecule interacting with Slp65, show reduced Igλ L chain germline transcription and reduced Igλ usage [35]. Fourth, transgenic expression of the constitutively active E41K-Btk mutant in IgH μ chain negative pro-B cells induces premature rearrangement and protein expression of Igκ L chain [34]. Based on fluorescence in situ hybridization (FISH) studies, it has been proposed that in pro-B cells distal V_κ and C_κ genes are separated by large distances and that the Igκ locus specifically undergoes contraction in small pre-B and immature B cells actively undergoing V_κ-J_κ recombination [36]. However, it remains unknown how pre-BCR-induced signals affect the accessibility, contraction, and topology of the V_κ region, or how they affect the long-range interactions of the κ regulatory elements involved in organizing these events.

In this study, we identified the effects of pre-BCR signaling on germline V_κ transcription and on the expression of TFs implicated in the regulation of Igκ gene rearrangement. We found that the decrease in pre-BCR signaling capacity in wild-type, Btk-deficient, Slp65-deficient, and Btk/Slp65 double-deficient pre-B cells was paralleled by a gradient of decreased expression of many TFs including Ikaros, Aiolos, Irf4, and (to a lesser extent) E2a, as well as by a decreased Igκ locus accessibility for recombination. Several of these factors can mediate long-range chromatin interactions and are known to occupy κ regulatory elements that regulate locus accessibility [37]–[40]. We therefore sought to analyze the effect of pre-BCR signaling on the higher order chromatin structure organized by these regulatory sequences at the Igκ locus. To this end, we performed chromosome conformation capture and sequencing (3C-seq) analyses [41] on

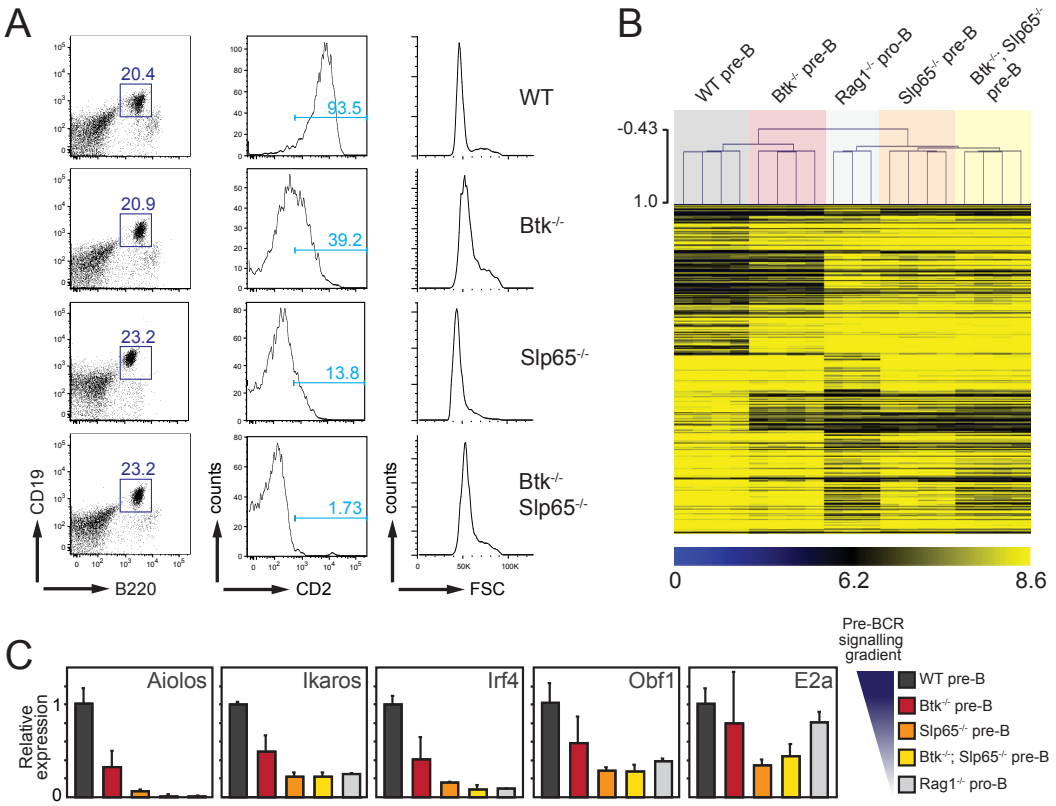


Figure 1. Gene expression profiling strategy for the identification of genes regulated by Btk/Slp65-mediated pre-BCR signaling. (A) FACS sorting strategy for purification of pre-B cell fractions from the indicated mice on a V_H81x transgenic *Rag1*$^{-/-}$ background. Lymphocytes were gated on the basis of forward/side scatter and B220+CD19+ pre-B cell fractions were sorted. Virtually all B220+CD19+ cells were cytoplasmic μ heavy chain positive [34], but showed genotype-dependent levels of expression of the CD2 differentiation marker, in agreement with previous findings [34]. (B) DNA microarray analysis of total mRNA from FACS-purified B220+CD19+ pre-B/pro-B cell fractions from the indicated mice. One-way ANOVA analysis (p=0.01) identified 266 significantly different genes. MeV hierarchical clustering of gene expression differences are represented in the heatmap. (C) Validation of the expression of TFs implicated in Igκ gene rearrangement. Total mRNA isolated from FACS-sorted B220+CD19+ pre-B/pro-B cell fractions from the indicated mice was analyzed by quantitative RT-PCR for expression of TFs. Expression levels were normalized to those of *Gapdh*, whereby the values in WT pre-B cells were set to one. Bars represent mean values and error bars denote standard deviations for four independent mice per group.

pro-B cells and pre-B cells from mice single or double deficient for Btk or Slp65 to evaluate the effects of this pre-BCR signaling gradient on Igκ locus topology. These 3C-seq experiments demonstrated that already in pro-B cells the κ enhancers robustly interact with the ~3.2 Mb $V_κ$ region and its flanking sequences, and that pre-BCR signaling induces accessibility by a functional redistribution of enhancer-mediated chromatin interactions within the $V_κ$ region.

Results

Identification of Genes Regulated by Pre-BCR Signaling

Whereas mice deficient for the pre-BCR signaling molecules Btk and Slp65 have a partial block at the pre-B cell stage [42],[43], in Btk/Slp65 double-deficient mice, only very few pre-B cells show progression to the

Table 1. Genes differentially expressed between WT, Btk, or Slp65 single or double mutant V_H81X Tg *Rag1*$^{-/-}$ pre-B cells or *Rag1*$^{-/-}$ pro-B cells.

ID Probe Set	Accession Number	Gene	Description of Function	*p* Value[a]	Fold Change (Btk KO[b])	Fold Change (Slp65 KO)	Fold Change (BtkSlp65 KO)	Fold Change (Rag1 KO)
Genes known to be up-regulated in signaling-deficient pre-B cells								
10463123	NM_009345	Dntt	N addition VDJ recombination	4.26E-08	8.46	16.99	22.49	39.19
10438064	NM_016982	Vpreb1	VpreB SLC component	7.58E-06	5.52	6.80	6.32	5.73
10438060	ENSMUST00000100136	Igll1	λ5 SLC component	2.17E-04	3.38	3.81	4.17	3.69
10427628	NM_008372	Il7r	IL-7 cytokine receptor	n.s.[c]	1.08	1.62	1.56	1.91
Genes known to be down-regulated in signaling-deficient pre-B cells								
10500677	NM_013486	Cd2	cell adhesion	2.57E-04	−4.82	−5.35	−20.52	−24.26
10469278	NM_008367	Il2ra	IL2 cytokine receptor CD25	1.60E-03	−5.21	−8.79	−15.90	−16.26
10450154	NM_010378	H2-Aa	MHC class II	1.45E-04	−2.04	−5.75	−13.16	−19.80
10562132	NM_001043317	Cd22	Siglec- family receptor	3.32E-05	1.29	1.14	−1.98	−6.06
Transcription regulators and V(D)J recombination								
10390640	NM_011771	Ikzf3	Aiolos DNA binding factor	7.05E-08	−1.66	−3.99	−29.34	−26.09
10384020	NM_017401	Polm	Polumerase mu	2.35E-06	−1.91	−4.17	−10.09	−12.25
10502510	NM_010723	Lmo4	DNA binding factor	1.70E-03	−1.90	−3.56	−5.70	−7.22
10404389	NM_013674	Irf4	DNA binding factor	7.87E-05	−1.60	−2.16	−4.75	−5.29
10438415	ENSMUST00000103752	Igl-V2	Ig V lambda light chain	6.90E-05	−3.41	−3.87	−4.57	−4.81
10438405	M94350	Igl-V1	Ig V lambda light chain	1.58E-06	−3.42	−3.07	−3.98	−6.20
10562812	NM_019866	Spib	Spi-B DNA binding factor	3.88E-04	−1.57	−1.94	−3.16	−3.22
10364559	NM_007880	Arid3a	Bright DNA binding factor	1.10E-03	−1.80	−2.16	−3.04	−3.18
10594001	NM_019689	Arid3b	DNA binding factor	5.74E-04	−1.79	−2.51	−2.91	−3.53
10554370	NM_175433	Zfp710	DNA binding factor	9.03E-03	−1.50	−1.64	−2.14	−2.99
10374333	NM_001025597	Ikzf1	Ikaros DNA binding factor	5.19E-05	−1.31	−1.77	−1.99	−2.44
10560964	NM_011138	Pou2f2	Oct-2 DNA binding factor	6.20E-03	−1.24	−1.30	−1.84	−2.56
10517090	NM_001080819	Arid1a	DNA binding factor	3.60E-03	−1.05	−1.25	−1.21	−2.16
10371662	NM_011461	Sp1c	Pu.1 Dna binding factor	n.s.	−1.04	−1.05	1.13	1.15
10585276	NM_011136	Pou2af1	OBF/OcaB DNA binding factor	n.s.	−1.10	−1.22	−1.22	−1.53
10359770	NM_011137	Pou2f1	DNA binding factor	n.s.	−1.10	−1.22	−1.22	−1.53
10370837	NM_011548	E2a	helix-loop-helix DNA binding factor	n.s.	−1.18	−1.15	−1.37	−1.68
10399691	NM_010496	Id2	inhibitor hlh DNA binding factor	n.s.	−1.39	−2.78	−3.18	−4.01
10509163	NM_008321	Id3	inhibitor hlh DNA binding factor	n.s.	1.46	1.39	1.29	−1.01
10576034	NM_008320	Irf8	DNA binding factor	n.s.	1.36	1.07	−1.07	−1.63
10512669	NM_008782	Pax5	DNA binding factor	n.s.	1.04	−1.22	−1.32	−1.92
10485372	NM_009019	Rag1	V(D)J recombination	n.s.	−1.45	−1.63	−2.03	−1.50
10485370	NM_009020	Rag2	V(D)J recombination	n.s.	−1.63	−1.21	−1.76	−1.72

[a]*p* value in ANOVA analysis.
[b]Fold change times up-regulated or down-regulated (−) when compared with WT (V_H81X Tg *Rag1*$^{-/-}$) pre-B cells.
Groups are Rag1 KO *Rag1*$^{-/-}$ pro-B cells; Btk KO *Btk*$^{-/-}$ V_H81X Tg *Rag1*$^{-/-}$ pre-B cells; Slp65 KO *Slp65*$^{-/-}$ V_H81X Tg *Rag1*$^{-/-}$ pre-B cells; BtkSlp65 KO *Btk*$^{-/-}$*Slp65*$^{-/-}$ V_H81X Tg *Rag1*$^{-/-}$ pre-B cells.
[c]n.s., $p>0.01$.
doi:10.1371/journal.pbio.1001791.t001

immature B cell stage characterized by functional IgL chain gene recombination [44]. To enable analysis of the effects of pre-BCR signaling on (i) the expression of genes involved in Igκ gene rearrangement and on (ii) long-distance chromatin interactions in the Igκ locus in pre-B cells in the absence of Igκ gene recombination events, we bred Btk and Slp65 single- and double-deficient mice on the *Rag1*$^{-/-}$ background. In these mice, progression of B cell progenitors to the pre-B cell stage was conferred by the transgenic, functionally rearranged V_H81x IgH μ chain, which ensures pre-BCR expression and cellular proliferation. The absence of functional Rag1 protein precludes IgL chain gene rearrangement and cells are completely arrested at the small pre-B cell stage (Figure 1A).

We performed genome-wide expression profiling of FACS-purified B220+CD19+ pre-B cell fractions from wild-type (WT), Btk, and Slp65 single- and double-deficient V_H81x transgenic *Rag1*$^{-/-}$ mice (Figure 1A). In these experiments non-V_H81x transgenic *Rag1*$^{-/-}$ pro-B cells served as controls. One-way ANOVA analysis using MeV software ($p<0.01$) [45] revealed that 266 genes were differentially expressed between the five groups of pro-B/pre-B cells (Figure 1B). When compared with WT V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells, 174 genes were up-regulated, whereby the average values of the fold increase were ~1.70, ~3.28, ~3.36, and ~3.47 for *Btk*$^{-/-}$, *Slp65*$^{-/-}$, *Btk*$^{-/-}$*Slp65*$^{-/-}$ V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells and non-V_H81x transgenic *Rag1*$^{-/-}$ pro-B cells, respectively (see Table S1). A similar gradient of gene expression changes was apparent from the average values of the fold change for the 192 significantly down-regulated genes, which were ~1.65, ~2.29, ~3.79, and ~4.15 in the four groups of pre-B/pro-B cells, respectively (see Table S2). In a hierarchical clustering analysis of the five groups of B cell precursors, the expression profiles of *Btk*$^{-/-}$*Slp65*$^{-/-}$ V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells and non-V_H81x transgenic *Rag1*$^{-/-}$ pro-B cells were very similar (Figure 1B). This implies that expression of the 266 genes is not substantially influenced by pre-BCR-mediated proliferation, which is still induced in pre-B cells lacking both Btk and Slp65 [44],[46] but not in *Rag1*$^{-/-}$ pro-B cells. Consistent with these findings, gene distance matrix analysis revealed a clear gene expression gradient among the five groups of pre-B/pro-B cells, in which *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B and *Rag1*$^{-/-}$ pro-B cells again showed highly comparably expression signatures (Figure S1).

In agreement with previous findings [34],[43],[46], pre-BCR signaling-defective pre-B cells

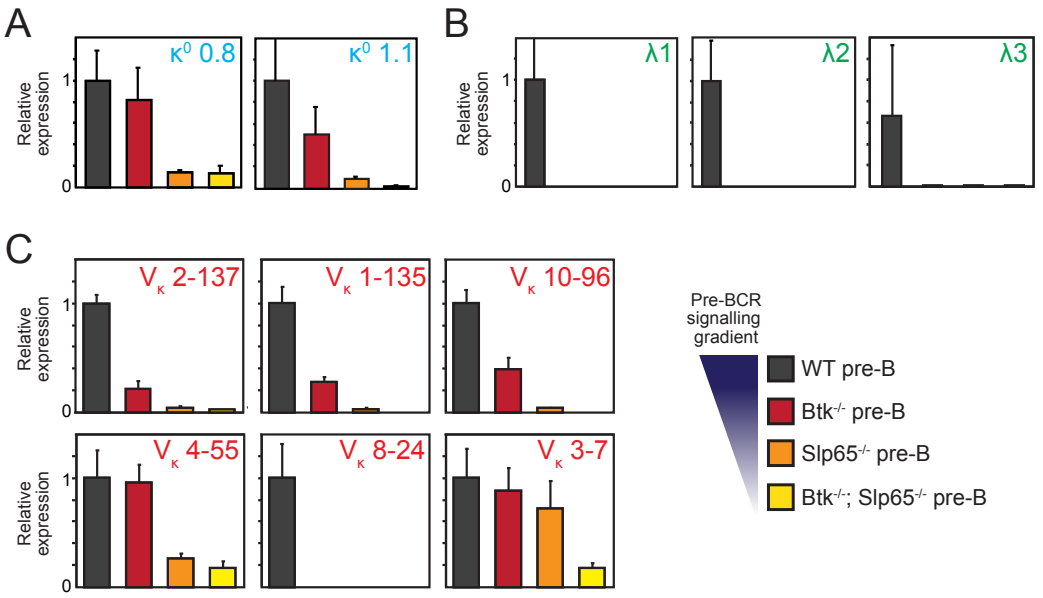


Figure 2. Reduction of Btk/Slp65-mediated pre-BCR signaling is associated with progressive loss of Igκ GLT. Quantitative RT-PCR analysis for κ^0 (A), λ^0 (B), and V_κ GLT (C) of FACS-sorted B220+CD19+ pre-B/pro-B cell fractions from the indicated mice on a V_H81x transgenic *Rag1*$^{-/-}$ background. Expression levels were normalized to those of *Gapdh*, whereby the values in WT pre-B cells were set to one. Bars represent mean values and error bars denote standard deviations for four independent mice per group.

manifested increased expression of *Dntt*, encoding terminal deoxynucleotidyl transferase and the SLC components Vpre (*Vpreb1*) and λ5 (*Igll1*), as well as decreased expression of the cell surface markers Cd2, Cd22, Cd25(IL-2R), and MHC class II (Table 1). Btk and Slp65 single-deficient and particularly double-deficient pre-B cells failed to up-regulate various genes known to be involved in IgL chain recombination, such as *Ikzf3* (Aiolos), *Ikzf1* (Ikaros), *Irf4*, *Spib*, *Pou2f2* (Oct2), polymerase-μ [47], as well as *Hivep1* encoding the Mbp-1 protein, which has been shown to bind to the κ enhancers [48]. In addition, pre-BCR signaling influenced the expression levels of many other DNA-binding or modifying factors that were not previously associated with IgL chain recombination, including *Lmo4*, *Zfp710*, *Arid1a/3a/3b*, the lysine-specific demethylases *Aof1* and *Phf2*, *Prdm2* (a H3K9 methyltransferase), the *sik1* gene encoding a histon deacetylase (HDAC) kinase, *Hdac5*, *Hdac8*, and the DNA repair protein gene *Rev1* (Table S2). We did not find significant differences in the expression of several other TFs implicated in Ig gene recombination—for example, *Obf1*/*Oca-B*, *Pax5*, *E2a*, and *Irf8* (Table 1). In addition, in signaling-deficient pre-B cells, we found reduced transcription of genes encoding several signaling molecules (e.g., *Rasgrp1*, *Rapgefl1*, *Ralgps2*, *Blk*, *Traf5*, *Hck*, *Nfkbia* (IκBα), *Syk*, *Csk*), cell surface markers (Cd38, Cd72, Cd74, Cd55, and Notch2), or genes regulating cell survival (*Bmf* and *Bcl2l1* encoding BclXL) (Table S2). Interestingly, we observed concomitant up-regulation of signaling molecules that are also associated with the T cell receptor (*Lat*, *Zap70*, and *Prkcq* (PKCθ); Table S1).

Next, we used quantitative RT-PCR to confirm the observed differential expression of several TFs. Expression levels of these genes were indeed significantly reduced in a pre-BCR signaling-dependent manner, especially for Aiolos, Ikaros, and Irf4, with residual expression levels in *Btk*$^{-/-}$*Slp65*$^{-/-}$ V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells that were ~1%, ~20%, and ~9% of those observed in WT V_H81x *Rag1*$^{-/-}$ mice, respectively (Figure 1C). In addition, we found moderate effects on Obf1 (Oca-B) and E2a with residual expression levels of ~28% and ~44%, respectively. In chromatin immunoprecipitation (ChIP) assays, we observed in pre-B cells substantial binding of E2a protein to the intronic and 3′ κ enhancer regions and to the three $V_κ$ regions analyzed. Under conditions of reduced pre-BCR signaling activity, E2a binding to the enhancers was essentially maintained (3′Eκ) or reduced (iEκ), but E2a binding to the $V_κ$ regions was lost (Table S3). Consistent with the significant reduction of Ikaros expression in *Slp65*$^{-/-}$ pre-B cells, Ikaros binding to both κ enhancers and $V_κ$ regions was undetectable in these cells (Table S3).

Taken together, from these findings we conclude that the five groups of pro-B/pre-B cells, representing a gradient of progressively diminished pre-BCR signaling, show in parallel a gradient of diminished modulation of many genes that signify pre-B cell differentiation, including key genes implicated in Igκ gene recombination.

Progressively Diminished $V_κ$ and $J_κ$ GLTs in Btk$^{-/-}$*, Slp65*$^{-/-}$*, and Btk*$^{-/-}$*Slp65*$^{-/-}$ *Pre-B Cells*

In these expression profiling studies, we only detected limited differences in germline transcription (GLT) over unrearranged $J_κ$ and $V_κ$ gene segments, which is thought to reflect locus accessibility [12]. However, we previously showed by serial-dilution RT-PCR that the levels of $κ^0$ 0.8 and $κ^0$ 1.1 germline transcripts, which are initiated in different regions 5′ of $J_κ$ and spliced to the $C_κ$ region [49], are apparently normal in *Btk*$^{-/-}$ pre-B cells, modestly reduced in *Slp65*$^{-/-}$ pre-B cells, and severely reduced in *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells [34]. We could confirm these findings for κ0 GLT by quantitative RT-PCR assays on FACS-purified B220+CD19+ pro-B/pre-B cell fractions (Figure 2A). In agreement with our reported findings [34], we also found that *Btk*$^{-/-}$and *Slp65*$^{-/-}$ pre-B cells have defective $λ^0$ transcription, which is initiated 5′ of the $J_λ$ segments (Figure 2B) [49]. GLT across the $V_κ$ region showed a similar pattern of sensitivity to pre-BCR signaling: decreased transcription of six individual $V_κ$ regions tested ($V_κ$3–7, $V_κ$8–24, $V_κ$4–55, $V_κ$10–96, $V_κ$1–35, and $V_κ$2–137) correlated with decreased pre-BCR signaling activity (Figure 2C) in the pre-B cells of the four groups of mice. GLT over unrearranged $V_λ$1 and $V_λ$2 segments was strongly reduced in the absence of Btk or Slp65, as detected by the expression arrays (Table 1).

These observations indicate that Igκ locus accessibility, a hallmark of recombination-competent antigen receptor loci, is progressively reduced under conditions of diminishing pre-BCR signaling.

Pre-BCR Signaling Induces Modulation of Long-Range Chromatin Interactions at the Igκ Locus

Accessibility of antigen receptor loci for V(D)J recombination is thought to be initiated by enhancers, in part through long-range chromatin interactions with promoters of noncoding transcription, resulting in the

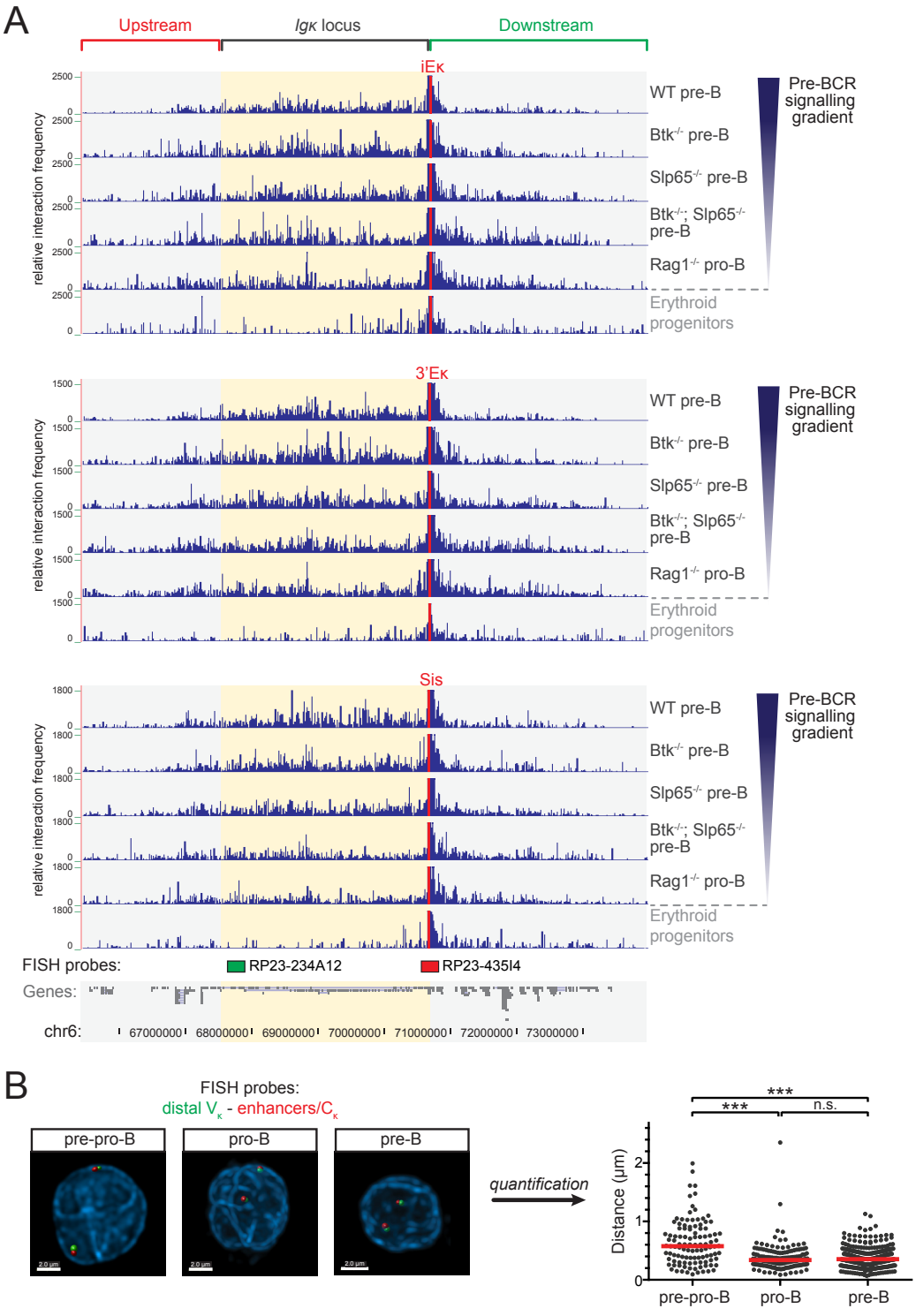


Figure 3. *(Legend at the bottom of the next page)*

activation of germline transcription [8]. Because pre-BCR signaling affects the expression of GLT and various nuclear proteins that mediate long-range chromatin interactions and bind the κ enhancers, it is conceivable that pre-BCR signaling induces changes in the enhancer-mediated higher order chromatin structure of the Igκ locus that facilitates $V_κ$ gene accessibility.

We therefore performed 3C-Seq analyses on FACS-purified B220+CD19+ fractions from the same five groups of mice (WT, *Btk*$^{-/-}$, *Slp65*$^{-/-}$, and *Btk*$^{-/-}$*Slp65*$^{-/-}$ V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells, as well as *Rag1*$^{-/-}$ pro-B cells). Erythroid progenitors were analyzed in parallel as a non-lymphoid control, in which the Igκ locus was not contracted. Genome-wide chromatin interactions were measured for three regulatory elements involved in the control of Igκ locus accessibility and recombination: the iEκ and 3′Eκ enhancers [50]–[52] and the Sis element [20], which contain binding sites for Ikaros/Aiolos, E2a, and Irf4 [16],[17],[20],[38],[53].

In WT pre-B cells, all three regulatory elements showed extensive long-range chromatin interactions within the $V_κ$ region and substantially less interactions with regions up- or downstream of the ~3.2 Mb Igκ domain (Figure 3A; see Figure S2, Figure S3, and Figure S4 for line graphs), confirming previous observations [21]. Under conditions of reduced pre-BCR signaling activity, the three Igκ regulatory elements still showed strong interactions with the $V_κ$ region. Surprisingly, even in the complete absence of pre-BCR signaling in *Rag1*$^{-/-}$ pro-B cells, long-range interactions were still observed at frequencies well above those seen in non-lymphoid cells, suggesting that a contracted Igκ locus topology is not strictly dependent on pre-BCR signaling (Figure 3A, Figure S2, Figure S3, and Figure S4). Next, we used 3D DNA FISH analyses using BAC probes hybridizing to the distal $V_κ$ and $C_κ$/enhancer regions to confirm that Igκ locus contraction was similar in *Rag1*$^{-/-}$ pro-B cells and V_H81x transgenic *Rag1*$^{-/-}$ pre-B cells (both showing a contracted topology, compared with non-contracted pre–pro-B cells deficient for the TF E2a; Figure 3B).

Nevertheless, we did observe that pre-BCR signaling induced clear differences in interaction frequencies. Whereas an increase in pre-BCR signaling was associated with a decrease in the interaction frequencies between the two κ enhancers and regions flanking the Igκ locus (as also revealed by more detailed images of selected regions upstream and downstream of the Igκ domain; see Figure S5), the overall interaction frequency within the Igκ domain appeared unchanged (Figure S3, Figure S4, and Figure S5). Remarkably, interactions with the Sis element showed quite an opposite pattern: pre-BCR signaling correlated with increased overall interactions within the Igκ domain and did not substantially affect interaction frequencies in the Igκ flanking regions (Figure S2 and Figure S5).

Taken together, these analyses show that (i) the Igκ locus is already contracted at the pro-B cell stage and that (ii) pre-BCR signaling induces changes in long-range chromatin interactions, both within the Igκ locus and in the flanking regions.

Pre-BCR Signaling Enhances Interactions of 3′Eκ and Sis, But Not iEκ, with $V_κ$+ Fragments

The differential effects of pre-BCR signaling on long-range chromatin interactions of the iEκ, 3′Eκ, and Sis elements clearly emerged in a quantitative analysis of the 3C-seq datasets (Figure 4A; see Materials

Figure 3. 3C-Seq analysis of long-range chromatin interactions within the Igκ locus and flanking regions. (A) Overview of long-range interactions revealed by 3C-Seq experiments performed on the indicated cell fractions, representing a gradient of pre-BCR signaling, whereby the iEκ element (top), the 3′Eκ element (center), or the Sis element (bottom) was used as a viewpoint. Shown are the relative interaction frequencies (average of two replicate experiments) for the indicated genomic locations. The ~8.4 Mb region containing the Igκ locus (yellow shading) and flanking regions (cyan shading) is shown and genes and genomic coordinates are given (bottom). The locations of the two BAC probes used for 3D DNA-FISH are indicated by a green (distal $V_κ$ probe) and red (proximal $C_κ$/enhancer probe) rectangle (bottom). Pre-B cell fractions were FACS-purified from the indicated mice on a V_H81x transgenic *Rag1*$^{-/-}$ background (see Figure 1 for gating strategy). Erythroid progenitor cells were used as a non-lymphoid control. (B) 3D DNA-FISH analysis comparing locus contraction in cultured bone-marrow–derived *E2a*$^{-/-}$ pre-pro-B, *Rag1*$^{-/-}$ pro-B, and V_H81x *Rag1*$^{-/-}$ pre-B cells (see Figure S6 for phenotype of IL-7 cultured B-lineage cells). Locations of the BAC probes used are indicated at the bottom of panel A. Representative images for all three cell types are shown on the left, quantifications (>100 nuclei counted per cell type) on the right. The red lines indicate the median distance between the two probes. Statistical significance was determined using a Mann–Whitney U test (***$p<0.001$; n.s., not significant, $p\geq0.05$).

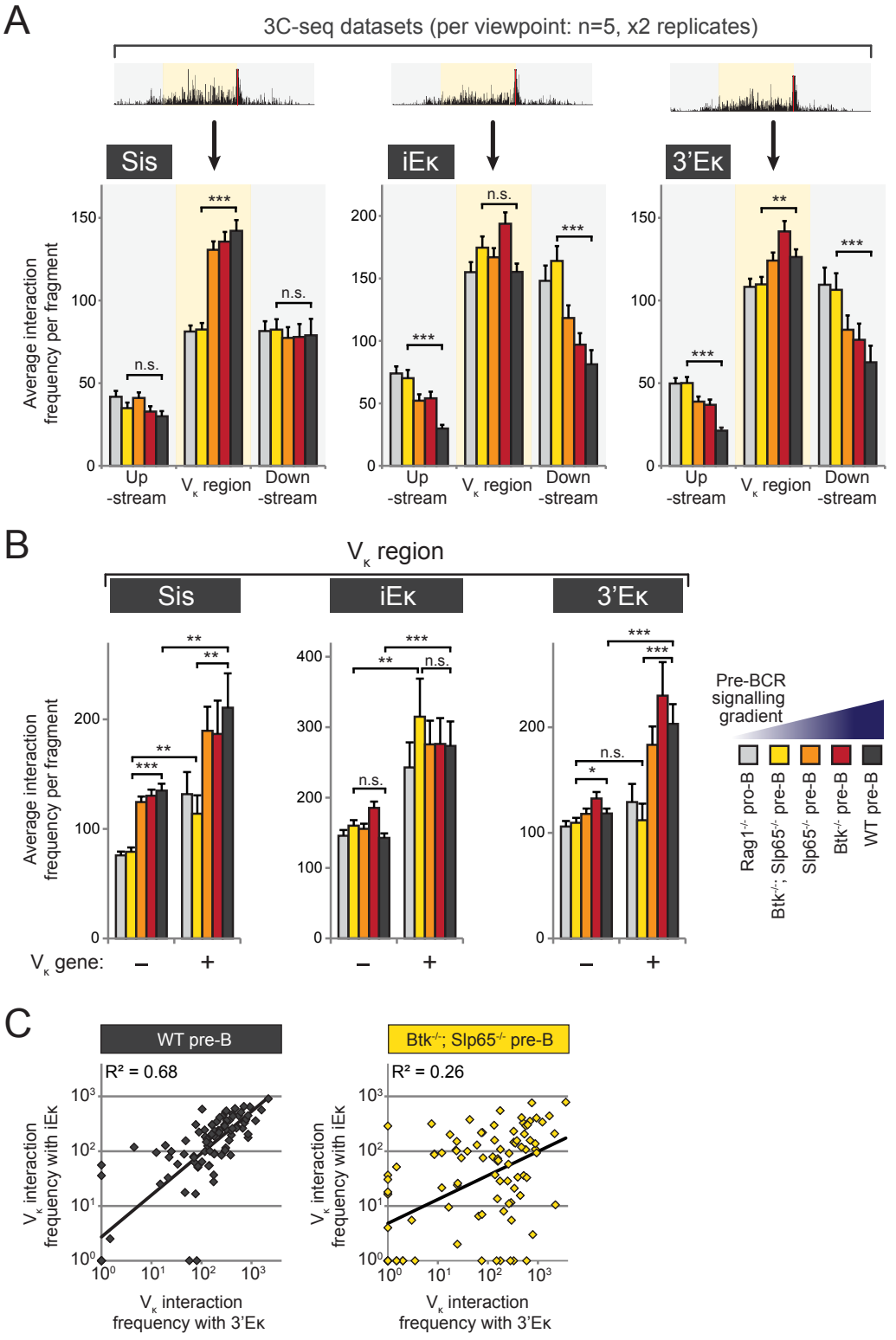


Figure 4. *(Legend at the bottom of the next page)*

and Methods for a detailed description of the quantification methods used). When pre-BCR signaling was absent (*Rag1*$^{-/-}$ pro-B cells) or very low (*Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells), the average interaction frequencies were similar within the ~3.2 Mb V_κ region and the ~3.2 Mb downstream flanking region, for all three regulatory elements. Interaction frequencies with the upstream flanking region were lower, consistent with the larger chromosomal distance to the three viewpoints. The presence of increasing levels of Btk/Slp65-mediated pre-BCR signaling was associated with reduced interaction of iEκ and 3'Eκ with the Igκ flanking regions and with increased interaction of the Sis element and (to a lesser extent) 3'Eκ with the V_κ region (Figure 4A). As a result, for all three regulatory elements pre-BCR signaling resulted in a preference for interaction with fragments inside the V_κ region over fragments outside the V_κ region (Figure S7).

We next focused our analysis on the V_κ region and compared fragments that harbor a functional V_κ gene (V_κ+ fragment) and those that do not (V_κ– fragment). When pre-BCR signaling was absent (*Rag1*$^{-/-}$ pro-B cells) or very low (*Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells), the average interaction frequencies of the Sis or iEκ elements with V_κ+ fragments were higher than with V_κ– fragments. The average interaction frequencies of 3'Eκ with V_κ+ and V_κ– fragments, however, were similar (Figure 4B). Upon pre-BCR signaling, the Sis element showed an increase in interaction frequencies with both V_κ+ and V_κ– fragments, with nevertheless an interaction preference for V_κ+ fragments. In contrast, interaction frequencies between the iEκ element and V_κ+ or V_κ– fragments were not modulated by pre-BCR signaling at all (Figure 4B). The 3'Eκ element exhibited yet another profile: pre-BCR signaling induced increased interaction frequencies specifically with V_κ+ fragments, while interactions with V_κ– fragments were not notably modulated by pre-BCR signaling (Figure 4B). When we separately analyzed non-functional pseudo-V_κ genes, we found for the Sis and 3'Eκ elements that the interaction patterns with functional and non-functional V_κ genes were similar (Figure S8). In contrast, the iEκ enhancer did show an overall increased interaction frequency with V_κ functional genes, compared with non-functional V_κ genes, a phenomenon which was again independent from pre-BCR signaling (Figure S8).

The finding that interactions of V_κ genes with the intronic enhancer are already robust in pro-B cells, while those with the 3'κ enhancer are dependent on pre-BCR signaling, suggested that for individual V_κ genes pre-BCR signaling may result in more similar interaction frequencies with the two enhancers. To investigate this, we examined for all individual V_κ genes the correlation between their 3C-seq interaction frequencies with the iEκ and 3'κ elements and found that these were highly correlated in WT pre-B cells (R^2 = 0.68; Figure 4C). Correlation was severely reduced when pre-BCR signaling was low in *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells (R^2 = 0.26; Figure 4C). Similar pre-BCR signaling-dependent correlations were observed between V_κ-interactions with the Sis element and those with the two enhancers (Figure S9). As the Sis element particularly suppresses recombination of the proximal V_κ3 family, we investigated interaction correlations specifically for this V_κ family. Similar to our findings for all V_κ genes, a subanalysis showed strong correlations for the interactions of V_κ3 family genes with iEκ, 3'κ, and Sis in WT pre-B cells, which were diminished when pre-BCR signaling was low, except for iEκ–Sis correlations, which were pre-BCR signaling-independent (Figure S9).

In summary, we conclude that pre-BCR signaling induces a redistribution of long-range interactions of the iEκ, 3'Eκ, and Sis elements, thereby restricting interactions towards the V_κ gene region. Moreover, upon pre-BCR signaling the long-range interactions mediated by 3'Eκ and Sis—but not those mediated by iEκ—become enriched for fragments harboring a V_κ gene, demonstrating increased proximity

Figure 4. Modulation of long-range chromatin interactions within the Igκ locus by pre-BCR signaling. Quantitative analysis of 3C-Seq datasets using the three indicated κ regulatory elements as viewpoints. (A) Average long-range chromatin interaction frequencies (from two replicate 3C-seq experiments) with upstream (~2.0 Mb), V_κ (~3.2 Mb), and downstream (~3.2 Mb) regions, as defined in Figure 3A, for the five B-cell precursor fractions representing a pre-BCR signaling gradient. Average interaction frequencies per region were calculated as the average number of 3C-Seq reads per restriction fragment within that region. See Materials and Methods section for more details. (B) Average interaction frequencies within the V_κ region were determined for fragments that do not (–) contain a functional V_κ gene and for those that do contain a functional V_κ gene (+). (C) Correlation plots of average interaction frequencies of the two enhancer elements with the 101 functional V_κ genes for WT pre-B cells (left) versus *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells (right). On the log scale, frequencies <1 were set to 10^0. Statistical significance was determined using a Mann–Whitney U test (*p<0.05; **p<0.01; ***p<0.001; n.s., not significant, p≥0.05).

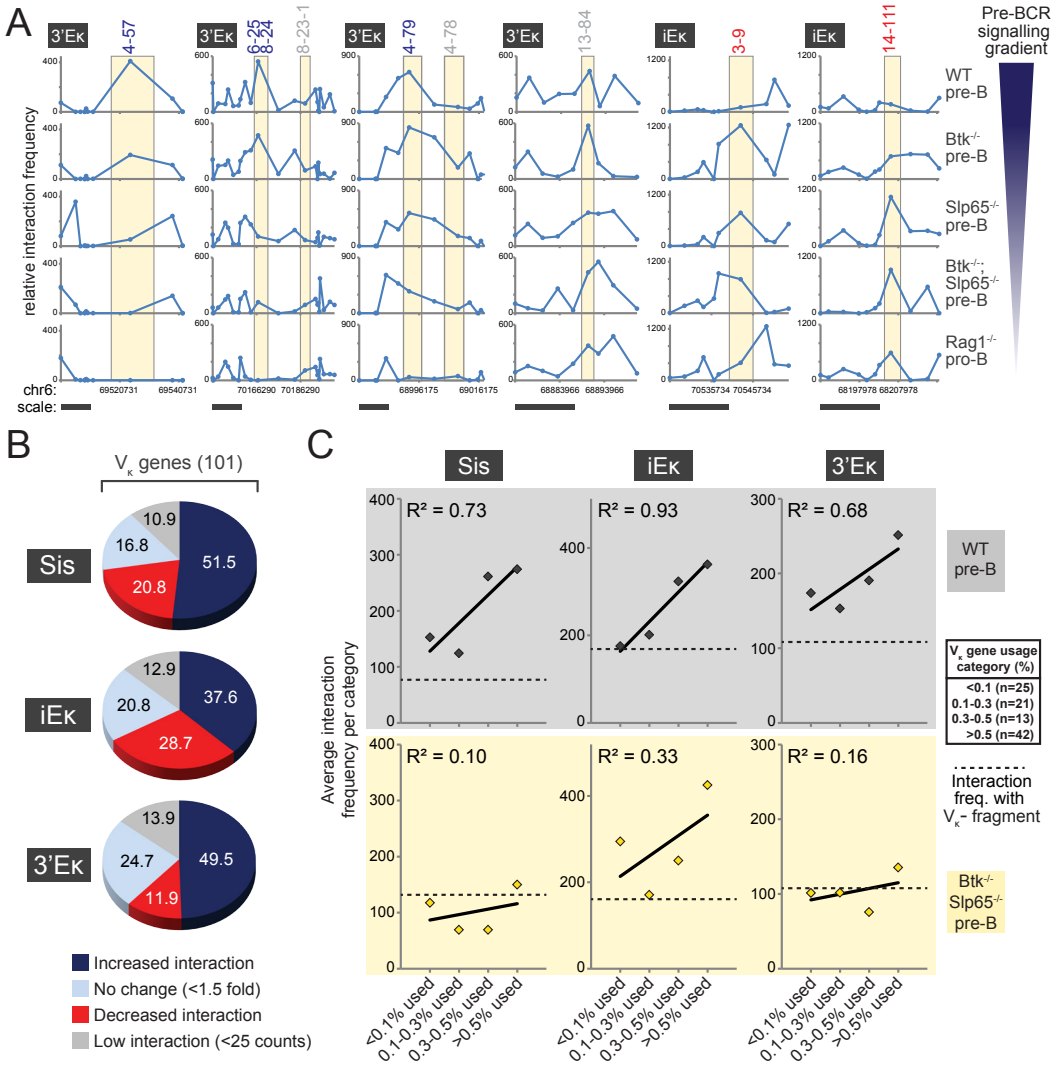


Figure 5. Long-range chromatin interactions of κ regulatory elements correlate with $V_κ$ gene usage. (A) Selected examples of genomic regions containing $V_κ$+ fragments, showing increased ($V_κ$4–57, $V_κ$6–25, $V_κ$8–24, $V_κ$4–79), stable ($V_κ$8–23–1, $V_κ$4–78, $V_κ$13–84), or decreased ($V_κ$3–9, $V_κ$14–111) 3C-seq interaction frequencies with 3'Eκ or iEκ upon pre-BCR signaling. Averaged 3C-seq signals are plotted as a line graph, with the individual data points representing the center of the BglII restriction fragments. Yellow shading marks the BglII fragment on which the $V_κ$ gene(s) is located. $V_κ$ gene(s) are indicated (top) and chromosomal coordinates and scale bars (10 kb) are plotted (bottom). (B) Classification of $V_κ$+ fragments, based on the effect of pre-BCR signaling on their interactions with the three κ regulatory elements indicated. Increase and decrease were defined as >1.5-fold change of interaction frequencies detected in WT pre-B cells versus *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells. (C) Correlation of average interaction frequencies (for the three κ regulatory elements indicated) with four $V_κ$ usage categories ranging from low (<0.1%) to high usage (>0.5%, listed in the table on the right). Diamonds represent average interaction frequencies for *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells (yellow) and WT pre-B cells (grey). The dotted line in each graph depicts the average interaction frequency with fragments that do not contain a functional $V_κ$ ($V_κ$–). Primary $V_κ$ gene usage data were taken from [54].

of 3'Eκ and Sis to $V_κ$ genes. Finally, for individual $V_κ$ genes, the interactions with iEκ, 3'Eκ, and Sis become highly correlated upon pre-BCR signaling, indicating that pre-BCR signals result in regulatory coordination

between these three elements that govern Igκ locus recombination. In contrast, interactions between genes of the proximal $V_κ3$ family, Sis and iEκ—but not 3'κ—appear to be coordinated already in the absence of pre-BCR signaling.

Long-Range Chromatin Interactions of κ Regulatory Elements Correlate with $V_κ$ Usage

Next, we investigated the effects of pre-BCR signaling on the interaction frequencies of individual functional $V_κ$ genes with the three κ regulatory elements (Figure 5A,B). The 3C-seq patterns of the majority (~91%) of the 101 individual $V_κ$+ fragments showed evidence for interaction with one or more of the κ regulatory elements (>25 average counts). When comparing *Btk*$^{-/-}$*Slp65*$^{-/-}$ with WT pre-B cells, we observed that for a large proportion (~38–52%) of $V_κ$+ fragments, interaction frequencies increased upon pre-BCR signaling (Figure 5B). Smaller proportions of $V_κ$+ fragments showed a decrease (~12–29%) or were not significantly affected by pre-BCR signaling (~17–25% with <1.5-fold change). The observed increase or decrease was not related to proximal or distal location of the $V_κ$ genes, nor to their sense or antisense orientation (not shown). Distributions of the three different classes of $V_κ$+ fragments showed substantial differences between the κ regulatory elements. For the Sis and 3'Eκ elements, more $V_κ$+ fragments showed increased than decreased interactions (Figure 5B), in agreement with the signaling-dependent increase in average interaction frequencies of all $V_κ$+ fragments (Figure 4B). In contrast, for the iEκ viewpoint, $V_κ$+ fragments showing increased and decreased interactions were more equal in number, consistent with the limited effects of pre-BCR signaling on overall iEκ interaction frequencies of all $V_κ$+ fragments (Figure 4B).

Although antigen receptor recombination is in principle regarded as a random process, a significant skewing of the primary Igκ repertoire of C57BL/6 mice was recently reported: one third of the $V_κ$ genes was shown to account for >85% of the $V_κ$ segments used by B cells [54]. To assess whether a correlation exists between usage of $V_κ$ genes and their interaction frequencies with κ regulatory elements, we divided the $V_κ$ genes into four usage categories (<0.1%, 0.1–0.3%, 0.3–0.5%, and >0.5%) and calculated their average 3C-Seq interaction frequencies with Sis, iEκ, and 3'κ (Figure 5C). In WT pre-B cells, $V_κ$ usage showed a strong positive correlation with 3C-Seq interaction frequencies for all three regulatory elements (R^2 = ~0.7–0.9; Figure 5C). These correlations were pre-BCR signaling-dependent, since in *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells, they were reduced (for iEκ; R^2 = 0.33) or absent (for Sis and 3'κ; R^2 = 0.10 and R^2 = 0.16, respectively) (Figure 5C).

Collectively, our results indicate that specifically the most frequently used $V_κ$ genes are the main interaction targets of κ regulatory elements, whereby pre-BCR signaling completely underlies this specificity for the Sis and 3'Eκ elements, and to a lesser extent for iEκ.

Long-Range Interactions with κ Regulatory Elements Correlate with the Presence of Ctcf, Ikaros, E2a, and H3K4 Hypermethylation

Next, we investigated whether long-range interactions between κ regulatory elements and the $V_κ$ region correlated with the presence of the TFs Ctcf [21], Ikaros [55], and E2a [56], which have been implicated in Igκ locus recombination [21],[37],[55],[57],[58]. Notably, Ikaros and E2a both strongly bind all three κ regulatory elements, while the Sis element is also occupied by Ctcf ([21]; unpublished data).

Remarkably, we found similar striking correlations between the presence of *in vivo* binding sites for each of these TFs (as determined by ChIP experiments; see Materials and Methods for the relevant references) and long-range chromatin interactions with the κ regulatory elements (Figure 6A–C), even though Ctcf sites are mostly located in between $V_κ$ genes [21] and Ikaros/E2a sites were frequently found close to $V_κ$ gene promoter regions ([2]; Figure 7A). Even when pre-BCR signaling was absent (*Rag1*$^{-/-}$ pro B cells) or very low (*Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells), the average interaction frequencies of the κ regulatory elements with fragments containing Ctcf, Ikaros, or E2a bindings sites were higher than those without binding sites. Irrespective of the presence or absence of bindings sites for these TFs, we found that upon pre-BCR signaling interaction frequencies with the Sis element increased and those with the iEκ did not change. In contrast, for the 3'Eκ we found that pre-BCR signaling specifically increased interaction frequencies with fragments occupied by Ctcf, Ikaros, or E2a. Finally, we found that the presence of di- or trimethylation of histone 3 lysine 4 (H3K4Me2/3), an epigenetic signature associated with locus accessibility [59] and Rag-binding [60],[61], also correlated with increased interaction frequencies with κ regulatory elements, revealing a similar pre-BCR signaling dependency as seen for the TFs analyzed (Figure 6D).

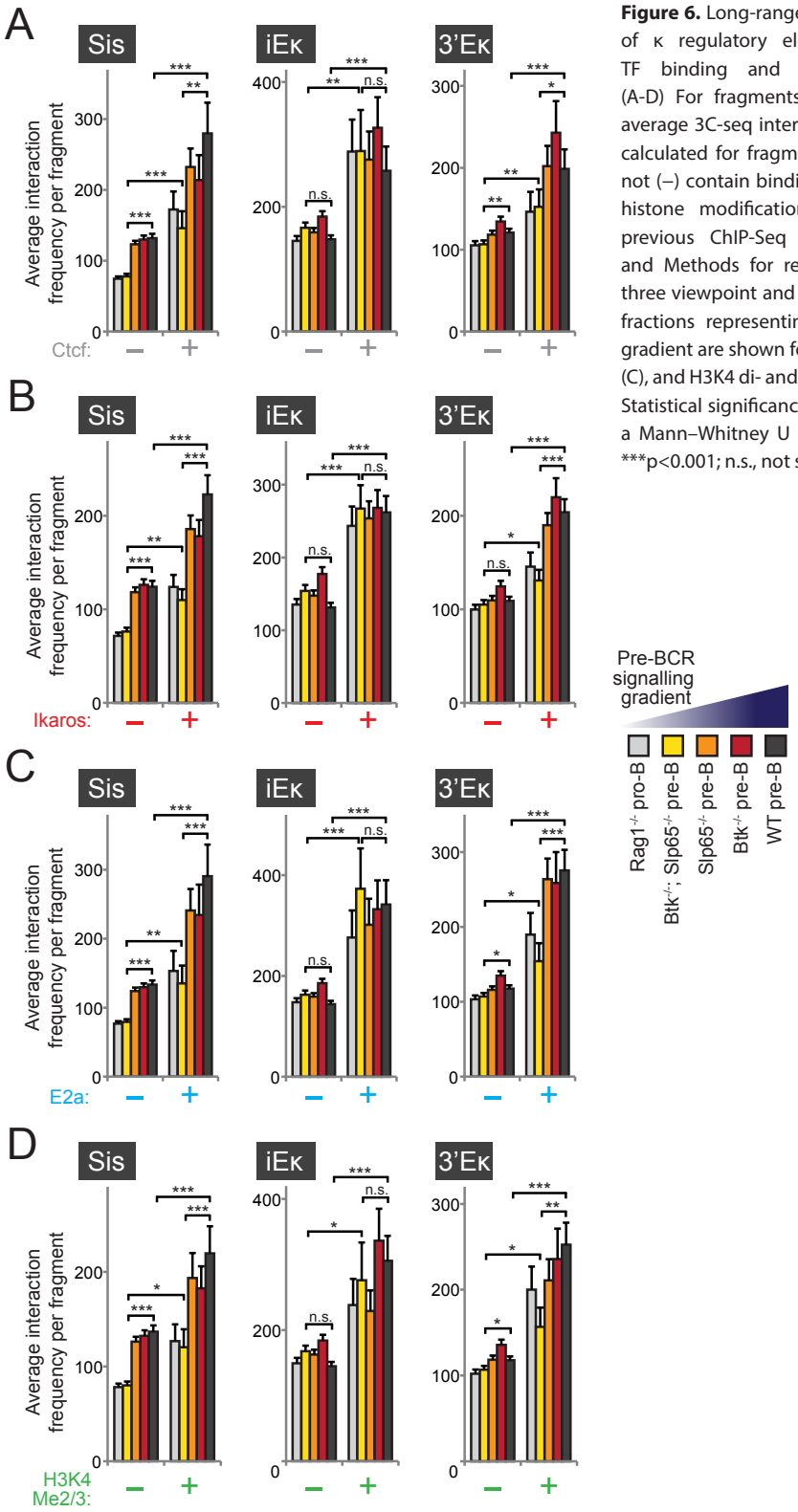


Figure 6. Long-range chromatin interactions of κ regulatory elements correlate with TF binding and histone modifications. (A-D) For fragments within the $V_κ$ region, average 3C-seq interaction frequencies were calculated for fragments that did (+) or did not (–) contain binding sites for TFs or H3K4 histone modifications (as determined by previous ChIP-Seq studies; see Materials and Methods for references). Data for the three viewpoint and the five B-cell precursor fractions representing a pre-BCR signaling gradient are shown for Ctcf (A), Ikaros (B), E2a (C), and H3K4 di- and tri-methylation (Me2/3). Statistical significance was determined using a Mann–Whitney U test (*p<0.05; **p<0.01; ***p<0.001; n.s., not significant, p≥0.05).

We conclude that the presence of essential TFs or H3K4Me2/3 in the V_κ region strongly correlates with the formation of long-range chromatin interactions with the κ regulatory elements, and that for the Sis and 3'Eκ elements this interaction preference is further enhanced by pre-BCR signaling.

Proximity of V_κ Genes to E2a Binding Sites Correlates with High V_κ Usage and Increased Long-Range Chromatin Interactions

Since the long-range interactions with κ regulatory elements correlated with the presence of TFs implicated in Igκ recombination, we next asked whether the κ regulatory elements preferentially interacted with V_κ genes that are in close proximity to binding sites for Ctcf, Ikaros, or E2a.

Strikingly, the majority of functional V_κ genes (95/101) was found to have an Ikaros binding site in close proximity—that is, located on the same 3C-seq restriction fragment (average length of ~3 kb, unpublished data) (Figure 7A). Proximity of V_κ genes to an E2a binding site (37%) or H3K4Me2/3 positive region (~28%) is more selective, while only a small fraction of V_κ genes are close to Ctcf binding sites (~12%) ([22]; Figure 7A). All V_κ genes marked by E2a, Ctcf, H3K4Me2/3, or a combination of these also contain an Ikaros binding site. Frequently used V_κ genes (>1.0% usage; 33/101 genes) were located in two separate regions, a proximal and a distal region, which also contained virtually all E2a and H2K4Me2/3-marked V_κ genes (Figure 7A).

We found that V_κ genes marked by both Ikaros and E2a were used substantially more often than those only bound by Ikaros (Figure 7B), suggesting that these V_κ genes are preferentially targeted for V_κ-to-J_κ gene rearrangement. Our 3C-seq analyses showed that in WT pre-B cells, interaction frequencies with the three κ regulatory elements were higher for Ikaros/E2a-marked V_κ genes compared to genes marked by Ikaros binding alone (Figure 7C). In fact, V_κ+ restriction fragments containing an Ikaros binding site but not an E2a binding site showed interaction frequencies similar to V_κ– restriction fragments. Under conditions of very low pre-BCR signaling (in *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells), we observed strongly reduced interaction frequencies of V_κ+ E2a binding restriction fragments with the Sis and 3'Eκ elements. These interaction frequencies were in the same range as those of V_κ– fragments or V_κ+ fragments that harbored an Ikaros site only (Figure 7C). Interaction frequencies with the iEκ enhancer, however, were independent of pre-BCR signaling. As shown in Figure 7D, for the majority of Ikaros/E2a-marked V_κ+ fragments (65%), pre-BCR signaling was associated with increased interactions with the Sis and 3'Eκ elements (comparing wild-type and *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells). In these analyses, only ~13.5% and ~5.4% of Ikaros/E2a-marked V_κ+ fragments showed a decreased interaction frequency upon pre-BCR signaling. In contrast, almost equal proportions of Ikaros/E2a-marked V_κ+ fragments showed increased (~37%) and decreased (~30%) interactions with iEκ upon pre-BCR signaling.

Taken together, these data reveal strong positive correlations between the presence of E2a binding sites, V_κ usage, and long-range chromatin interactions with κ regulatory elements in pre-B cells. Remarkably, for the iEκ element, these correlations are largely independent of Btk/Slp65-mediated pre-BCR signaling, whereas for the 3'Eκ they are completely dependent on signaling.

Discussion

During B-cell development the pre-BCR checkpoint is known to regulate the expression of many genes, part of which control the increase in Igκ locus accessibility to the V(D)J recombinase complex. However, it remained unknown how pre-BCR signaling events affect accessibility in terms of Igκ locus contraction and topology.

Here we identified numerous genes involved in IgL chain recombination, chromatin modification, signaling, and cell survival to be aberrantly expressed in pre-B cells lacking the pre-BCR signaling molecules Btk and/or Slp65. We found that GLT over the V_κ region, reflecting V_κ accessibility, is strongly reduced in these cells. We used 3C-Seq to show that in pro-B cells both the intronic and the 3' κ enhancers frequently interact with the ~3.2 Mb V_κ region, as well as with Igκ flanking sequences, indicating that the Igκ locus is already contracted at the pro-B cell stage. 3C-Seq analyses in wild-type and Btk/Slp65 single- and double-deficient pre-B cells demonstrated that pre-BCR signaling significantly affects Igκ locus topology. First, pre-BCR signaling reduces the interactions of the intronic and 3'κ enhancers with Igκ flanking regions, effectively focusing enhancer action towards the V_κ region to facilitate V_κ-to-J_κ recombination. Second, pre-BCR signaling strongly increases nuclear proximity of the 3'κ enhancer to V_κ genes, whereby this increase

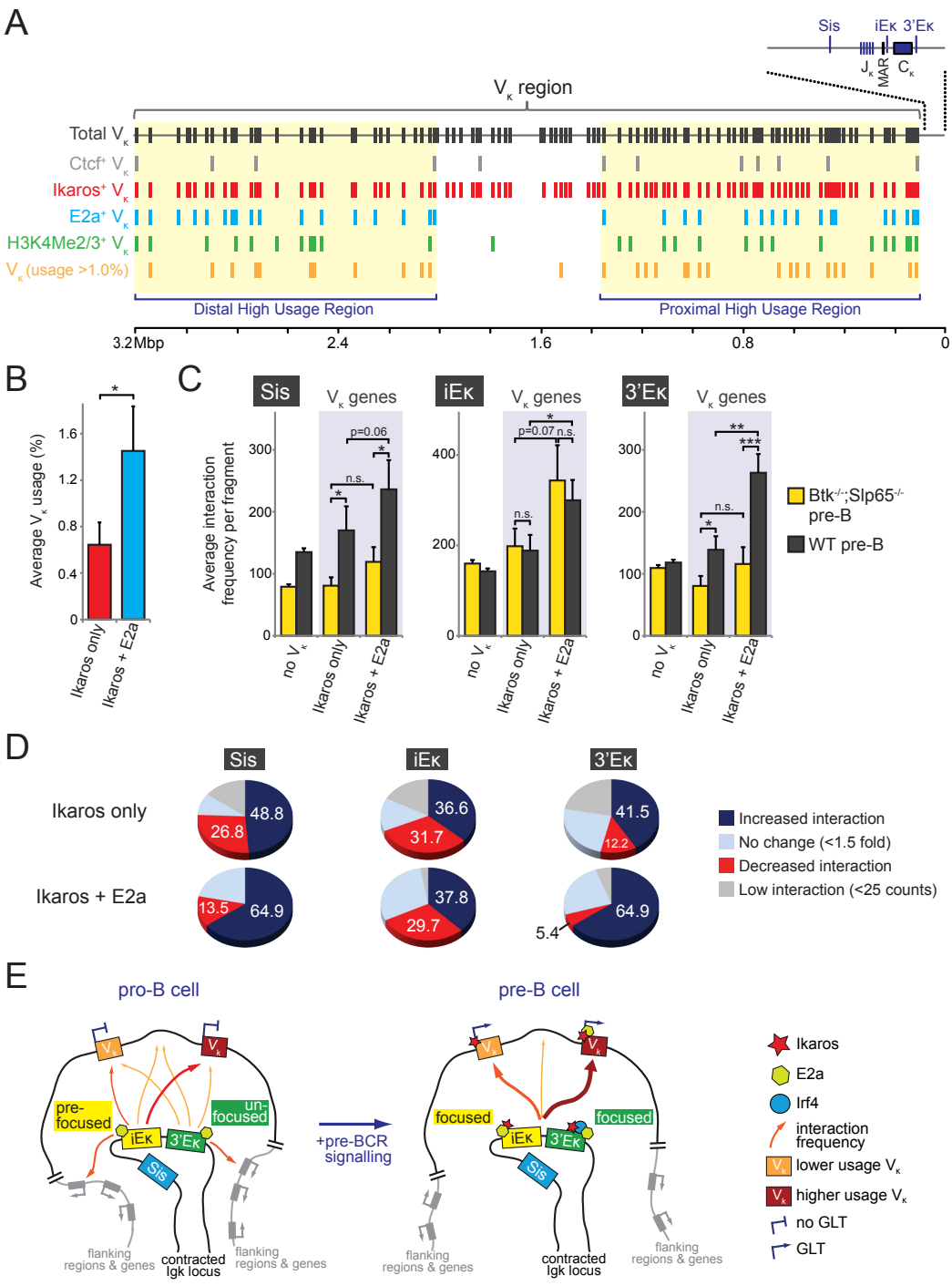


Figure 7. *(Legend at the bottom of the next page)*

is more substantial for more frequently used V_κ genes and for V_κ genes close to a binding site for the basic helix-loop-helix protein E2a. Third, pre-BCR signaling augments interactions between κ regulatory elements

and fragments within the V_κ region bound by the key B-cell TFs Ikaros and E2a and the architectural protein Ctcf. Fourth, pre-BCR signaling has limited effects on interactions of the intronic κ enhancer with fragments within the Igκ locus, as this enhancer already displays interaction specificity for functional V_κ genes and TF-bound regions in pro-B cells. Fifth, pre-BCR signaling has limited effects on the interactions between the intronic or 3'κ enhancers and fragments that do not contain a V_κ gene or an Ikaros, E2a, or Ctcf binding site, emphasizing the specificity of pre-BCR signaling-induced changes in Igκ locus topology. Sixth, pre-BCR signaling appears to induce mutual regulatory coordination between the three regulatory elements, as their interaction profiles with individual V_κ genes become highly correlated upon signaling. Finally, pre-BCR signaling increases interactions of the Sis element with DNA fragments in the Igκ locus, irrespective of the presence of a V_κ gene or TF. Collectively, our findings demonstrate that pre-BCR signals relayed through Btk and Slp65 are required to create a chromatin environment that facilitates proper Igκ locus recombination. This multistep process is initiated by up-regulation of key TFs like Aiolos, Ikaros, Irf4, and E2a. These proteins are then recruited to or further accumulate at the Igκ locus and its regulatory elements, resulting in a specific fine-tuning of enhancer-mediated locus topology that increases locus accessibility to the Rag recombinase proteins.

Importantly, the presence of strong lineage-specific interaction signals between the C_κ/enhancer region and distal V_κ genes in pro-B cells indicates that the Igκ locus is already contracted at this stage. In contrast to a previous microscopy study indicating that Igκ locus contraction did not occur until the small pre-B cell stage [36], our 3D DNA FISH analysis indeed detected similar nuclear distances between distal V_κ and the C_κ/enhancer region in cultured pro-B and pre-B cells. Recently Hi-C was employed to study global early B cell genomic organization whereby substantial interaction frequencies were found between the intronic κ enhancer and the V_κ region in pro-B cells [40]. E2a-deficient pre–pro-B cells, which are not yet fully committed to the B-cell lineage [62], showed very few interactions among the iEκ and the distal part of the V_κ region [40], resembling the interactions we observed in nonlymphoid cells (Figure 3A). Accordingly, 3D-FISH analysis showed that the Igκ locus adopted a noncontracted topology in these pre–pro-B cells (Figure 3B). These data indicate that Igκ locus contraction is already achieved in pro-B cells and depends on the presence of E2a. Supporting this notion, active histone modifications and E2a were already detected at the κ enhancers and V_κ genes at the pro-B cell stage [56],[63], whereby E2a was frequently found at the base of long-range chromatin interactions together with Ctcf and Pu.1, possibly acting as "anchors" to organize genome topology [40]. The observed correlation between E2a binding, V_κ gene usage and iEκ proximity in pro-B cells (Figure 5C, Figure 7C) further strengthens an early critical role for E2a in regulating Igκ locus topology, V_κ gene accessibility, and recombination.

Figure 7. Proximity of V_κ genes to E2a binding sites correlates with frequencies of long-range interactions. (A) Schematic representation of the Igκ locus, showing the location of all functional V_κ (grey, top), J_κ and C_κ gene segments, and the κ regulatory elements Sis, iEκ, and 3'Eκ. MAR, matrix attachment region. V_κ genes within close proximity (as defined by colocalization on the same 3C-Seq restriction fragment) to the indicated TFs or H3K4 hypermethylation (as detected by previous ChIP-seq studies; see Materials and Methods for references) are shown. At the bottom, highly used (>1.0% used) V_κ gene segments are depicted (orange), which cluster within two large high-usage domains (yellow shading). Primary V_κ gene usage data was taken from [54]. (B) Average usage of V_κ genes marked only by an Ikaros binding site or those marked by binding sites of both Ikaros and E2a. (C) Comparison of average interaction frequencies (for the three κ regulatory elements indicated) between V_κ– fragments (no V_κ), V_κ+ fragments containing an Ikaros binding site only, and V_κ+ fragments containing both an Ikaros and E2a binding site. Bars represent average frequencies for *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells (yellow) and WT pre-B cells (grey). (D) Classification of V_κ+ fragments, containing an Ikaros binding site only (top) or containing both an Ikaros and E2a binding site (bottom), based on the effect of pre-BCR signaling on their interactions with the three κ regulatory elements indicated. Increase and decrease were defined as >1.5-fold change of interaction frequencies detected in WT pre-B cells versus *Btk*$^{-/-}$*Slp65*$^{-/-}$ pre-B cells. (E) Proposed model of pre-BCR signaling-mediated changes in κ enhancer action. In pro-B cells (left) the enhancers show minimal coordination and their interactions are not yet (fully) focused on the V_κ genes. Upon pre-BCR signaling and differentiation to pre-B cells (right), TFs bind the locus to coordinate enhancer action and focus their interactions to the V_κ genes, inducing germline transcription (GLT) and accessibility to the V(D)J recombinase. See Discussion for more details. Statistical significance was determined using a Mann–Whitney U test (*p<0.05; **p<0.01; ***p<0.001; n.s., not significant, p≥0.05).

Our 3C-seq experiments revealed that pre-BCR signaling is not required to induce long-range interactions between the κ regulatory elements and distal parts of the $V_κ$ locus, indicating that TFs strongly induced by signaling—that is, Aiolos, Ikaros, and Irf4—are not strictly necessary to form a contracted Igκ locus. Prime candidates for achieving Igκ locus contraction at the pro-B cell stage are E2a and Ctcf, as they have been implicated in regulating Ig locus topology [21],[40],[64],[65] and E2a already marks frequently used $V_κ$ genes at the pro-B cell stage (Figure 7), although we did observe reduced E2a expression and binding to the iEκ enhancer and $V_κ$ genes when pre-B cell signaling was low (Figure 1 and Table S3), suggesting that pre-BCR signaling is required for high-level E2a occupancy of the $V_κ$ genes. We previously reported that Igκ gene recombination can occur in the absence of Ctcf and that Ctcf mainly functions to limit interactions of the κ enhancers with proximal $V_κ$ regions and to prevent inappropriate interactions between these strong enhancers and elements outside the Igκ locus [21]. Because at the pro-to-pre–B cell transition Aiolos, Ikaros, and Irf4 are recruited to the Igκ locus and histone acetylation and H3K4 methylation increases [17],[38],[63],[66], we hypothesize that pre-BCR–induced TFs act upon an E2a/Ctcf-mediated topological scaffold to further refine the long-range chromatin interactions of the κ regulatory elements. Hereby, these TFs mainly act to focus and to coordinate the interactions of the two κ enhancers to the $V_κ$ gene segments, in particular to frequently used $V_κ$ genes, thereby increasing their accessibility for recombination (see Figure 7E for a model of pre-BCR signaling-induced changes in Igκ locus accessibility).

In this context, our 3C-seq data show that the two κ enhancer elements have distinct roles. Both 3′Eκ and iEκ elements manifest interaction specificity for highly used, E2a-marked, $V_κ$ genes. However, whereas iEκ already shows this specificity in pro-B cells (although pre-BCR signaling does augment this specificity), 3′Eκ only does so in pre-B cells upon pre-BCR signaling. These observations indicate that iEκ is already "prefocused" at the pro-B cell stage and that pre-BCR signals are required to fully activate and focus the 3′Eκ to allow synergistic promotion of Igκ recombination by both enhancers (see Figure 7E) [52]. In agreement with such distinct sequential roles, iEκ and not the 3′Eκ was found to be required for the initial increase in Igκ locus accessibility, which occurred upon binding of E2a only [37],[38],[67]. The 3′Eκ on the other hand requires binding of pre-BCR signaling-induced Irf4 to promote locus accessibility [19],[38], followed by further recruitment of E2a to both κ enhancers and highly used $V_κ$ genes (Table S3 and [38],[57]).

The Sis regulatory element was shown to dampen proximal $V_κ$–$J_κ$ rearrangements and to specify the targeting of Igκ transgenes to centromeric heterochromatin in pre-B cells [20]. As Sis is extensively occupied by the architectural Ctcf protein and deletion of Sis or Ctcf both resulted in increased proximal $V_κ$ usage [21],[23], it was postulated that Sis functions as a barrier element to prevent the κ enhancers from too frequently targeting proximal $V_κ$ genes for recombination. In this context, we now provide evidence that interactions between the proximal $V_κ$ genes, Sis, and iEκ—but not 3′κ—are already coordinated before pre-BCR signaling occurs (Figure S9). Perhaps not surprisingly, Sis-mediated long-range chromatin interactions displayed a pattern and pre-BCR signaling response that was different from the κ enhancers. Unlike for the enhancers, upon pre-BCR signaling, Sis-mediated interactions with regions outside the Igκ locus were maintained and interaction within the $V_κ$ region increased, irrespective of the presence of $V_κ$ genes or TF binding sites. Because Sis is involved in targeting the non-recombining Igκ allele to heterochromatin [20], the observed interaction pattern of the Sis element might reflect its action in pre-B cells to sequester the non-recombining Igκ locus and target it towards heterochromatin. This might also explain the increased interaction frequencies of Sis with highly used $V_κ$ genes upon pre-BCR signaling (Figures 5C and 7C), as such highly accessible genes likely require an even tighter association with Sis and heterochromatin to prevent undue recombination.

Surprisingly, we observed a striking correlation between Ikaros binding and $V_κ$ gene location (94% of $V_κ$ genes were in close proximity to an Ikaros binding site; Figure 7A). Although Ikaros and Aiolos have a positive role in regulating gene expression during B-cell development [55],[58] and Ikaros is required for IgH and IgL recombination [39],[58], Ikaros has also been reported to silence gene expression through its association with pericentromeric heterochromatin [68] or through recruitment of repressive cofactor complexes [69],[70]. Recruitment of Ikaros to the Igκ locus was found increased in pre-B cells as compared to pro-B cells [63], in agreement with its up-regulation in pre-B cells (Figure 1). Furthermore, Ikaros binds the Sis element, where it was suggested to mediate heterochromatin targeting of Igκ alleles by the Sis region [20]. Aiolos, although not essential for B-cell development like Ikaros [58],[71], is strongly induced by pre-B cell signaling and has been reported to cooperate with Ikaros in regulation gene expression [27]. Although their synergistic role during IgL chain recombination has not been extensively studied, the Ikaros/Aiolos

ratio changes upon pre-BCR signaling (Figure 1). Increased recruitment of Ikaros/Aiolos to V_κ genes and the κ enhancers likely increases Igκ locus accessibility and contraction (see Figure 6), as Ikaros was very recently shown to be essential for IgL recombination [58]. On the other hand, it is conceivable that on the non-recombining allele, increased recruitment of Ikaros/Aiolos to V_κ genes and the Sis region could facilitate silencing of this allele. Further investigations using allele-specific approaches [72] will be required to clarify the allele-specific action of the Sis element during Igκ recombination.

In summary, by investigating the effects of a pre-BCR signaling gradient—rather than deleting individual TFs—we have taken a more integrative approach to study the regulation of Igκ locus topology. Our 3C-Seq analyses in wild-type, Btk, and Slp65 single- and double-deficient pre-B cells show that interaction frequencies between Sis, iEκ, or 3'Eκ and the V_κ region are already high in pro-B cells and that pre-BCR signaling induces accessibility through a functional redistribution of long-range chromatin interactions within the V_κ region, whereby the iEκ and 3'Eκ enhancer elements play distinct roles.

Materials and Methods

Mice

V_H81x transgenic mice [73] on the *Rag1*$^{-/-}$ background [74] that were either wild-type, *Btk*$^{-/-}$ [75], *Slp65*$^{-/-}$ [42], or *Btk*$^{-/-}$*Slp65*$^{-/-}$ have been previously described [34]. Mice were crossed on the C57BL/6 background for >8 generations, bred, and maintained in the Erasmus MC animal care facility under specific pathogen-free conditions and were used at 6–13 wk of age. Experimental procedures were reviewed and approved by the Erasmus University Committee of Animal Experiments.

Flow Cytometry

Preparation of single-cell suspensions and incubations with monoclonal antibodies (mAbs) were performed using standard procedures. Bone marrow B-lineage cells were purified using fluorescein isothiocyanate (FITC)-conjugated anti-B220(RA3-6B2) and peridinin chlorophyll protein (PCP)-conjugated anti-CD19, together with biotinylated mAbs specific for lineage markers Gr-1, Ter119, and CD11b and APC-conjugated streptavidin as a second step to further exclude non-B cells. Cells were sorted with a FACSARia (BD Biosciences). The following mAbs were used for flow cytometry: FITC-, PerCP–anti-B220 (RA3-6B2), phycoerythrin (PE)–anti-CD2 (LFA-2), PCP-, allophycocyanin (APC)- or APC–Cy7–anti-CD19 (ID3), PE-, or APC anti-CD43 (S7). All these antibodies were purchased from BD Biosciences or eBiosciences. Samples were acquired on an LSRII flow cytometer (BD Biosciences) and analyzed with FlowJo (Tree Star) and FACSDiva (BD Biosciences) software.

Quantitative RT-PCR and DNA Microarray Analysis

Extraction of total RNA, reverse-transcription procedures, design of primers, and cDNA amplification have been described previously [21]. Gene expression was analyzed using an ABI Prism 7300 Sequence Detector and ABI Prism Sequence Detection Software version 1.4 (Applied Biosystems). All PCR primers used for quantitative RT-PCR of TFs or κ^0, λ^0, and V_κ GLT are described in [21], except for *Obf1* (forward 5'-CCTGGCCACCTACAGCAC-3', reverse 5'-GTGGAAGCAGAAA CCTCCAT-3', obtained from the Roche Universal Probe Library).

Biotin-labeled cRNA was hybridized to the Mouse Gene 1.0 ST Array according to the manufacturer's instructions (Affymetrix); data were analyzed with BRB-ArrayTools (version 3.7.0, National Cancer Institute) using Affymetrix CEL files obtained from GCOS (Affymetrix). The RMA approach was used for normalization. The TIGR MultiExperiment Viewer software package (MeV version 4.8.1) was used to perform data analysis and visualize results [45]. One-way ANOVA analysis of the five experimental groups of B cells was used to identify genes significantly different from wild-type V_H81X Tg *Rag1*$^{-/-}$ pre-B cells ($p<0.01$).

Chromatin Immunoprecipitation (ChIP)

ChIP experiments were performed as previously described [76] using FACS sorted bone marrow pre-B cell

fractions (0.3–2.0 million cells per ChIP). Antibodies against E2a (sc-349, Santa Cruz Biotechnology) and Ikaros (sc-9861, Santa Cruz Biotechnology) were used for immunoprecipitation. Purified DNA was analyzed by quantitative RT-PCR as described above. Primer sequences are available on request.

Chromosome Conformation Capture Coupled to High-Throughput Sequencing (3C-Seq)

3C-Seq experiments were essentially carried out as described previously [21],[41]. For 3C-Seq library preparation, BglII was used as the primary restriction enzyme and NlaIII as a secondary restriction enzyme. 3C-seq template was prepared from WT E13.5 fetal liver erythroid progenitors and FACS-sorted bone marrow pro-B cell or pre-B cell fractions (see above) from pools of 4–6 mice. In total, between 1 and 8 million cells were used for 3C-seq analysis. Primers for the Sis, iEκ, and 3'Eκ viewpoint-specific inverse PCR were described previously [21]. 3C-seq libraries were sequenced on an Illumina Hi-Seq 2000 platform. 3C-Seq data processing was performed as described elsewhere [41],[77]. Two replicate experiments were sequenced for each genotype and viewpoint, and normalized interaction frequencies per BglII restriction fragment were averaged between the two experiments.

For quantitative analysis, the Igκ locus and surrounding sequences were divided into three parts (mm9 genome build): a ~2 Mb upstream region (chr6:65,441,978–67,443,029; 759 fragments), a ~3.2 Mb V_κ region (chr6:67,443,034–70,801,754; 1,290 fragments) and a downstream ~3.2 Mb region (chr6:70,801,759–73,993,074; 1,143 fragments). For each cell type (as described above) sequence read counts within individual BglII restriction fragments were normalized for differences in library size (expressed as "reads per million"; see [74]) and averaged between the two replicates before further use in the various calculations. Very small BglII fragments (<100 bp) were excluded from the analysis. Fragments in the immediate vicinity of the regulatory elements (chr6:70,659,392–70,693,183; 10 fragments) were also excluded because of high levels of noise around the viewpoint, a characteristic of all 3C-based experiments. V_κ gene coordinates (both functional genes and pseudogenes) were obtained from IMGT [11] and NCBI (Gene ID: 243469) databases. V_κ gene usage data (C57BL/6 strain, bone marrow) were obtained from [54]. ChIP-seq datasets were obtained from [21] (Ctcf), [55] (Ikaros), and [56] (E2a, H3K4Me2, and H3K4Me3). V_κ genes were scored positive for TF binding sites or for a histone modification, if they were located on the same BglII restriction fragment (corresponding to the 3C-Seq analysis).

3D DNA Immuno-FISH

Rag1$^{-/-}$ pro-B and *Rag1*$^{-/-}$;V_H81X pre-B cells were isolated from femoral bone marrow suspensions by positive enrichment of CD19+ cells using magnetic separation (Miltenyi Biotec). Cells were cultured for 2 wk in Iscove's Modified Dulbecco's medium containing 10% fetal calf serum, 200 U/ml penicillin, 200 mg/ml streptomycin, 4 nM L-glutamine, and 50 μM β-mercaptoethanol, supplemented with IL-7 and stem cell factor at 2 ng/ml. *E2a*$^{-/-}$ hematopoietic progenitors were grown as described previously [78]. Prior to 3D-FISH analysis, cells were characterized by flow cytometric analysis of CD43, CD19, and CD2 surface marker expression to verify their phenotype (Figure S6).

3D DNA FISH was performed as described previously [79] with BAC clones RP23-234A12 and RP23-435I4 (located at the distal end of the V_κ region and at the C_κ/enhancer region, respectively; Figure 3A) obtained from BACPAC Resources (Oakland, CA). Probes were directly labeled with Chromatide Alexa Fluor 488-5 dUTP and Chromatide Alexa Fluor 568-5 dUTP (Invitrogen) using Nick Translation Mix (Roche Diagnostics GmbH).

Cultured primary cells were fixed in 4% paraformaldehyde, and permeabilized in a PBS/0.1% Triton X-100/0.1% saponin solution and subjected to liquid nitrogen immersion following incubation in PBS with 20% glycerol. The nuclear membranes were permeabilized in PBS/0.5% Triton X-100/0.5% saponin prior to hybridization with the DNA probe cocktail. Coverslips were sealed and incubated for 48 h at 37°C, washed, and mounted on slides with 10 μl of Prolong gold anti-fade reagent (Invitrogen).

Pictures were captured with a Leica SP5 confocal microscope (Leica Microsystems). Using a 63× lens (NA 1.4),

we acquired images of ~70 serial optical sections spaced by 0.15 µm. The datasets were deconvolved and analyzed with Huygens Professional software (Scientific Volume Imaging, Hilversum, the Netherlands). The 3D coordinates of the center of mass of each probe were transferred to Microsoft Excel, and the distances separating each probe were calculated using the equation: √(Xa−Xb)2+(Ya−Yb)2+(Za−Zb)2, where X, Y, and Z are the coordinates of object a or b.

Statistical Analysis

Statistical significance was analyzed using a nonparametric Mann–Whitney U test (IBM SPSS Statistics 20). The p values<0.05 were considered significant.

Accession Numbers

3C-seq and microarray expression datasets have been submitted to the Sequence Read Archive (SRA, accession number SRP032509) and Gene Expression Omnibus (GEO, accession number GSE53896), respectively.

Supporting Information

Supplementary Figures S1-S9 and Supplementary Tables S1-S3 are available at the PLOS Biology website.

Acknowledgments

We thank H. Jumaa (Ulm, Germany), J. Kearney (Birmingham, AL), and C. Murre (San Diego, CA) for kindly providing *Slp65*$^{-/-}$, V_H81x transgenic, and *E2a*$^{-/-}$ mice, respectively. We thank D. Nemazee (La Jolla, CA) for providing detailed V_κ usage data. We also thank Z. Özgür, C.E.M. Kockx, Rutger Brouwer, and Mirjam van den Hout (Biomics, Erasmus MC), M. Pescatori (Bioinformatics, Erasmus MC), and P.F. van Loo, I. Bergen, and V. Ta (Pulmonary Medicine, Erasmus MC) for their contributions.

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: RS MVZ ES FG RWH. Performed the experiments: RS MJWdB MBR CRdA. Analyzed the data: RS RWH MBR SY WvIJ PK. Wrote the paper: RS RWH.

Funding Statement

This work was partly supported by Fundação para a Ciência e a Tecnologia (to CRA), the International Association for Cancer Research (AICR 10-0562, to RWH), EpiGenSys/ERASysBio +/FP7 (NL: NWO, UK: BSRC, D: BMBF, to PK), the Center of Biomedical Genetics and the EU 6th Framework Programme EuTRACC Consortium (Project LSHG-CT-2007-037455; FG, ES, and RS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Jung D, Giallourakis C, Mostoslavsky R, Alt FW (2006) Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. Annu Rev Immunol 24: 541–570.
2. Bossen C, Mansson R, Murre C (2012) Chromatin topology and the regulation of antigen receptor assembly. Annu Rev Immunol 30: 337–356.
3. Herzog S, Reth M, Jumaa H (2009) Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. Nat Rev Immunol 9: 195–205.
4. Hendriks RW, Middendorp S (2004) The pre-BCR checkpoint as a cell-autonomous proliferation switch. Trends Immunol 25: 249–256.
5. Jhunjhunwala S, van Zelm MC, Peak MM, Murre C (2009) Chromatin architecture and the generation of antigen receptor diversity. Cell 138: 435–448.
6. Ji Y, Resch W, Corbett E, Yamane A, Casellas R, et al. (2010) The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell 141: 419–431.
7. Perlot T, Alt FW (2008) Cis-regulatory elements and epigenetic changes control genomic rearrangements of the IgH locus. Adv Immunol 99: 1–32.
8. Cobb RM, Oestreich KJ, Osipovich OA, Oltz EM (2006) Accessibility control of V(D)J recombination. Adv Immunol 91: 45–109.
9. Oestreich KJ, Cobb RM, Pierce S, Chen J, Ferrier P, et al. (2006) Regulation of TCRbeta gene assembly by a promoter/enhancer holocomplex. Immunity 24: 381–391.
10. Seitan VC, Krangel MS, Merkenschlager M (2012) Cohesin, CTCF and lymphocyte antigen receptor locus rearrangement. Trends Immunol 33: 153–159.
11. Lefranc MP, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, et al. (2005) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 33: D593–597.

12. Yancopoulos GD, Alt FW (1985) Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell 40: 271–281.
13. Abarrategui I, Krangel MS (2009) Germline transcription: a key regulator of accessibility and recombination. Adv Exp Med Biol 650: 93–102.
14. Schlissel MS, Baltimore D (1989) Activation of immunoglobulin kappa gene rearrangement correlates with induction of germline kappa gene transcription. Cell 58: 1001–1007.
15. Murre C (2005) Helix-loop-helix proteins and lymphocyte development. Nat Immunol 6: 1079–1086.
16. Muljo SA, Schlissel MS (2003) A small molecule Abl kinase inhibitor induces differentiation of Abelson virus-transformed pre-B cell lines. Nat Immunol 4: 31–37.
17. Lu R, Medina KL, Lancki DW, Singh H (2003) IRF-4,8 orchestrate the pre-B-to-B transition in lymphocyte development. Genes Dev 17: 1703–1708.
18. Ma S, Turetsky A, Trinh L, Lu R (2006) IFN regulatory factor 4 and 8 promote Ig light chain kappa locus activation in pre-B cell development. J Immunol 177: 7898–7904.
19. Johnson K, Hashimshony T, Sawai CM, Pongubala JM, Skok JA, et al. (2008) Regulation of immunoglobulin light-chain recombination by the transcription factor IRF-4 and the attenuation of interleukin-7 signaling. Immunity 28: 335–345.
20. Liu Z, Widlak P, Zou Y, Xiao F, Oh M, et al. (2006) A recombination silencer that specifies heterochromatin positioning and ikaros association in the immunoglobulin kappa locus. Immunity 24: 405–415.
21. Ribeiro de Almeida C, Stadhouders R, de Bruijn MJ, Bergen IM, Thongjuea S, et al. (2011) The DNA-binding protein CTCF limits proximal Vkappa recombination and restricts kappa enhancer interactions to the immunoglobulin kappa light chain locus. Immunity 35: 501–513.
22. Ribeiro de Almeida C, Stadhouders R, Thongjuea S, Soler E, Hendriks RW (2012) DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation. Blood 119: 6209–6218.
23. Xiang Y, Zhou X, Hewitt SL, Skok JA, Garrard WT (2011) A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. J Immunol 186: 5356–5366.
24. Pan X, Papasani M, Hao Y, Calamito M, Wei F, et al. (2013) YY1 controls Igkappa repertoire and B-cell development, and localizes with condensin on the Igkappa locus. EMBO J 32: 1168–1182.
25. Melchers F (2005) The pre-B-cell receptor: selector of fitting immunoglobulin heavy chains for the B-cell repertoire. Nat Rev Immunol 5: 578–584.
26. Li Z, Dordai DI, Lee J, Desiderio S (1996) A conserved degradation signal regulates RAG-2 accumulation during cell division and links V(D)J recombination to the cell cycle. Immunity 5: 575–589.
27. Thompson EC, Cobb BS, Sabbattini P, Meixlsperger S, Parelho V, et al. (2007) Ikaros DNA-binding proteins as integral components of B cell developmental-stage-specific regulatory circuits. Immunity 26: 335–344.
28. Nakayama J, Yamamoto M, Hayashi K, Satoh H, Bundo K, et al. (2009) BLNK suppresses pre-B-cell leukemogenesis through inhibition of JAK3. Blood 113: 1483–1492.
29. Herzog S, Hug E, Meixlsperger S, Paik JH, DePinho RA, et al. (2008) SLP-65 regulates immunoglobulin light chain gene recombination through the PI(3)K-PKB-Foxo pathway. Nat Immunol 9: 623–631.
30. Amin RH, Schlissel MS (2008) Foxo1 directly regulates the transcription of recombination-activating genes during B cell development. Nat Immunol 9: 613–622.
31. Novobrantseva TI, Martin VM, Pelanda R, Muller W, Rajewsky K, et al. (1999) Rearrangement and expression of immunoglobulin light chain genes can precede heavy chain expression during normal B cell development in mice. J Exp Med 189: 75–88.
32. Melchers F, ten Boekel E, Seidl T, Kong XC, Yamagami T, et al. (2000) Repertoire selection by pre-B-cell receptors and B-cell receptors, and genetic control of B-cell development from immature to mature B cells. Immunol Rev 175: 33–46.
33. Schlissel MS (2004) Regulation of activation and recombination of the murine Igkappa locus. Immunol Rev 200: 215–223.
34. Kersseboom R, Ta VB, Zijlstra AJ, Middendorp S, Jumaa H, et al. (2006) Bruton's tyrosine kinase and SLP-65 regulate pre-B cell differentiation and the induction of Ig light chain gene rearrangement. J Immunol 176: 4543–4552.
35. Dingjan GM, Middendorp S, Dahlenborg K, Maas A, Grosveld F, et al. (2001) Bruton's tyrosine kinase regulates the activation of gene rearrangements at the lambda light chain locus in precursor B cells in the mouse. J Exp Med 193: 1169–1178.
36. Roldan E, Fuxa M, Chong W, Martinez D, Novatchkova M, et al. (2005) Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. Nat Immunol 6: 31–41.
37. Inlay MA, Tian H, Lin T, Xu Y (2004) Important roles for E protein binding sites within the immunoglobulin kappa chain intronic enhancer in activating Vkappa Jkappa rearrangement. J Exp Med 200: 1205–1211.
38. Lazorchak AS, Schlissel MS, Zhuang Y (2006) E2A and IRF-4/Pip promote chromatin modification and transcription of the immunoglobulin kappa locus in pre-B cells. Mol Cell Biol 26: 810–821.
39. Reynaud D, Demarco IA, Reddy KL, Schjerven H, Bertolino E, et al. (2008) Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros. Nat Immunol 9: 927–936.
40. Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, et al. (2012) Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nat Immunol 13: 1196–1204.
41. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, et al. (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc 8: 509–524.
42. Jumaa H, Wollscheid B, Mitterer M, Wienands J, Reth M, et al. (1999) Abnormal development and function of B lymphocytes in mice deficient for the signaling adaptor protein SLP-65. Immunity 11: 547–554.
43. Middendorp S, Dingjan GM, Hendriks RW (2002) Impaired precursor B cell differentiation in Bruton's tyrosine kinase-deficient mice. J Immunol 168: 2695–2703.
44. Jumaa H, Mitterer M, Reth M, Nielsen PJ (2001) The absence of SLP65 and Btk blocks B cell development at the preB cell receptor-positive stage. Eur J Immunol 31: 2164–2169.
45. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, et al. (2006) TM4 microarray software suite. Methods Enzymol 411: 134–193.
46. Kersseboom R, Middendorp S, Dingjan GM, Dahlenborg K, Reth M, et al. (2003) Bruton's tyrosine kinase cooperates with the B cell linker protein SLP-65 as a tumor suppressor in Pre-B cells. J Exp Med 198: 91–98.
47. Bertocci B, De Smet A, Berek C, Weill JC, Reynaud CA (2003) Immunoglobulin kappa light chain gene rearrangement is impaired in mice deficient for DNA polymerase mu. Immunity 19: 203–211.
48. Baldwin AS Jr, LeClair KP, Singh H, Sharp PA (1990) A large protein containing zinc finger domains binds to related sequence elements in the enhancers of the class I major histocompatibility complex and kappa immunoglobulin genes. Mol Cell Biol 10: 1406–1414.

49. Engel H, Rolink A, Weiss S (1999) B cells are programmed to activate kappa and lambda for rearrangement at consecutive developmental stages. Eur J Immunol 29: 2167–2176.
50. Gorman JR, van der Stoep N, Monroe R, Cogne M, Davidson L, et al. (1996) The Ig(kappa) enhancer influences the ratio of Ig(kappa) versus Ig(lambda) B lymphocytes. Immunity 5: 241–252.
51. Xu Y, Davidson L, Alt FW, Baltimore D (1996) Deletion of the Ig kappa light chain intronic enhancer/matrix attachment region impairs but does not abolish V kappa J kappa rearrangement. Immunity 4: 377–385.
52. Inlay M, Alt FW, Baltimore D, Xu Y (2002) Essential roles of the kappa light chain intronic enhancer and 3′ enhancer in kappa rearrangement and demethylation. Nat Immunol 3: 463–468.
53. Greenbaum S, Zhuang Y (2002) Identification of E2A target genes in B lymphocyte development by using a gene tagging-based chromatin immunoprecipitation system. Proc Natl Acad Sci U S A 99: 15030–15035.
54. Aoki-Ota M, Torkamani A, Ota T, Schork N, Nemazee D (2012) Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. J Immunol 188: 2305–2315.
55. Ferreiros-Vidal I, Carroll T, Taylor B, Terry A, Liang Z, et al. (2013) Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. Blood 121: 1769–1782.
56. Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, et al. (2010) A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat Immunol 11: 635–643.
57. Sakamoto S, Wakae K, Anzai Y, Murai K, Tamaki N, et al. (2012) E2A and CBP/p300 act in synergy to promote chromatin accessibility of the immunoglobulin kappa locus. J Immunol 188: 5547–5560.
58. Heizmann B, Kastner P, Chan S (2013) Ikaros is absolutely required for pre-B cell differentiation by attenuating IL-7 signals. J Exp Med; e-pub. December 2013.
59. Feeney AJ (2011) Epigenetic regulation of antigen receptor gene rearrangement. Curr Opin Immunol 23: 171–177.
60. Liu Y, Subrahmanyam R, Chakraborty T, Sen R, Desiderio S (2007) A plant homeodomain in RAG-2 that binds Hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. Immunity 27: 561–571.
61. Matthews AG, Kuo AJ, Ramon-Maiques S, Han S, Champagne KS, et al. (2007) RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature 450: 1106–1110.
62. Bain G, Robanus Maandag EC, te Riele HP, Feeney AJ, Sheehy A, et al. (1997) Both E12 and E47 allow commitment to the B cell lineage. Immunity 6: 145–154.
63. Goldmit M, Ji Y, Skok J, Roldan E, Jung S, et al. (2005) Epigenetic ontogeny of the Igk locus during B cell development. Nat Immunol 6: 198–203.
64. Degner SC, Verma-Gaur J, Wong TP, Bossen C, Iverson GM, et al. (2011) CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. Proc Natl Acad Sci U S A 108: 9566–9571.
65. Guo C, Yoon HS, Franklin A, Jain S, Ebert A, et al. (2011) CTCF-binding elements mediate control of V(D)J recombination. Nature 477: 424–430.
66. Xu CR, Feeney AJ (2009) The epigenetic profile of Ig genes is dynamically regulated during B cell differentiation and is modulated by pre-B cell receptor signaling. J Immunol 182: 1362–1369.
67. Inlay MA, Lin T, Gao HH, Xu Y (2006) Critical roles of the immunoglobulin intronic enhancers in maintaining the sequential rearrangement of IgH and Igk loci. J Exp Med 203: 1721–1732.
68. Brown KE, Guest SS, Smale ST, Hahm K, Merkenschlager M, et al. (1997) Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. Cell 91: 845–854.
69. Kim J, Sif S, Jones B, Jackson A, Koipally J, et al. (1999) Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. Immunity 10: 345–355.
70. Koipally J, Renold A, Kim J, Georgopoulos K (1999) Repression by Ikaros and Aiolos is mediated through histone deacetylase complexes. EMBO J 18: 3090–3100.
71. Schmitt C, Tonnelle C, Dalloul A, Chabannon C, Debre P, et al. (2002) Aiolos and Ikaros: regulators of lymphocyte development, homeostasis and lymphoproliferation. Apoptosis 7: 277–284.
72. Holwerda SJ, van de Werken HJ, Ribeiro de Almeida C, Bergen IM, de Bruijn MJ, et al. (2013) Allelic exclusion of the immunoglobulin heavy chain locus is independent of its nuclear localization in mature B cells. Nucleic Acids Res 41: 6905–6916.
73. Martin F, Chen X, Kearney JF (1997) Development of VH81X transgene-bearing B cells in fetus and adult: sites for expansion and deletion in conventional and CD5/B1 cells. Int Immunol 9: 493–505.
74. Mombaerts P, Iacomini J, Johnson RS, Herrup K, Tonegawa S, et al. (1992) RAG-1-deficient mice have no mature B and T lymphocytes. Cell 68: 869–877.
75. Hendriks RW, de Bruijn MF, Maas A, Dingjan GM, Karis A, et al. (1996) Inactivation of Btk by insertion of lacZ reveals defects in B cell development only past the pre-B cell stage. EMBO J 15: 4862–4872.
76. Stadhouders R, Thongjuea S, Andrieu-Soler C, Palstra RJ, Bryne JC, et al. (2012) Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. EMBO J 31: 986–999.
77. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B (2013) r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res 41: e132.
78. Ikawa T, Kawamoto H, Wright LY, Murre C (2004) Long-term cultured E2A-deficient hematopoietic progenitor cells are pluripotent. Immunity 20: 349–360.
79. Sayegh CE, Jhunjhunwala S, Riblet R, Murre C (2005) Visualization of looping involving the immunoglobulin heavy-chain locus in developing B cells. Genes Dev 19: 322–327.

Chapter 9

General Discussion

Parts of this Chapter were published in:

Transcription
2012; 3:181-6
&
Blood
2012; 119:6209-18

The human genome and its 'promise'

Fourteen years have passed since US President Bill Clinton and UK Prime Minister Tony Blair announced the completion of the first draft sequence of the human genome, which they called "the most wondrous map ever produced by humankind"[1]. This qualification illustrates the almost infinite optimism that surrounded the publication of the human genome sequence at the start of this millennium. It was expected to rapidly revolutionize the way we diagnose, prevent and treat major human diseases[1].

The availability of the entire human genome sequence surely has revolutionized the life sciences. It has had a tremendous impact on our knowledge of human genome anatomy, genetic variation among humans, human evolutionary history and spurred amazing technological advances in genome sequencing. In the last decade, sequencing costs have dropped >10,000 fold and the time it takes to sequence a complete human genome is now a matter of days instead of years[2]. However, it is only fair to state that it has not yet fulfilled it promises in terms of radically changing the way we practice medicine[3,4]. Although the human genome sequence has greatly facilitated the discovery of new disease genes and mutations, knowledge of virtually every letter of our genetic code has thus far not led to the major medical breakthroughs that some expected it to bring about. Nevertheless, several significant advances have been made, including predicting drug response in individuals, the development of several new cancer drugs and the identification of major risk factors for several diseases[4,5].

An important reason for this apparent 'lack of major translational value' of the human genome sequence is one that was difficult to predict upfront. After its completion, scientists began to realize that the functional information content of our genome is much more complicated than previously anticipated. Traditionally, the major focus of biomedical research has been with the protein-coding part of our genome (consisting of 22.000 genes), which has long been known to play a critical role in human biology and disease. We now know, in part due to the availability of the human genome sequence, that the non-coding part of our genome (>95% of our total genome) is just as important when it comes to understanding human biology and disease[2]. It turns out that obtaining the sequence itself was just the first step: it provided the framework required to begin understanding how our genome really works. As a consequence, scientific focus is now shifting[6] towards investigating the non-coding part of our genome (exemplified by the creation of large-scale consortia such as ENCODE[7] and FANTOM[8]): how the vast and largely unexplored non-coding part functions and influences the protein-coding part and how disease and trait-associated mutations affect genome function. Obtaining such knowledge will be an important step towards the realization of the human genome sequence as the key to revolutionize our understanding of human biology in both health and disease[9]. The experiments described in this thesis can be viewed as part of this ongoing effort to understand the function of our entire genome.

Studying gene regulation during blood cell differentiation

We have focused our efforts on an import aspect of genome biology: gene regulation by sequence-specific DNA-binding proteins called transcription factors (TFs). Hematopoietic differentiation processes were chosen as a model system because of our laboratory's extensive experience with such systems and the relatively easy way of obtaining primary cells from mice and humans. Our aims were twofold. First, we wanted to obtain specific insight into how TFs dictate gene expression patterns and V(D)J-recombination during hematopoiesis. As the disruption of TF function and gene regulation during hematopoietic development is a common cause of disease (e.g. leukemia), such knowledge also has potential implications for human health. Second, we hoped to uncover novel general principles underlying mammalian gene regulation.

Novel repressors of erythroid gene expression that keep activators in check

Proper timing of gene activation is critical for development. We and others have previously shown that in erythroid progenitors, prior to their full activation, genes of the late erythroid transcriptome (e.g. globins, membrane proteins) are already bound by the LDB1 TF complex that is responsible for their activation upon terminal differentiation[10,11]. This observation raises an interesting question: what prevents these genes from being prematurely activated? Another member of the LDB1-complex, the ETO2 repressor, was found to be a key player in keeping late erythroid genes poised for activation in progenitors[10,12-14]. Upon terminal erythroid

differentiation, ETO2-mediated repression of these genes appeared to be lost, although mechanistically it remained largely unclear how ETO2 achieves this temporal repression.

We have further investigated ETO2-mediated gene repression in erythroid progenitors using a combination of proteomics and genomics approaches (Chapter 2). This way we were able to confirm known interactions and identify new ETO2 protein partners involved in gene silencing. One of our most interesting hits was the IRF2BP2 protein, which until then had not been implicated in regulating erythroid differentiation. Possibly together with the GFI1B TF, this ETO2-IRF2BP2 axis confers potent repression of the LDB1-complex bound late erythroid genes in progenitor cells, which is lost upon terminal differentiation.

An important question is what exactly leads to this loss of repression. Current literature favours a loss of repressor gene expression at later stages of erythropoiesis as a plausible answer. In terms of expression levels, *Cbfa2t3*/ETO2 levels are known to be downregulated upon differentiation[10,12,14,15], and we have shown that *Irf2bp2* transcription follows a similar pattern (Chapter 2). Using a combination of G1E and ES model systems of erythroid differentiation, Fujiwara et al.[16] show that *Cbfa2t3* is occupied and activated by a GATA2-containing LDB1-complex in early erythroid progenitors. Upon GATA1 induction and GATA2-to-GATA1 switching, *Cbfa2t3* expression is repressed, potentially mediated by a loss of TAL1 (the LDB1-complex's main activator[12,14]) occupancy[16]. We reveal that similar events appear to take place at the *Irf2bp2* locus, suggesting that *Cbfa2t3* and *Irf2bp2* are both silenced upon terminal erythroid differentiation when

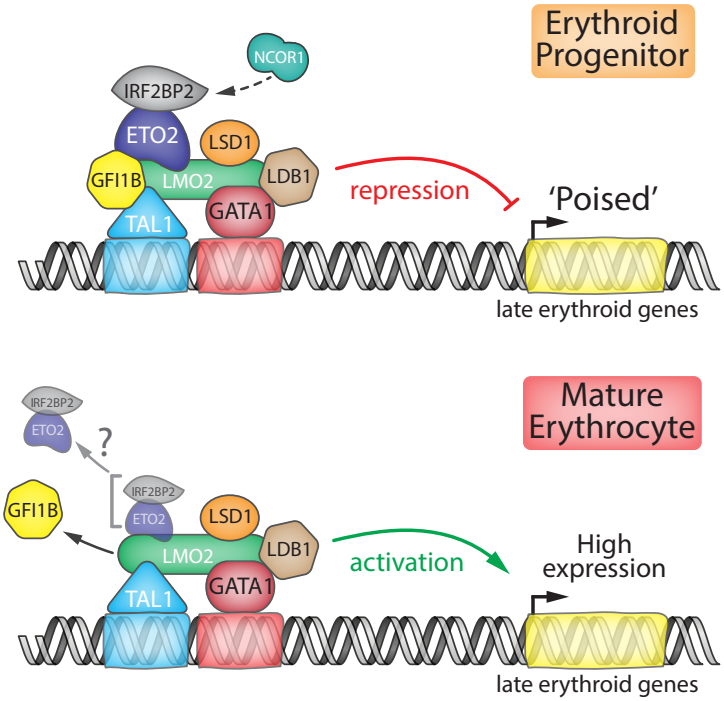


Figure 1. Model of LDB1-complex target gene repression by the ETO2-IRF2BP2 axis during erythroid differentiation. In erythroid progenitors, the LDB1-complex is already recruited to its 'late erythroid' target genes (e.g. heme biosynthesis genes) before their full activation is required. At this stage, premature target gene activation is prevented by the cooperative action of the ETO2-IRF2BP2 subunits (and possibly GFI1B). ETO2-IRF2BP2 repression might be mediated by the NCOR1 repressor complex (Upper part). In mature terminally differentiating erythrocytes, gene repression by ETO2-IRF2BP2 (and GFI1B) is lost – resulting in the full activation of LDB1-complex target genes (Lower part). Published evidence favours a loss of ETO2-IRF2BP2 expression upon erythroid maturation, resulting in either an absolute loss of genomic occupancy or a relative one as compared to the levels of the activator subunits TAL1 and LDB1 (the latter phenomenon is illustrated by the smaller size and increased transparency of the ETO2-IRF2BP2 proteins within the LDB1-complex). The schematic representation does not take into account the right size dimensions of the double helix with regard to the schematized proteins.

their repression of late erythroid genes needs to be relieved. Studies employing ChIP experiments[10,12,14] have shown that ETO2 occupancy of late erythroid target genes decreases; either in absolute terms or in a relative fashion compared to the levels of activators (i.e. TAL1, LDB1). IRF2BP2 genomic occupancy during erythroid differentiation would be expected to follow the pattern observed for ETO2, although this remains to be experimentally demonstrated (Figure 1).

Our generation of an IRF2BP2-deficient mouse model revealed for the first time that *Irf2bp2* is an essential gene, as live IRF2BP2-null mice were rarely obtained. The few knockout animals that were born alive never reached 5 weeks of age and displayed severe growth retardation. These observations prompt the development of a conditional *Irf2bp2* allele to study IRF2BP2 function in adult mice. During embryonic development, we could not detect any gross abnormalities when examining IRF2BP2-null embryos. Closer inspection did reveal a modest defect in fetal liver erythroid development, indicating that the absence of IRF2BP2 resulted in a partial developmental block of erythroid maturation *in vivo*. The flow cytometry analyses we conducted are however still somewhat superficial. Pinpointing the exact nature of the defect will require further experimentation and would greatly benefit from the availability of a conditional *Irf2bp2* knockout model system to study adult erythropoiesis in the absence of IRF2BP2. Outstanding issues include the possible role of IRF2BP2 in the enucleation process (as suggested by our experiments in Chapter 2) and whether IRF2BP2, like ETO2[17], is important for stress erythropoiesis. In light of its proposed interaction with IRF2[18], a TF important for a broad range of hematopoietic lineages[19,20], it will be of interest to also investigate the other hematopoietic lineages in IRF2BP2-deficient mice or embryos. A preliminary analysis of fetal liver and bone marrow hematopoietic cells in E15-18 IRF2BP2-null embryos indicated abnormal development of the granulocyte and monocyte lineages (R. Stadhouders, J.C. Boisset, C. Robin, E. Soler, unpublished data).

From a more general perspective, our results confirm that multimeric regulatory complexes (such as the LDB1-complex) feature a dynamic interplay between activating and repression components that determines lineage-specific gene expression. Molecular dissection of these complexes will remain an important strategy in revealing how these 'gene regulatory machines' operate, in both normal and disease cellular states. However, there are important aspects of TF biology that current '-omics' approaches (such as those employed in Chapter 2) cannot elucidate. As their general nature is static and '-omics' technology often measures an average of a large population of cells, we still have very limited insight into the actual dynamics of a TF complex regulating its target genes during the developmental maturation of a cell. Answering important questions such as 'How dynamic are associations between individual TF complex components and between the TF complex and chromatin?', 'Are there smaller sub-complexes within a large TF complex that compete with each other for chromatin occupancy (e.g. a repressing and activating sub-complex)?', 'What kind of heterogeneity in TF dynamics is present within a population of cells?' will require another technical leap forward, posing major challenges for the coming years.

3C-Seq facilitates the discovery of long-range regulatory connections

The development of Chromosome Conformation Capture (3C) methodology by Dekker et al. in 2002[21] has had a dramatic impact on the study of genome biology. 3C and its 4C, 5C and Hi-C derivatives (reviewed by De Wit and De Laat[22]) enable researchers to investigate the spatial conformation of a genome, allowing them to study the 3D organisation of entire chromosomes but also to identify individual chromatin loops that connect a promoter to an enhancer. Fuelled by our observation that the LDB1-complex mainly regulates its target genes from a distance[10], we made an effort to adapt the existing microarray-based 4C method[23] to next-generation sequencing platforms, including the development of bioinformatics tools for data analysis (Chapters 3 and 4). This has greatly improved the resolution and throughput of the 4C technique, as was also shown by analogous efforts from the De Laat laboratory (called '4C-Seq'[24]). The 3C-Seq method we developed has been an indispensable tool for the in-depth characterization of chromatin topology at the *Myb* and *Igκ* loci described in Chapters 5, 6, 7 and 8 of this thesis. Many other laboratories have now also successfully implemented 3C- and 4C-Seq techniques into their research efforts[25-29], emphasizing the broad relevance of this technology for studying gene regulation.

Myb oncogene regulation from a distance: opportunity for therapy?

The *Myb* proto-oncogene plays a pivotal role in hematopoietic development: *Myb*-deficient mice exhibit

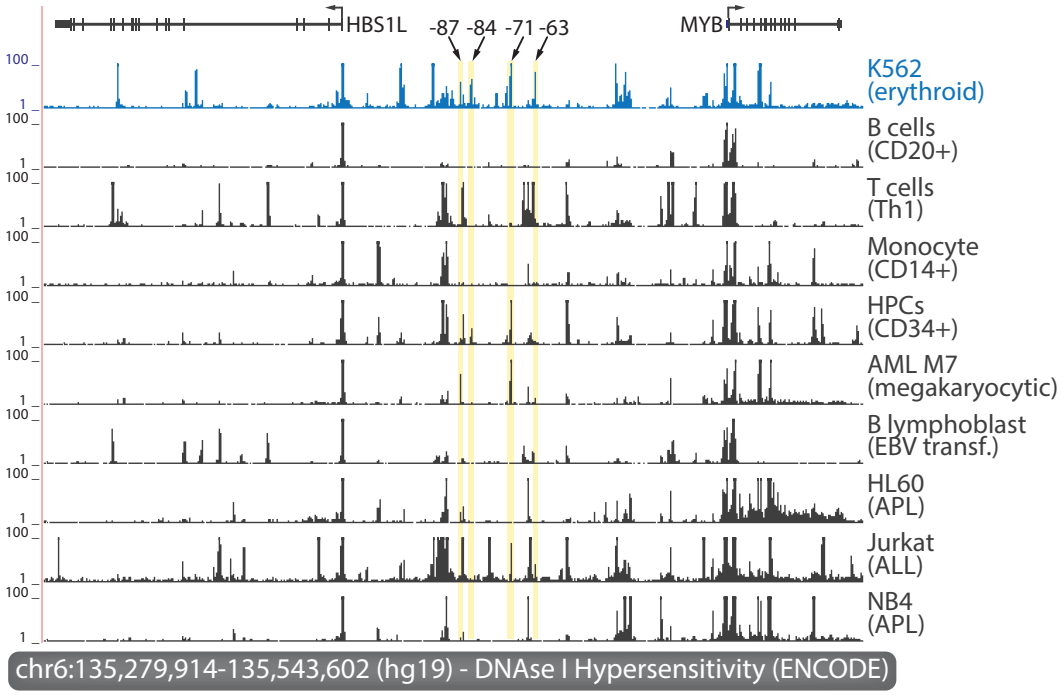


Figure 2. DNAse I hypersensitive sites are present in the *HBS1L-MYB* intergenic region of several (malignant) hematopoietic cell types. DNAse I hypersensitivity (as measured by DNAse I-Seq[98]) profiles of several malignant and non-malignant hematopoietic cell types. Depicted is the human *HBS1L-MYB* intergenic region on chromosome 6 (hg19 coordinates shown). Locations of the four conserved LDB1-complex binding sites (-87, -84, -71 and -63 upstream of the *MYB* transcription start site) are indicated. HPCs, Hematopoietic Progenitor Cells; AML, Acute Myeloid Leukemia; EBV, Epstein-Barr Virus; APL, Acute Promyelocytic Leukemia; ALL, Acute Lymphoblastic Leukemia

a complete lack of virtually all definitive hematopoietic cells, resulting in embryonic death due to a fatal anemia[30]. *Myb*, encoding the DNA binding TF c-Myb, is highly expressed in hematopoietic stem/progenitor cells and is rapidly silenced upon their terminal differentiation. In accordance with this expression pattern, c-Myb is thought to maintain a proliferative cellular phenotype by controlling genes involved in proliferation and survival. Downregulation of *Myb* is necessary for the initiation of terminal differentiation, as is illustrated by the ability of c-Myb to block differentiation and promote leukemogenesis when overexpressed[31].

Despite the clear importance of accurately regulating c-Myb levels, *Myb* transcriptional control remained poorly understood. Our combination of ChIP-Seq, 3C-Seq and RNAi experiments in mouse and human erythroid progenitors (described in Chapters 5 and 6) provides the first comprehensive analysis of the *Myb/MYB* regulatory landscape and its dynamics during a hematopoietic differentiation process. In erythroid progenitors expressing high levels of *Myb*, a cluster of distal enhancers bound by the LDB1-complex and the key erythroid KLF1 TF is responsible for *Myb* activation. We propose the formation of an active chromatin hub (ACH) in these progenitors, resulting in the spatial clustering of regulatory complexes and the general transcriptional machinery around *Myb* to induce its activation. This 3D-organisation is lost upon differentiation, when *Myb* expression is downregulated (Chapter 5).

One remaining question is whether similar intergenic (clusters of) regulatory elements control *Myb* expression in non-erythroid hematopoietic cell types. Several observations suggest that this might indeed be the case. First, other studies have detected TF binding within the murine *Myb-Hbs1l* intergenic region of hemangioblast cells[32], hematopoietic stem/progenitor cells[33] and leukemic myeloblast cells[34]. Second, analysis of publically available DNaseI-Seq datasets[35] suggests that many human hematopoietic cell types (both primary cells and leukemic cell lines) might also bear intergenic regulatory sequences

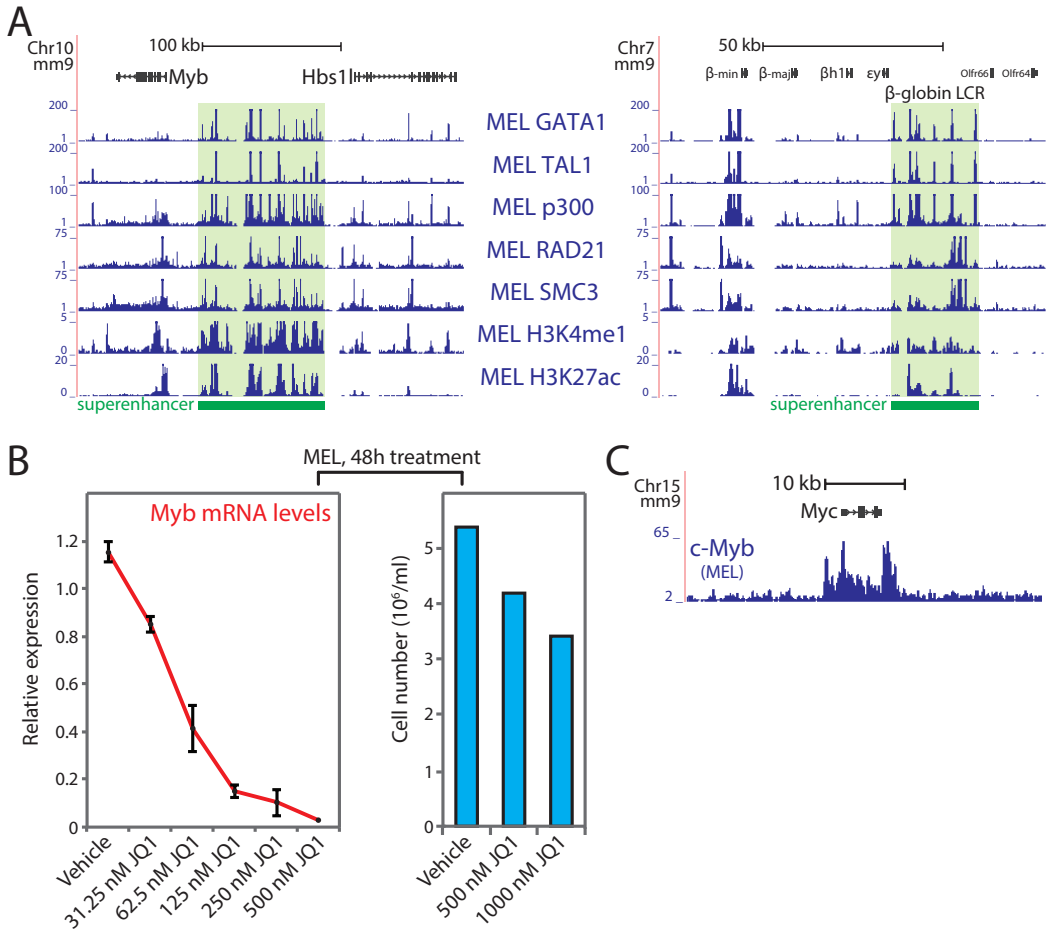


Figure 3. A putative *Myb-Hbs1l* intergenic superenhancer in erythroid progenitors. (A) Enrichments (as measured by ChIP-Seq[98] in mouse erythroleukemic (MEL) cells) of key erythroid TFs (GATA1 and TAL1), the histone acetyltransferases p300, Cohesin complex subunits (RAD21 and SMC3) and enhancer-associated histone modifications (H3K4Me1 and H3K27Ac) at the *Myb-Hbs1l* intergenic region (left) and the β-globin locus (containing a known superenhancer; right). (Putative) superenhancer domains are marked by green shading. (B) MEL cells were treated for 48 hours with various concentrations of a BRD4 inhibitor (JQ1), after which cell numbers (right) and *Myb* mRNA levels were measured (by quantitative PCR and normalized versus *Rnh1* levels; left). Measurements are plotted as average values of two independent experiments (error bars denote s.d.). (C) Enrichment of c-Myb TF binding (as measured by ChIP-Seq[98]) in MEL cells) at the *Myc* locus. LCR, Locus Control Region

controlling *MYB* (Figure 2). We therefore speculate that conserved regulatory elements within the *Myb-Hbs1l* intergenic region control *Myb* expression in a broad range of hematopoietic cell types. Note however that DNAseI hypersensitivity at the conserved LDB1-complex binding sites we identified in primary human erythroid progenitors (-87, -84, -71 and -63, also present in K562 erythroleukemic cells) appears to be largely restricted to erythroid cells (Figure 2).

Our studies of *Myb/MYB* regulation suggest the involvement of multiple types of regulatory elements (enhancers, promoters and insulators) spread over a large genomic region that synergistically modulate gene transcription via chromatin looping. These observations are reminiscent of β-globin gene regulation by a distal LCR and surrounding insulator elements[36]. However, whether the *Myb-Hbs1l* intergenic

regulatory elements represent a bona fide LCR remains to be experimentally verified. The regulatory architecture of the intergenic region also corresponds to the recently identified class of superenhancers, which were found to be enriched near key cell identity genes and oncogenes[37,38]. Superenhancers bear a striking resemblance to LCR regions[38,39], and it remains presently unclear whether superenhancers are LCRs or whether they are a bona fide novel class of regulatory elements. Originally identified through their unusually high occupancy by the Mediator complex[38], many TFs, histone modifying complexes, histone modifications and Cohesin also show disproportionally high enrichments at superenhancers, as compared to 'typical' enhancers[40]. Indeed, the *Myb-Hbs1l* intergenic region displays essentially all chromatin hallmarks of a superenhancer (Figure 3A), similar to the β-globin LCR that was previously designated as a superenhancer[39].

An interesting aspect of superenhancers is their hypersensitivity to small molecule inhibitors targeting BRD4, another gene regulatory protein highly enriched at superenhancers[37]. Treatment of leukemic cells or leukemia-engrafted mice with BRD4 inhibitors yielded promising results, leading to the selective elimination of malignant cells by crippling the expression of key oncogenes such as *Myc*[41-43]. BRD4 inhibition to target the putative intergenic *Myb* superenhancer therefore seems a plausible therapeutic approach to induce *Myb* downregulation in blood cancers that depend on c-Myb for proliferation (e.g. T cell[44] and myeloid leukemias[45]). Indeed, preliminary studies on murine erythroleukemic cells reveal a potent inhibition of *Myb* expression and cell proliferation upon BRD4 inhibition (Figure 3B). Intriguingly, Zuber et al.[42,45] noted that gene expression changes in myeloid leukemic cells observed after BRD4 inhibition were very similar to those detected upon RNAi-mediated c-Myb depletion, including a strong downregulation of *Myc* expression. Previous work on a myeloid leukemic cell line also found c-Myb to act upstream of c-Myc[46] and c-Myb binds the *Myc* promoter in MEL cells (Figure 3C). Together, these observations suggest the existence of a druggable BRD4/c-Myb/c-Myc axis essential for leukemia maintenance.

Silencing *Myb* expression: who's responsible and how?

Downregulation of *Myb*/*MYB* is a well-documented prerequisite for terminal hematopoietic differentiation[31]. Our findings of a general LDB1-complex displacement from the intergenic enhancers and a loss of ACH formation upon erythroid maturation (Chapters 5 and 6) provide a direct mechanism for the loss of *Myb* expression. Nevertheless, it remains unclear how TF displacement is achieved.

A global loss of LDB1-complex levels upon terminal differentiation as a reason for displacement can directly be excluded, as it is firmly established that the core LDB1-complex components are absolutely essential for erythroid maturation and remain associated with numerous erythroid target genes (e.g. globins) throughout terminal differentiation[10,47,48]. One remaining and plausible explanation is the recruitment of other TFs to *Myb* regulatory elements at the onset of terminal differentiation. Such factors could displace activating factors due to direct competition for DNA binding, by altering local chromatin structure or through post-translational modifications (e.g. ubiquitination, resulting in activator degradation). We propose to focus future efforts aimed at identifying such proteins on two candidate TFs in particular.

A first potential *Myb* repressor is GFI1B, a DNA-binding TF required for erythroid development[49]. During erythroid development, GFI1B appears to predominantly function as a repressor through its association with the histone demethylase LSD1[50]. Depletion of GFI1B in cultures of differentiating human erythroid progenitors causes a delay in maturation; overexpressing GFI1B in these cultures accelerates differentiation[51]. Similarly, erythroid-specific deletion of the *Aof2* gene (encoding LSD1) in mice results in ineffective fetal liver erythropoiesis and derepression of a stem/progenitor cell gene expression program[52]. Interestingly, Saleque et al.[50] already showed that the GFI1B-LSD1 complex binds the *Myb* promoter in MEL cells, and they observed a further increase in GFI1B binding when cells were induced to differentiate[50]. Preliminary experiments in MEL cells confirmed these results: increased GFI1B-LSD1-RUNX1 (the latter TF was recently shown to interact with LSD1[53]) occupancy of the *Myb* promoter and intergenic enhancers was detected in differentiated MEL cells (data not shown). Experiments aimed at disrupting the GF1B-LSD1(-RUNX1) complex during erythroid differentiation will have to reveal whether *Myb* repression truly depends on these TFs.

The second candidate *Myb* repressors are the ZEB TFs: ZEB1 (or δEF1) and ZEB2 (or SIP1). The rationale behind this implication of ZEB factors in *Myb* silencing originated from an analysis of TF motifs present in LDB1-complex binding regions that (partially) lost complex occupancy upon MEL cell differentiation

(including the *Myb* intergenic regulatory elements; J.C. Bryne and E. Soler, unpublished results). The ZEB DNA-binding motif, an E-box variant[54], was found to be one of the most significantly overrepresented motifs, suggesting ZEB factors might occupy (a subset of) these sites during erythropoiesis. Furthermore, both ZEB proteins have been implicated in gene repression[55,56]. Interestingly, ZEB TF expression is induced by TGFβ signaling[57,58]. This signaling pathway regulates terminal erythroid differentiation by inhibiting proliferation and stimulating subsequent cellular maturation[59], a series of events analogous to the presumed consequences of *Myb* downregulation in maturing erythroid progenitors (see the Discussion section of Chapter 6 and references therein). We hypothesize that the upregulation of ZEB TF expression by TGFβ signaling might lead to increased ZEB TF recruitment to *Myb* regulatory elements, which could in turn result in *Myb* downregulation and the initiation of erythroid maturation. Until now, ZEB1 and ZEB2 have not been implicated in red blood cell development, although ZEB-deficient mouse models have uncovered important roles for ZEB proteins in hematopoietic stem/progenitor cell differentiation[60] and T cell development[61]. Experiments investigating the function of ZEB TFs and their protein partners in erythroid cells are currently ongoing. Finally, it should be noted that both repressor candidates might operate simultaneously or even synergistically: like GFI1B, ZEB1 was previously found to repress gene expression through its interaction with LSD1[62].

Myb intron 1 contains a versatile regulatory element bound by the insulator protein CTCF

Particularly captivating aspects of *Myb* transcriptional regulation are the regulatory events occurring at a site within its first intron. Early work on *Myb* regulation in leukemic cell lines pointed at the *Myb* first intron as a region containing a block to transcription elongation[63,64]. These studies showed that in hematopoietic progenitors (expressing high levels of *Myb* mRNA) this intronic elongation block is overcome, although the underlying mechanisms remain elusive. Our studies described in Chapter 5 validate this model by demonstrating the presence of an RNAPII CTD phosphorylation switch within the first intron of *Myb*. Initiating RNAPII (Ser5P) travels ±2.5kb from the TSS across exon 1 into the first intron, where it is converted into an elongation-competent RNAPII (Ser2P). Our data suggest that Ser5P-Ser2P switching is achieved by enhancer-bound transcription and elongation factors (e.g. CDK9) that are brought into close proximity of the promoter and first intron through chromatin looping. To our knowledge, this was the second study providing evidence for the regulation of transcriptional elongation by distal enhancers; the first demonstration came from work on the β-globin LCR[65]. A recent study from the Rosenfeld laboratory confirmed the concept of long-range regulation of RNAPII elongation by enhancers on a genome-wide level[66].

Several studies have reported proteins binding in the vicinity of the intronic elongation switch site, including ETS1[67], c-Jun[68], NF-κB[69] and ERα[70,71]. We expanded this list by showing that the CTCF insulator protein also binds this region and is required for high-level *Myb* expression. Nevertheless, the exact role of CTCF at the intronic elongation switch site is not entirely clear yet, although the evolutionary highly conserved nature of the CTCF binding motif suggests it is functionally important. We propose a dual role for CTCF in regulating *Myb* transcription. Considering the precise localization of the CTCF binding site immediately adjacent of the initiation-elongation switch site, it is tempting to speculate that CTCF directly interferes with RNAPII elongation. How CTCF establishes an elongation block is unknown, but it might represent a more widespread phenomenon: Paredes et al.[72] reported a strong genome-wide correlation between promoter-proximal (5'UTR) CTCF binding and high RNAPII pausing indexes. The presence of CTCF was also reported to promote RNAPII pausing within gene bodies to regulate alternative splicing[73]. A second potential function for CTCF involves the 3D chromatin structure of the *Myb* locus. Studies from the Blobel and Dean laboratories have demonstrated that in the β-globin locus LDB1-dimerization mediates chromatin loop formation between regulatory elements bound by the LDB1-complex[74,75]. While spatial clustering of *Myb* intergenic enhancers via such a mechanism is possible, it cannot explain chromatin looping with the 5' region of *Myb*, since it is completely devoid of LDB1-complex binding. Previous studies have implicated CTCF in orchestrating chromosome conformation[76] (also see Chapter 7), and we observed a strong presence of CTCF-bound genomic regions (including the first intron) within the *Myb* ACH. We propose that the highly enriched intronic CTCF site acts as an anchor that brings the intergenic region (containing the enhancers and other CTCF-occupied regions) in close nuclear proximity to the *Myb* promoter and initiation-elongation switch site (Chapter 5). Notably, LDB1 depletion resulted in a general reduction of chromatin looping between the *Myb* promoter and intergenic elements, suggesting that LDB1

and CTCF might act synergistically in ACH formation and/or maintenance.

Definitive proof of the precise individual contributions of the intergenic regulatory elements (in particular the -81 enhancer), the elongation block region and the intronic CTCF site is still lacking. Recent advances in genome editing techniques (e.g. CRISPR/Cas9 technology) have made it feasible to modify endogenous genomic sequences in a medium-throughput fashion[77]. Removing or mutating regulatory elements in hematopoietic cell lines or mice can provide the toolset required to functionally evaluate the role of key regulatory elements in ACH formation and (overcoming) the RNAPII elongation block. However, as redundancy among enhancers regulating the same target gene has been reported before[78,79], it might prove difficult to dissect the function of the individual intergenic enhancers. Using recombineering technology[80] we have created a mouse strain that lacks the -81 intergenic *Myb* enhancer (R. Jorna, R. Stadhouders, F. Grosveld, E. Soler; unpublished results). A first experiment on fetal liver erythroid cells obtained from homozygous -81$^{-/-}$ animals showed a ±30% reduction of *Myb* mRNA levels compared to wildtype cells (Figure 4). This observation supports the enhancer status of the -81 TF binding site, although it also shows that its presence is not an absolute requirement for *Myb* expression in red blood cells.

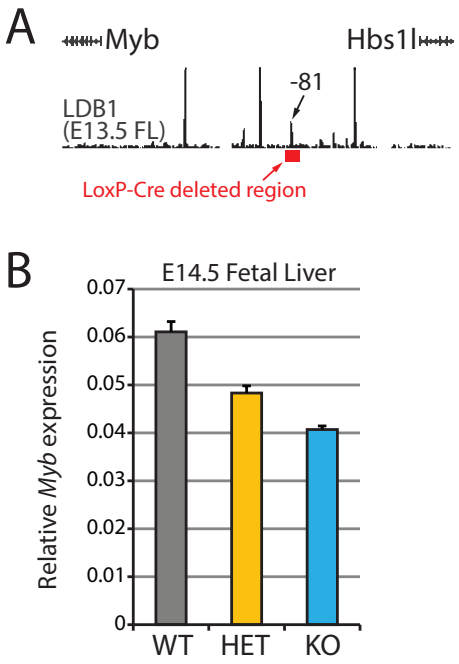


Figure 4. Deletion of the -81 *Myb* enhancer *in vivo*. (A) *Myb-Hbs1l* intergenic LDB1 occupancy in primary erythroid cells is shown (as measured by ChIP-Seq; cells were obtained from murine E13.5 fetal liver (FL)). Recombineering technology was employed to insert LoxP sites flanking the -81 enhancer region (as determined by the 347 bp of genomic sequence occupied by the LDB1 ChIP-Seq peak). Cre-mediated excision was used to delete the -81 LDB1-complex binding site (red rectangle). (B) *Myb* mRNA expression levels in E14.5 FL erythroid cells from wildtype mice (WT; n=8) and heterozygous (HET; n=13) or homozygous (KO; n=11) -81 enhancer-deleted animals were determined using quantitative PCR (normalized versus *Rnh1* levels). Error bars denote s.d.

Functional follow-up of a genetic association: clinical significance of *MYB* enhancers?

At the time we started our experiments on the murine *Myb-Hbs1l* intergenic interval, it did not escape our notice that the corresponding human region was found to contain genetic variants strongly associated with differences in clinically relevant erythroid traits[81,82]. Responsible for identifying the association between *HBS1L-MYB* intergenic polymorphisms and erythroid traits (in particular elevated HbF levels) is the laboratory of Swee Lay Thein[81,83]. Despite the strength and reproducibility of the association (the intergenic variants explain 19.4% of the variation in HbF levels among Europeans[84], which is unusually high for variants obtained through association studies[85]), its exact biological basis remained unclear. Together with the Thein group, we were able to show that these polymorphisms interfere with conserved enhancers that activate *MYB* transcription, providing a first molecular mechanism for the genotype-phenotype association (Chapter 6). Our conclusion of *MYB* as the 'culprit' gene was further reinforced by the recent identification of a child with a rare complete loss of *HBS1L* function that did not exhibit any hematological abnormalities and a normal distribution of Hb subtypes[86].

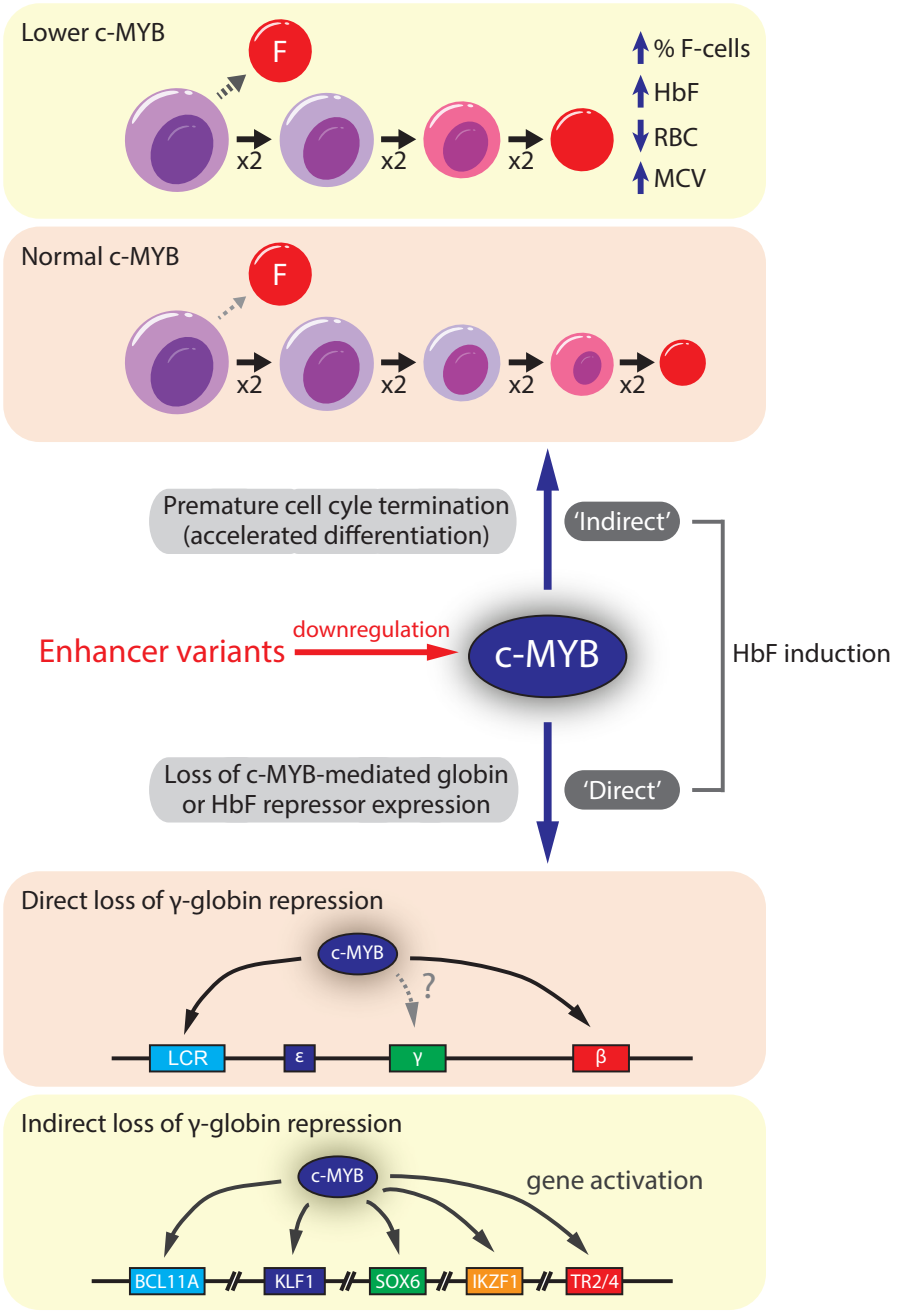


Figure 5. Dual model explaining the de-repression of HbF levels and modulation of erythroid traits when *MYB* levels are reduced. Lower c-MYB levels (e.g. as a consequence of the enhancer variants described in Chapter 6) can lead to HbF induction via increased premature cell cycle termination ('indirect', top part), resulting in the generation of more F-cells and therefore higher overall HbF levels. Fewer proliferation cycles ('x2', indicating cell division) will result in a lower red blood cell count (RBC) and a larger mean cell volume (MCV). Alternatively, lower *MYB* levels could result in a loss of proper transcriptional regulation at the β-globin locus and HbF repressor genes as these loci are bound by c-Myb in murine cells ('direct', lower part). Reduced activation by c-MYB of known HbF repressors (e.g. *BCL11A*, *KLF1*) or disrupted regulation at the β-globin locus could result in γ-globin gene reactivation and subsequent HbF induction.

As is the case for the majority of loci obtained from association studies (including the *HBS1L-MYB* intergenic interval), functional characterization of the involved polymorphisms has been complicated due to their non-genic localization[87]. The main issue that needs to be resolved is the proper identification of gene(s) that mediate the association between genetic variants and phenotype. As non-coding variants often localize to (distal) regulatory elements[88], 3C-based techniques provide an effective strategy to physically link phenotype-associated variants to specific genes. It is important to realize here that linear genomic distance is not always a reliable predictor of regulatory connections between genes and distal regulatory regions[89,90]. Investigators should make use of 3C(-Seq) technology when trying to identify genes that are controlled by regulatory elements harboring phenotype-associated variants. The value of 3C(-derived) approaches in this context was recently advocated by studies of obesity-linked variants that fall in an intron of the *FTO* gene. Although *FTO* itself was immediately regarded as the causative gene, subsequent studies using 4C-Seq pointed at the neighbouring *IRX3* gene instead[27,91]. Follow-up experiments indeed identified *IRX3*, and not *FTO*, as the gene responsible for the association between obesity and the intronic *FTO* variants[27].

The therapeutic implications of a pharmacological means to stimulate HbF synthesis in adults have long been recognized[92]. Fetal hemoglobin ($\alpha_2\gamma_2$), normally produced in minor quantities in adult red blood cells, has the ability to compensate for a lack of functional adult hemoglobin ($\alpha_2\beta_2$). This compensation significantly ameliorates the severity of diseases caused by a quantitative or qualitative lack of the adult β globin chain. Decades of dedicated research have uncovered important aspects of γ-globin gene regulation, including the identification of TFs that silence its expression in adults (reviewed by Sankaran and Orkin[93]). c-MYB is one of those TFs, and its expression levels have been shown to negatively correlate with γ-globin expression and HbF production (see Discussion section of Chapter 6 and references therein). However, direct targeting of TFs to induce HbF production in adults has remained challenging, as conventional TFs such as c-MYB have thus far been 'undruggable'[94]. An alternative strategy for lowering the levels of HbF repressor proteins is to interfere with their expression at the transcriptional level. One way of accomplishing this, apart from the BRD4-mediated strategy described above, was recently described in a study by the Orkin laboratory. Using advanced genome-editing technology, they deleted the enhancers of an HbF repressor gene (*Bcl11a*) to induce fetal globin gene expression[95]. Our characterization of *MYB* regulatory elements has provided a mechanistic basis for the future development of similar HbF-inducing strategies by interfering with *MYB* expression. Important for such approaches to become feasible will be to ensure they target *MYB* expression (almost) exclusively in erythroid cells. Since *MYB* is essential for many hematopoietic cell types[30] and intergenic regulatory potential appears to be present in other hematopoietic cell types (Figure 2), future studies will have to further validate the erythroid-specific nature of the *HBS1L-MYB* enhancers described in this thesis. Nevertheless, as moderately reduced *Myb* levels are tolerated by the erythroid system[30,45], exploring this approach in a therapeutic context might be worth pursuing.

How c-MYB represses HbF levels and modifies many other erythroid traits (e.g. erythroid cell size) is not yet fully understood. This can be partially attributed to a lack of genome-wide binding and proteomics experiments reported for c-Myb/c-MYB, although a few groups have performed gene expression profiling after c-MYB depletion in erythroid progenitors[96,97]. We mined these data, along with a publically available ENCODE c-Myb ChIP-Seq dataset from MEL cells[98] to explore possible mechanisms used by c-MYB to control γ-globin expression and a broad range of erythroid parameters. We propose two non-mutually exclusive modes of action (an 'indirect' and a 'direct' one) through which c-MYB could regulate HbF levels, of which one also explains its control over the other erythroid traits (Figure 5, also see Chapter 6 – Supplementary Figure 7). The 'indirect' mechanism is based on a small shift of the late stage erythroid proliferation/differentiation balance due to lower c-MYB levels. Previous loss-of-function studies have reported slower cell-cycle progression and accelerated differentiation kinetics in maturing erythroid progenitors[96,97,99]. Accordingly, c-Myb was found to bind and regulate several key cell-cycle regulators (e.g. *Bcl2*, *Cdk6*, *Myc*). Accelerated differentiation could favor premature cell-cycle termination during the proliferation cycles of terminal adult erythropoiesis. This could in turn result in the production of more erythroid cells that predominantly synthesize HbF (so-called "F-cells") before the switch to adult Hb synthesis occurs (Figure 5)[92]. In such a situation, lower c-MYB levels will lead to lower red blood cell counts (RBC; resulting from the reduced number of proliferation cycles) and a higher average cell size (MCV; as the erythrocytes are 'younger' red cells) (Figure 5)[100,101]. Indeed, these traits are genetically associated with the minor alleles of the intergenic SNPs[100]. The 'direct' mechanism involves direct transcriptional control of γ-globin (repressor) expression. c-Myb occupies the β-globin locus itself and was found to bind and activate several of the

established γ-globin repressor genes[93]. Reduced c-MYB levels could therefore lead to a loss of γ-globin repression, resulting in increased HbF production. Both models, although plausible, are still speculative at this point and require additional experimentation for proper validation.

B cell development, V(D)J-recombination and the role of CTCF

CTCF is an exceptional TF. The CTCF protein is absolutely essential for life, which can be attributed to its remarkably pleiotropic role in gene regulatory processes (reviewed by Ong and Corces[102]). Prominent among these roles is the ability of CTCF to mediate chromatin looping and shape 3D genome topology[103]. It does so, at least in part, by anchoring the architectural Cohesin complex to specific genomic locations[25,104,105]. We investigated the impact of *Ctcf* deletion on early B cell development, with a special emphasis on changes in gene expression and chromatin looping in the context of V(D)J-recombination (Chapter 7).

Perhaps not very surprising, deletion of the *Ctcf* gene at the early pro-B cell stage (using an *mb1*-Cre strain) severely impaired B cell development, resulting in a block at the pre-B cell developmental checkpoint. CTCF was found to control the expression of almost 200 genes, which is likely to account for the reduced size, impaired proliferation and the complete developmental arrest of CTCF-deficient pre-B cells. We next turned our attention to the V(D)J-recombination process occurring at the heavy and light chain immunoglobulin (*Ig*) loci. Since V(D)J-recombination requires extensive changes in 3D locus topology (often referred to as locus 'contraction'[106]), CTCF immediately became a prime suspect for orchestrating *Ig* locus contraction. These suspicions were further fuelled by extensive CTCF occupancy of the *Igh* and *Igκ* loci[107]. Unexpectedly, we and other laboratories (Chapter 7 and Degner et al.[108]) found that the presence of CTCF is not lineage- or stage-specific, nor is it strictly required for *Ig* locus contraction and rearrangement per se. However, at the same time our experiments did reveal the importance of CTCF for proper *Igκ* locus rearrangement: CTCF mediates chromatin looping between κ locus regulatory elements and the $V_κ$ gene domain, thereby preventing κ enhancer promiscuity and ensuring the generation of a diverse $V_κ$ repertoire (Figure 6). Research from other laboratories investigating CTCF and Cohesin action at other antigen receptor loci largely confirmed our conclusions (reviewed by Choi and Feeney[109]).

One aspect of our analysis of V(D)J-recombination in CTCF-deficient B cells that puzzled us was the seemingly differential impact of *Ctcf* deletion on *Igh* and *Igκ* rearrangement. A virtually complete depletion of CTCF protein from pro-B cells did not prevent the generation of $Igμ^+$ pre-B cells that used both proximal and distal V_H segments. An initial low-resolution screening of V_H germline transcription and gene usage indicated no evident *Igh* recombination defects in CTCF-depleted pro-B cells, although recombination efficiencies appeared reduced (also see Discussion section in Chapter 7). As rearrangement defects of the *Igκ* locus appeared much more prominent, we focused our in-depth analysis on the latter. Nevertheless, we could not formally exclude a role for CTCF in *Igh* locus conformation: residual CTCF protein could still remain associated with the chromatin, or CTCF might already (pre-)establish *Igh* locus compaction very early after B cell commitment before efficient *mb1*-Cre deletion occurs. Degner et al.[108] provided a more thorough investigation of *Igh* locus topology under CTCF-depleting conditions and reported a modest reduction in locus contraction. Intriguingly, Guo et al.[110] showed that germline deletion of CTCF binding sites in the V_H-D_H intergenic region (a region analogous to the CTCF-occupied Sis/Cer region in the *Igκ* locus) resulted in *Igh* enhancer blocking defects that remarkably parallel those observed at the *Igκ* locus upon *Ctcf* (Chapter 7) or Sis/Cer[111,112] deletion. In conclusion, the gene regulatory and topology-organizing functions of CTCF are important for the generation of a diverse antibody repertoire – although other factors (e.g. PAX5 and YY1, both essential for *Igh* locus contraction[113,114] and found to interact with CTCF[115,116]) are required.

In B cells, the *Igh* and *Igκ* loci contain over 60 CTCF-binding sites that might influence the genomic architecture of these loci (Figure 6). The identification of PAX5-activated intergenic repeat (PAIR) elements co-occupied by CTCF in the *Igh* locus just upstream of distal V_H3609 genes[117], together with the observation that many proximal V_H gene segments are located within 100 bp of CTCF binding sites[118], suggests that proximity of a CTCF binding site may affect the probabilities of individual V_H genes to encounter a D_H-J_H element for recombination. With regard to CTCF occupancy, the *Igh* and *Igκ* loci are remarkably different: whereas approximately 36% of V_H genes have a CTCF binding site within 1 kb distance, this is the case for only 2% of $V_κ$ genes (Figure 6). Moreover, only 22% of $V_κ$ genes have a CTCF binding site within a genomic window of 5 kb, which is substantially higher for V_H genes (60%). Thus, although CTCF binds many sites throughout the *Igκ* locus, the majority of CTCF occupancy is located relatively far from $V_κ$ segments. It has

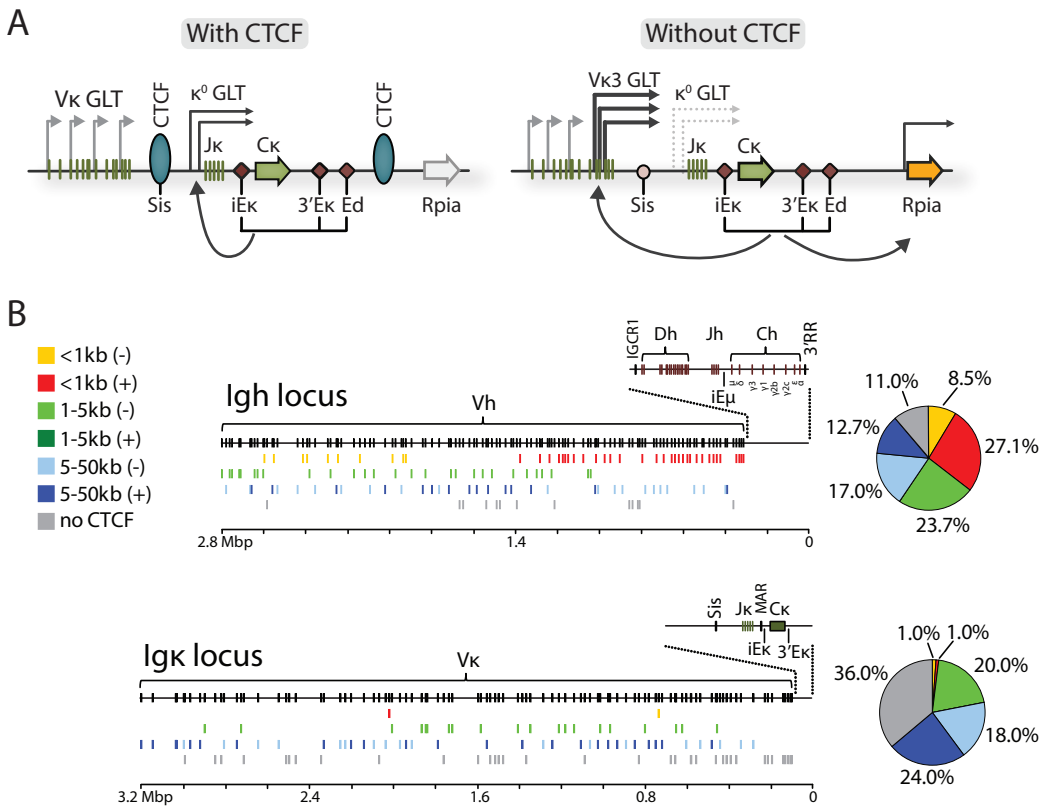


Figure 6. CTCF function at the *Igκ* locus and CTCF occupancy of the *Ig* loci in B cells. (A) Model of CTCF function at the *Igκ* locus. CTCF binding at the Sis element in the $V_κ$-$J_κ$ intergenic region is important for proper germline transcription (GLT) over the $J_κ$ region, limits proximal $V_κ$ recombination, and restricts κ enhancer (iEκ, 3'Eκ and Ed) interactions to the *Igκ* locus (left). Conditional deletion of *Ctcf* in the mouse (right) leads to increased proximal $V_κ$3-family transcription and rearrangement, reduced $J_κ$ GLT and increased expression of the neighbouring *Rpia* gene (orange arrow). (B) CTCF occupancy relative to functional V genes in the *Igh* (upper part) and the *Igκ* (lower part) loci. V genes were grouped according to the location of the nearest CTCF site (in kilo-basepairs, kb) and whether this site was located upstream (-) or downstream (+) relative to V gene transcriptional orientation. Genes were assigned to the 'no CTCF' group if no CTCF binding site was present in the intergenic regions up to their neighbouring genes or if the nearest binding site was >50kb away. V, D, J and C segments, as well as key regulatory elements are depicted in germline configuration at the top of each part; categorized color-coded V genes are aligned directly below. On the right, a pie chart represents the distribution of the different V gene groups expressed as percentages of the total number of V genes in the locus. A scale bar at the bottom of each part measures the size of the locus in mega-basepairs (Mbp).

been suggested that V regions are spatially organized as rosettes by CTCF, whereby CTCF binding adjacent to a V gene increases its recombination probability[118,119]. Because the *Igh* and *Igκ* loci show large differences in the proximity of CTCF binding sites to V genes, in such a model the 2 loci would be very differently organized in 3D nuclear space to provide appropriate access of individual V regions to the proposed recombination center[120]. Interestingly, it was recently proposed that E2A proteins might modulate *Igκ* locus topology by acting as anchors in a mechanism similar to that put forward for CTCF in the *Igh* locus[121]. This hypothesis still needs to be tested but is supported by the remarkable, non-random distribution of E2A binding sites across the *Igκ* locus, more specifically within 200 bp of the 5' or the 3' end of $V_κ$ regions[121]. Additionally, E2A binding to κ enhancers is required for efficient *Igκ* rearrangement[122]. Further support for a key role of E2A comes from our observation that E2A-bound $V_κ$ genes were more frequently used for

recombination in pre-B cells and more frequently involved in long-range interactions with κ enhancers (see Chapter 8 and discussion below).

Pre-BCR signaling: inducing TFs to 'focus' enhancers?

The assembly of a pre-BCR after productive *Igh* rearrangement is essential for the pro-to-pre B cell transition. Pre-BCR signaling triggers clonal expansion and subsequent cellular differentiation through several downstream pathways (see reviews by Hendriks and Middendorp[123] and Herzog et al.[124]). Signals from one of these pathways are relayed through the SLP65 adapter protein and the BTK enzyme, which function in a cooperative fashion to switch cell fate from proliferation to differentiation[123,125,126,127]. Maturation towards the small pre-B cell stage involves the activation of several key TFs by pre-BCR signals (including Foxo TFs[128], Ikaros[129,130], Aiolos and IRF4/8[131]; see Chapter 8). These proteins downregulate pre-BCR expression, re-activate the recombination machinery and induce *Ig* light chain locus accessibility, culminating in *Ig* light chain recombination[123,124]. TFs regulating *Ig* locus accessibility and recombination do so by interacting with local cis-regulatory elements, such as promoters, enhancers, or insulators[132,133]. The long-range chromatin interactions involved in this process are thought to be crucial for proper V(D)J recombination and orchestrate changes in subnuclear localisation, germline transcription, histone acetylation and/or methylation, DNA demethylation, and contraction of antigen receptor loci[106,134].

In spite of the well-established induction of *Igκ* rearrangement by pre-BCR signals, how pre-BCR signaling events affect *Igκ* locus accessibility in terms of contraction and topology had not been addressed experimentally. As described in Chapter 8, we made use of a series of knockout mouse strains to model a pre-BCR signaling gradient *in vivo*, comparing several aspects of the *Igκ* recombination process in pre-B cells obtained from these mice with wildtype pre-B cells (normal levels of pre-BCR signaling) and pro-B cells (no pre-BCR signaling). As expected, pre-BCR signaling induced the expression of TFs required for *Igκ* rearrangement and stimulated germline transcription of V_κ and J_κ promoters. However, experiments addressing the 3D organisation of the *Igκ* locus in B cell populations isolated from the different mouse strains revealed several novel aspects of enhancer-mediated regulation of *Igκ* locus topology.

After extensive 3C-Seq analysis of the V_κ region we realized, to our surprise, that long-range chromatin interactions between distal V_κ genes and the proximal J_κ/enhancer region were already present in pro-B cells and were not generally affected by reduced pre-BCR signaling activity. These observations suggest that full *Igκ* locus contraction is already achieved in pro-B cells - in marked contrast with a published microscopy study reporting that *Ig* loci are in a contracted state in rearranging cells only[135]. We validated our 3C-based findings using 3D DNA FISH experiments, and a recent Hi-C study by the Murre laboratory confirmed the existence of extensive long-range chromatin interactions in the *Igκ* locus of pro-B cells[136]. Interestingly, the phenomenon of *Igκ* contraction in pro-B cells could very well explain the low levels of *Igκ* rearrangements previously detected in these cells[125,137]. It is presently unclear what causes the discrepancy between our study and the one by Roldan et al.[135]. Minor deviations in culturing conditions (e.g. co-culturing with ST2 cells[135]) might explain these differences. Regardless, the *Igκ* locus contraction we observed in pro-B cells was not simply an artefact of our *in vitro* culturing conditions: 3D DNA FISH analysis on freshly isolated pro and pre B cells clearly showed *Igκ* locus contraction at both developmental stages (M.B. Rother and M.C. van Zelm, unpublished results). We therefore conclude that antigen receptor locus contraction does not necessarily correlate with ongoing rearrangement.

Although pro-B cells carried contracted *Igκ* loci and showed similar patterns of long-range interactions between the V_κ region and κ enhancers, we did observe marked differences in enhancer-mediated *Igκ* locus topology under conditions of reduced pre-BCR signaling activity. Together, we refer to this pre-BCR signaling-induced redistribution of chromatin interactions as 'enhancer focusing' (see Chapter 8 - Figure 7E for a graphical summary). *Igκ* enhancer focusing involved restricting κ enhancer interactions to the V_κ region, promoting coordinated interactions of both enhancers with V_κ genes and increasing enhancer interaction affinity for a subset of highly used V_κ genes often bound by E2A and Ikaros. We hypothesize that focusing of the κ enhancers is likely mediated by a set of specific TFs: Ikaros/Aiolos, E2A and IRF4. At the pro-to-pre B cell transition, expression of these factors is induced (Ikaros/Aiolos and IRF4) or maintained at high levels (E2A) by pre-BCR signaling. Increased TF binding to the κ enhancers and V_κ region results in coordinated interactions of the iEκ and 3'Eκ enhancers with the V_κ genes, inducing germline transcription and accessibility to the V(D)J recombinase. Although CTCF expression (and *Igκ* locus occupancy[107]) is not

pre-BCR dependent, it is conceivable that CTCF cooperates with pre-BCR induced TFs in the processes mentioned above. Supporting this notion is the partial phenotypic overlap between CTCF and pre-BCR signaling deficient pre-B cells: both show increased interactions of the κ enhancers with regions up- and downstream of the *Igκ* locus.

An important remaining question concerns the apparent differential TF requirement of *Igκ* locus contraction and enhancer focusing. Our combined 3D DNA FISH and 3C-Seq experiments clearly demonstrated that *Igκ* locus contraction does not require pre-BCR signaling and therefore κ enhancer focusing. It is thus unlikely that TFs induced by pre-BCR signals are involved in mediating *Igκ* locus contraction in pro-B cells. Notably, we observed that contraction was not observed in E2A-deficient pre-pro B cells, suggesting that E2A is required for this process. Similar observations were made using Hi-C experiments[136]. In support of this hypothesis, E2A was already recruited to the *Igκ* locus and its regulatory elements in pro-B cells[136]. The latter also holds true for CTCF[107]. Although its role in mediating *Igκ* locus contraction in pro-B cells has not been formally tested yet, CTCF does not appear to be required to maintain contraction at the pre-B cell stage (Chapter 7). The role of other ubiquitous genome architectural proteins such as Cohesin and Mediator[138,139], including any functional redundancy among the different architectural protein subclasses, has not yet been addressed in this context. Considering their general importance for genome topology, it is plausible that these proteins are also important for the establishment of *Igκ* locus contraction. We therefore propose that contraction in pro-B cells is driven by E2A and is likely to involve the actions of architectural proteins. Pre-BCR induced TFs act upon this 'topological scaffold' to further refine it.

In summary, research described in Chapter 8 shows that pre-BCR signals relayed through BTK and SLP65 create a chromatin environment that facilitates proper *Igκ* locus recombination. The observed redistribution of enhancer-mediated chromatin interactions induced by pre-BCR signaling immediately suggests that these phenomena have functional consequences for the *Igκ* recombination process. Future studies will have to put our model to the test. Interesting angles to pursue could involve V_κ repertoire sequencing in $Btk^{-/-}$;$Slp65^{-/-}$ pre-B cells: is there indeed a skewed V_κ usage pattern under conditions of low pre-BCR signaling?; targeted deletions of E2A/Ikaros binding motifs near specific (highly used) V_κ genes: would this affect the rearrangement frequencies of this gene?; or the conditional ablation of E2A/Ikaros specifically at the pre-B cell stage: does this result in a loss of enhancer focusing and altered V_κ usage patterns? At present, it remains to be shown whether enhancer focusing as observed here at the *Igκ* locus by signal-responsive TFs is a more generally occurring phenomenon. We believe this to be a possibility, as the modulation of pre-existing chromatin interactions in a developmentally regulated fashion has been previously reported, e.g. at the *Hox* and *Shh* loci[140]. Pre-formed 'permissive' chromatin structures such as the contracted *Igκ* locus in pro-B cells could be beneficial by enabling cells to more rapidly respond to differentiation signals, as exemplified by pre-BCR signaling in the case of pro-B cells.

Concluding remarks

A key aspect of genome biology is the proper regulation of gene expression, which is found perturbed in many diseases and disorders. Here we conducted studies on differentiating hematopoietic cells, with the aim to generate new knowledge of the regulatory mechanisms operated by TFs and the impact of these processes on cellular differentiation. Our results provide a collection of novel insights into the biology of hematopoietic TF complexes and how they govern complex cellular processes such as the regulation of gene transcription and V(D)J recombination. Additionally, we have also developed and implemented new tools to study TF actions. Our main accomplishments are summarized below:

1) Identification of TFs that prevent premature gene activation during red blood cell development (Chapter 2)
2) Adaptation of microarray-based 4C technology to the Illumina next-generation sequencing platform ('3C-Seq') and the development of a bioinformatics toolkit for data analysis (Chapters 3 and 4)
3) Mapping and functional characterization of regulatory elements that control *Myb* proto-oncogene expression during erythroid differentiation (Chapter 5)
4) Revealing how common genetic variants associated with a plethora of clinically important erythroid traits interfere with *MYB* activation via distal enhancers (Chapter 6)

5) Demonstrating that the CTCF insulator protein is essential for early B cell development and the generation of a diverse antibody repertoire (Chapter 7)
6) Using an *in vivo* pre-BCR signaling gradient to show that pre-BCR signals act upon a pre-contracted *Igκ* locus to induce 'enhancer focusing' and proper *Igκ* locus recombination (Chapter 8)

Future studies are bound to answer the remaining and new questions that emerged from our work, as scientists work towards a common goal of fully understanding the inner workings of our genetic material. As our understanding advances, I believe we will see the fruits of this 'genomic labour' being applied to improve daily medical practice. It is encouraging to see that the first genomics-based diagnostic and treatment protocols are currently finding their way into the clinic[141-147].

References

1 Quackenbush, J. *The human genome : the book of essential knowledge*. (Imagine, 2011).
2 Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187-197, doi:10.1038/nature09792 (2011).
3 Evans, J. P., Meslin, E. M., Marteau, T. M. & Caulfield, T. Genomics. Deflating the genomic bubble. *Science* **331**, 861-862, doi:10.1126/science.1198039 (2011).
4 Marshall, E. Human genome 10th anniversary. Waiting for the revolution. *Science* **331**, 526-529, doi:10.1126/science.331.6017.526 (2011).
5 Collins, F. Has the revolution arrived? *Nature* **464**, 674-675, doi:10.1038/464674a (2010).
6 Mattick, J. S. Genome-sequencing anniversary. The genomic foundation is shifting. *Science* **331**, 874, doi:10.1126/science.1203703 (2011).
7 Consortium, E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640, doi:10.1126/science.1105136 (2004).
8 Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685-690, doi:10.1038/35055500 (2001).
9 Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251, doi:10.1016/j.cell.2013.02.014 (2013).
10 Soler, E. *et al.* The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes & development* **24**, 277-289, doi:10.1101/gad.551810 (2010).
11 Li, L. *et al.* Ldb1-nucleated transcription complexes function as primary mediators of global erythroid gene activation. *Blood* **121**, 4575-4585, doi:10.1182/blood-2013-01-479451 (2013).
12 Goardon, N. *et al.* ETO2 coordinates cellular proliferation and differentiation during erythropoiesis. *The EMBO journal* **25**, 357-366, doi:10.1038/sj.emboj.7600934 (2006).
13 Meier, N. *et al.* Novel binding partners of Ldb1 are required for haematopoietic development. *Development* **133**, 4913-4923, doi:10.1242/dev.02656 (2006).
14 Schuh, A. H. *et al.* ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Molecular and cellular biology* **25**, 10235-10250, doi:10.1128/MCB.25.23.10235-10250.2005 (2005).
15 Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136-3147, doi:10.1182/blood-2004-04-1603 (2004).
16 Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Molecular cell* **36**, 667-681, doi:10.1016/j.molcel.2009.11.001 (2009).
17 Chyla, B. J. *et al.* Deletion of Mtg16, a target of t(16;21), alters hematopoietic progenitor cell proliferation and lineage allocation. *Molecular and cellular biology* **28**, 6234-6247, doi:10.1128/MCB.00404-08 (2008).
18 Childs, K. S. & Goodbourn, S. Identification of novel co-repressor molecules for Interferon Regulatory Factor-2. *Nucleic acids research* **31**, 3016-3026 (2003).
19 Matsuyama, T. *et al.* Targeted disruption of IRF-1 or IRF-2 results in abnormal type I IFN gene induction and aberrant lymphocyte development. *Cell* **75**, 83-97 (1993).
20 Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Developmental cell* **23**, 796-811, doi:10.1016/j.devcel.2012.09.003 (2012).
21 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).
22 de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26**, 11-24, doi:10.1101/gad.179804.111 (2012).
23 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354, doi:10.1038/ng1896 (2006).
24 van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods* **9**, 969-972, doi:10.1038/nmeth.2173 (2012).
25 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 996-1001, doi:10.1073/pnas.1317788111 (2014).
26 Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334**, 222-225, doi:10.1126/science.1207194 (2011).
27 Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-375, doi:10.1038/nature13138 (2014).
28 Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics* **46**, 136-143, doi:10.1038/ng.2870 (2014).

29 Daniel, B. *et al.* The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes & development* **28**, 1562-1577, doi:10.1101/gad.242685.114 (2014).
30 Mucenski, M. L. *et al.* A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. *Cell* **65**, 677-689 (1991).
31 Ramsay, R. G. & Gonda, T. J. MYB function in normal and cancer cells. *Nature reviews. Cancer* **8**, 523-534, doi:10.1038/nrc2439 (2008).
32 Mylona, A. *et al.* Genome-wide analysis shows that Ldb1 controls essential hematopoietic genes/pathways in mouse early development and reveals novel players in hematopoiesis. *Blood* **121**, 2902-2913, doi:10.1182/blood-2012-11-467654 (2013).
33 Li, L. *et al.* Nuclear adaptor Ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. *Nature immunology* **12**, 129-136, doi:10.1038/ni.1978 (2011).
34 Zhang, J., Markus, J., Bies, J., Paul, T. & Wolff, L. Three murine leukemia virus integration regions within 100 kilobases upstream of c-myb are proximal to the 5' regulatory region of the gene through DNA looping. *Journal of virology* **86**, 10524-10532, doi:10.1128/JVI.01077-12 (2012).
35 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).
36 de Laat, W. & Grosveld, F. Spatial organization of gene expression: the active chromatin hub. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **11**, 447-459 (2003).
37 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).
38 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
39 Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17921-17926, doi:10.1073/pnas.1317023110 (2013).
40 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).
41 Delmore, J. E. *et al.* BET bromodomain inhibition as a therapeutic strategy to target c-Myc. *Cell* **146**, 904-917, doi:10.1016/j.cell.2011.08.017 (2011).
42 Zuber, J. *et al.* RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478**, 524-528, doi:10.1038/nature10334 (2011).
43 Mertz, J. A. *et al.* Targeting MYC dependence in cancer by inhibiting BET bromodomains. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 16669-16674, doi:10.1073/pnas.1108190108 (2011).
44 Lahortiga, I. *et al.* Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nature genetics* **39**, 593-595, doi:10.1038/ng2025 (2007).
45 Zuber, J. *et al.* An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. *Genes & development* **25**, 1628-1640, doi:10.1101/gad.17269211 (2011).
46 Schmidt, M., Nazarov, V., Stevens, L., Watson, R. & Wolff, L. Regulation of the resident chromosomal copy of c-myc by c-Myb is involved in myeloid leukemogenesis. *Molecular and cellular biology* **20**, 1970-1981 (2000).
47 Love, P. E., Warzecha, C. & Li, L. Ldb1 complexes: the new master regulators of erythroid gene transcription. *Trends in genetics : TIG* **30**, 1-9, doi:10.1016/j.tig.2013.10.001 (2014).
48 Papadopoulos, G. L. *et al.* GATA-1 genome-wide occupancy associates with distinct epigenetic profiles in mouse fetal liver erythropoiesis. *Nucleic acids research* **41**, 4938-4948, doi:10.1093/nar/gkt167 (2013).
49 Saleque, S., Cameron, S. & Orkin, S. H. The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. *Genes & development* **16**, 301-306, doi:10.1101/gad.959102 (2002).
50 Saleque, S., Kim, J., Rooke, H. M. & Orkin, S. H. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Molecular cell* **27**, 562-572, doi:10.1016/j.molcel.2007.06.039 (2007).
51 Garcon, L. *et al.* Gfi-1B plays a critical role in terminal differentiation of normal and transformed erythroid progenitor cells. *Blood* **105**, 1448-1455, doi:10.1182/blood-2003-11-4068 (2005).
52 Kerenyi, M. A. *et al.* Histone demethylase Lsd1 represses hematopoietic stem and progenitor cell signatures during blood cell maturation. *eLife* **2**, e00633, doi:10.7554/eLife.00633 (2013).
53 van Riel, B. *et al.* A novel complex, RUNX1-MYEF2, represses hematopoietic genes in erythroid cells. *Molecular and cellular biology* **32**, 3814-3822, doi:10.1128/MCB.05938-11 (2012).
54 Remacle, J. E. *et al.* New mode of DNA binding of multi-zinc finger transcription factors: deltaEF1 family members bind with two hands to two target sites. *The EMBO journal* **18**, 5073-5084, doi:10.1093/emboj/18.18.5073 (1999).
55 Verschueren, K. *et al.* SIP1, a novel zinc finger/homeodomain repressor, interacts with Smad proteins and binds to 5'-CACCT sequences in candidate target genes. *The Journal of biological chemistry* **274**, 20489-20498 (1999).
56 Sekido, R. *et al.* The delta-crystallin enhancer-binding protein delta EF1 is a repressor of E2-box-mediated gene activation. *Molecular and cellular biology* **14**, 5692-5700 (1994).
57 Comijn, J. *et al.* The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion. *Molecular cell* **7**, 1267-1278 (2001).
58 Nishimura, G. *et al.* DeltaEF1 mediates TGF-beta signaling in vascular smooth muscle cell differentiation. *Developmental cell* **11**, 93-104, doi:10.1016/j.devcel.2006.05.011 (2006).
59 Zermati, Y. *et al.* Transforming growth factor inhibits erythropoiesis by blocking proliferation and accelerating differentiation of erythroid progenitors. *Experimental hematology* **28**, 885-894 (2000).
60 Goossens, S. *et al.* The EMT regulator Zeb2/Sip1 is essential for murine embryonic hematopoietic stem/progenitor cell differentiation and mobilization. *Blood* **117**, 5620-5630, doi:10.1182/blood-2010-08-300236 (2011).
61 Higashi, Y. *et al.* Impairment of T cell development in deltaEF1 mutant mice. *The Journal of experimental medicine* **185**, 1467-1479 (1997).
62 Wang, J. *et al.* Opposing LSD1 complexes function in developmental gene activation and repression programmes. *Nature* **446**, 882-887, doi:10.1038/nature05671 (2007).
63 Bender, T. P., Thompson, C. B. & Kuehl, W. M. Differential expression of c-myb mRNA in murine B lymphomas by a block to

transcription elongation. *Science* **237**, 1473-1476 (1987).
64 Watson, R. J. A transcriptional arrest mechanism involved in controlling constitutive levels of mouse c-myb mRNA. *Oncogene* **2**, 267-272 (1988).
65 Sawado, T., Halow, J., Bender, M. A. & Groudine, M. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes & development* **17**, 1009-1018, doi:10.1101/gad.1072303 (2003).
66 Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**, 1581-1595, doi:10.1016/j.cell.2013.10.056 (2013).
67 Wang, L. G., Liu, X. M., Li, Z. R., Denstman, S. & Bloch, A. Differential binding of nuclear c-ets-1 protein to an intron I fragment of the c-myb gene in growth versus differentiation. *Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research* **5**, 1243-1251 (1994).
68 Dooley, S., Seib, T., Welter, C. & Blin, N. c-myb intron I protein binding and association with transcriptional activity in leukemic cells. *Leukemia research* **20**, 429-439 (1996).
69 Suhasini, M. & Pilz, R. B. Transcriptional elongation of c-myb is regulated by NF-kappaB (p50/RelB). *Oncogene* **18**, 7360-7369, doi:10.1038/sj.onc.1203158 (1999).
70 Drabsch, Y. *et al.* Mechanism of and requirement for estrogen-regulated MYB expression in estrogen-receptor-positive breast cancer cells. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 13762-13767, doi:10.1073/pnas.0700104104 (2007).
71 Mitra, P., Pereira, L. A., Drabsch, Y., Ramsay, R. G. & Gonda, T. J. Estrogen receptor-alpha recruits P-TEFb to overcome transcriptional pausing in intron 1 of the MYB gene. *Nucleic acids research* **40**, 5988-6000, doi:10.1093/nar/gks286 (2012).
72 Paredes, S. H., Melgar, M. F. & Sethupathy, P. Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index. *Bioinformatics* **29**, 1485-1487, doi:10.1093/bioinformatics/bts596 (2013).
73 Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74-79, doi:10.1038/nature10442 (2011).
74 Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233-1244, doi:10.1016/j.cell.2012.03.051 (2012).
75 Krivega, I., Dale, R. K. & Dean, A. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes & development*, doi:10.1101/gad.239749.114 (2014).
76 Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development* **20**, 2349-2354, doi:10.1101/gad.399506 (2006).
77 Gaj, T., Gersbach, C. A. & Barbas, C. F., 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology* **31**, 397-405, doi:10.1016/j.tibtech.2013.04.004 (2013).
78 Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490-493, doi:10.1038/nature09158 (2010).
79 Xiong, N., Kang, C. & Raulet, D. H. Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity* **16**, 453-463 (2002).
80 Copeland, N. G., Jenkins, N. A. & Court, D. L. Recombineering: a powerful new tool for mouse functional genomics. *Nature reviews. Genetics* **2**, 769-779, doi:10.1038/35093556 (2001).
81 Thein, S. L. *et al.* Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11346-11351, doi:10.1073/pnas.0611393104 (2007).
82 Andrews, N. C. Genes determining blood cell traits. *Nature genetics* **41**, 1161-1162, doi:10.1038/ng1109-1161 (2009).
83 Craig, J. E. *et al.* Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nature genetics* **12**, 58-64, doi:10.1038/ng0196-58 (1996).
84 Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nature genetics* **39**, 1197-1199, doi:10.1038/ng2108 (2007).
85 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).
86 Sankaran, V. G. *et al.* Rare complete loss of function provides insight into a pleiotropic genome-wide association study locus. *Blood* **122**, 3845-3847, doi:10.1182/blood-2013-09-528315 (2013).
87 Freedman, M. L. *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics* **43**, 513-518, doi:10.1038/ng.840 (2011).
88 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).
89 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).
90 Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507-1520, doi:10.1016/j.cell.2013.11.039 (2013).
91 Gorkin, D. U. & Ren, B. Genetics: Closing the distance on obesity culprits. *Nature* **507**, 309-310, doi:10.1038/nature13212 (2014).
92 Stamatoyannopoulos, G. Control of globin gene expression during development and erythroid differentiation. *Experimental hematology* **33**, 259-271, doi:10.1016/j.exphem.2004.11.007 (2005).
93 Sankaran, V. G. & Orkin, S. H. The switch from fetal to adult hemoglobin. *Cold Spring Harbor perspectives in medicine* **3**, a011643, doi:10.1101/cshperspect.a011643 (2013).
94 Koehler, A. N. A complex task? Direct modulation of transcription factors with small molecules. *Current opinion in chemical biology* **14**, 331-340, doi:10.1016/j.cbpa.2010.03.022 (2010).
95 Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253-257, doi:10.1126/science.1242088 (2013).
96 Bianchi, E. *et al.* c-myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression. *Blood* **116**, e99-110, doi:10.1182/blood-2009-08-238311 (2010).
97 Sankaran, V. G. *et al.* MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13.

Proceedings of the National Academy of Sciences of the United States of America **108**, 1519-1524, doi:10.1073/pnas.1018384108 (2011).
98 Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
99 Jiang, J. *et al.* cMYB is involved in the regulation of fetal hemoglobin production in adults. *Blood* **108**, 1077-1083, doi:10.1182/blood-2006-01-008912 (2006).
100 Menzel, S. *et al.* The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans. *Blood* **110**, 3624-3626, doi:10.1182/blood-2007-05-093419 (2007).
101 Sankaran, V. G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes & development* **26**, 2075-2087, doi:10.1101/gad.197020.112 (2012).
102 Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics* **15**, 234-246, doi:10.1038/nrg3663 (2014).
103 Holwerda, S. J. & de Laat, W. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120369, doi:10.1098/rstb.2012.0369 (2013).
104 Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801, doi:10.1038/nature06634 (2008).
105 Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433, doi:10.1016/j.cell.2008.01.011 (2008).
106 Jhunjhunwala, S., van Zelm, M. C., Peak, M. M. & Murre, C. Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**, 435-448, doi:10.1016/j.cell.2009.07.016 (2009).
107 Degner, S. C., Wong, T. P., Jankevicius, G. & Feeney, A. J. Cutting edge: developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development. *Journal of immunology* **182**, 44-48 (2009).
108 Degner, S. C. *et al.* CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 9566-9571, doi:10.1073/pnas.1019391108 (2011).
109 Choi, N. M. & Feeney, A. J. CTCF and ncRNA Regulate the Three-Dimensional Structure of Antigen Receptor Loci to Facilitate V(D)J Recombination. *Frontiers in immunology* **5**, 49, doi:10.3389/fimmu.2014.00049 (2014).
110 Guo, C. *et al.* CTCF-binding elements mediate control of V(D)J recombination. *Nature* **477**, 424-430, doi:10.1038/nature10495 (2011).
111 Xiang, Y., Zhou, X., Hewitt, S. L., Skok, J. A. & Garrard, W. T. A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. *Journal of immunology* **186**, 5356-5366, doi:10.4049/jimmunol.1003794 (2011).
112 Xiang, Y., Park, S. K. & Garrard, W. T. Vkappa gene repertoire and locus contraction are specified by critical DNase I hypersensitive sites within the Vkappa-Jkappa intervening region. *Journal of immunology* **190**, 1819-1826, doi:10.4049/jimmunol.1203127 (2013).
113 Fuxa, M. *et al.* Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes & development* **18**, 411-422, doi:10.1101/gad.291504 (2004).
114 Liu, H. *et al.* Yin Yang 1 is a critical regulator of B-cell development. *Genes & development* **21**, 1179-1189, doi:10.1101/gad.1529307 (2007).
115 Donohoe, M. E., Zhang, L. F., Xu, N., Shi, Y. & Lee, J. T. Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Molecular cell* **25**, 43-56, doi:10.1016/j.molcel.2006.11.017 (2007).
116 Medvedovic, J. *et al.* Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire. *Immunity* **39**, 229-244, doi:10.1016/j.immuni.2013.08.011 (2013).
117 Ebert, A. *et al.* The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* **34**, 175-187, doi:10.1016/j.immuni.2011.02.005 (2011).
118 Lucas, J. S., Bossen, C. & Murre, C. Transcription and recombination factories: common features? *Current opinion in cell biology* **23**, 318-324, doi:10.1016/j.ceb.2010.11.007 (2011).
119 Jhunjhunwala, S. *et al.* The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**, 265-279, doi:10.1016/j.cell.2008.03.024 (2008).
120 Ji, Y. *et al.* The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell* **141**, 419-431, doi:10.1016/j.cell.2010.03.010 (2010).
121 Bossen, C., Mansson, R. & Murre, C. Chromatin topology and the regulation of antigen receptor assembly. *Annual review of immunology* **30**, 337-356, doi:10.1146/annurev-immunol-020711-075003 (2012).
122 Inlay, M. A., Tian, H., Lin, T. & Xu, Y. Important roles for E protein binding sites within the immunoglobulin kappa chain intronic enhancer in activating Vkappa Jkappa rearrangement. *The Journal of experimental medicine* **200**, 1205-1211, doi:10.1084/jem.20041135 (2004).
123 Hendriks, R. W. & Middendorp, S. The pre-BCR checkpoint as a cell-autonomous proliferation switch. *Trends in immunology* **25**, 249-256, doi:10.1016/j.it.2004.02.011 (2004).
124 Herzog, S., Reth, M. & Jumaa, H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. *Nature reviews. Immunology* **9**, 195-205, doi:10.1038/nri2491 (2009).
125 Kersseboom, R. *et al.* Bruton's tyrosine kinase and SLP-65 regulate pre-B cell differentiation and the induction of Ig light chain gene rearrangement. *Journal of immunology* **176**, 4543-4552 (2006).
126 Jumaa, H., Mitterer, M., Reth, M. & Nielsen, P. J. The absence of SLP65 and Btk blocks B cell development at the preB cell receptor-positive stage. *European journal of immunology* **31**, 2164-2169, doi:10.1002/1521-4141(200107)31:7<2164::AID-IMMU2164>3.0.CO;2-S (2001).
127 Kersseboom, R. *et al.* Bruton's tyrosine kinase cooperates with the B cell linker protein SLP-65 as a tumor suppressor in Pre-B cells. *The Journal of experimental medicine* **198**, 91-98, doi:10.1084/jem.20030615 (2003).
128 Herzog, S. *et al.* SLP-65 regulates immunoglobulin light chain gene recombination through the PI(3)K-PKB-Foxo pathway. *Nature immunology* **9**, 623-631, doi:10.1038/ni.1616 (2008).
129 Heizmann, B., Kastner, P. & Chan, S. Ikaros is absolutely required for pre-B cell differentiation by attenuating IL-7 signals. *The*

Journal of experimental medicine **210**, 2823-2832, doi:10.1084/jem.20131735 (2013).
130 Schwickert, T. A. *et al.* Stage-specific control of early B cell development by the transcription factor Ikaros. *Nature immunology* **15**, 283-293, doi:10.1038/ni.2828 (2014).
131 Ma, S., Pathak, S., Trinh, L. & Lu, R. Interferon regulatory factors 4 and 8 induce the expression of Ikaros and Aiolos to down-regulate pre-B-cell receptor and promote cell-cycle withdrawal in pre-B-cell development. *Blood* **111**, 1396-1403, doi:10.1182/blood-2007-08-110106 (2008).
132 Perlot, T. & Alt, F. W. Cis-regulatory elements and epigenetic changes control genomic rearrangements of the IgH locus. *Advances in immunology* **99**, 1-32, doi:10.1016/S0065-2776(08)00601-9 (2008).
133 Cobb, R. M., Oestreich, K. J., Osipovich, O. A. & Oltz, E. M. Accessibility control of V(D)J recombination. *Advances in immunology* **91**, 45-109, doi:10.1016/S0065-2776(06)91002-5 (2006).
134 Seitan, V. C., Krangel, M. S. & Merkenschlager, M. Cohesin, CTCF and lymphocyte antigen receptor locus rearrangement. *Trends in immunology* **33**, 153-159, doi:10.1016/j.it.2012.02.004 (2012).
135 Roldan, E. *et al.* Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. *Nature immunology* **6**, 31-41, doi:10.1038/ni1150 (2005).
136 Lin, Y. C. *et al.* Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology* **13**, 1196-1204, doi:10.1038/ni.2432 (2012).
137 Novobrantseva, T. I. *et al.* Rearrangement and expression of immunoglobulin light chain genes can precede heavy chain expression during normal B cell development in mice. *The Journal of experimental medicine* **189**, 75-88 (1999).
138 Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).
139 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).
140 de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506, doi:10.1038/nature12753 (2013).
141 Yu, Y., Wu, B. L., Wu, J. & Shen, Y. Exome and whole-genome sequencing as clinical tests: a transformative practice in molecular diagnostics. *Clinical chemistry* **58**, 1507-1509, doi:10.1373/clinchem.2012.193128 (2012).
142 Pellini, M. Q&A: Michael Pellini on cancer diagnostics. Interview by Eric Bender. *Cancer discovery* **2**, 382, doi:10.1158/2159-8290.CD-ND2012-021 (2012).
143 Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-308, doi:10.1038/nature12981 (2014).
144 Dewey, F. E. *et al.* Clinical interpretation and implications of whole-genome sequencing. *JAMA : the journal of the American Medical Association* **311**, 1035-1045, doi:10.1001/jama.2014.1717 (2014).
145 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *The New England journal of medicine* **369**, 1502-1511, doi:10.1056/NEJMoa1306555 (2013).
146 Pirmohamed, M. Personalized Pharmacogenomics: Predicting Efficacy and Adverse Drug Reactions. *Annual review of genomics and human genetics*, doi:10.1146/annurev-genom-090413-025419 (2013).
147 Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. *Science* **344**, 641-645, doi:10.1126/science.1251102 (2014).

Summary

Complex developmental processes that involve cell proliferation and differentiation are orchestrated at the level of gene expression. Therefore, tight regulation of gene expression needs to be implemented and maintained to ensure proper embryogenesis as well as adult tissue homeostasis. Two important aspects of gene regulation were intensively studied in this thesis: 1) the actions of specialized gene regulatory proteins called transcription factors (TFs) and 2) how these TFs control cis regulatory elements, in particular enhancers, to establish chromatin looping and modulate gene expression.

In **Chapter 2** we present the identification and characterization of TFs that prevent premature gene activation in red blood cell progenitors. In these progenitor cells, late erythroid-specific genes are already bound by the LDB1 TF complex responsible for their activation upon differentiation. However, how gene activation by the LDB1-complex is suppressed in erythroid progenitors is still poorly understood. We show that the ETO2 and IRF2BP2 corepressors are recruited to these erythroid LDB1-complex target genes and cooperate to maintain them in a poised state until activation is required. Moreover, we could demonstrate that IRF2BP2 is important for erythropoiesis *in vivo*. These experiments establish new regulatory mechanisms and proteins that orchestrate erythroid development.

Regulatory elements can be located at large distances from their cognate target genes, complicated their functional characterization. In **Chapters 3 and 4**, we describe multiplexed 3C-Seq technology and the r3C-Seq data analysis pipeline, which offer a straightforward set of methods for the semi high-throughput characterization of long-range chromatin interactions. Together, these tools facilitate the discovery of new regulatory connections between genes and distal regulatory elements.

We next used 3C-Seq to characterize the potential regulatory function of an intergenic region between the *Hbs1l* and *Myb* genes, the latter encoding a TF (called c-Myb) crucial for hematopoietic development and leukemogenesis. *Myb* is furthermore an interesting potential therapeutic target, as reducing its expression levels results in the accumulation of fetal hemoglobin (HbF) in human red blood cells - a favorable trait that significantly ameliorates β-hemoglobinopathy disease severity. Our studies described in **Chapters 5 and 6** show that in red blood cells the *Myb-Hbs1l* intergenic region contains long-range *Myb* enhancer elements. These enhancers are operated by key erythroid TFs, and local genetic variants common among humans were shown to have a negative impact on enhancer function and *MYB* levels. As a consequence, people bearing these specific polymorphisms have reduced erythroid *MYB* expression and higher HbF levels, suggesting that modulation of intergenic *MYB* enhancer activity could provide a new therapeutic strategy to treat β-hemoglobinopathies.

We also applied 3C-seq to developing B cell populations and used it to describe the three-dimensional (3D) conformation of the immunoglobulin (*Ig*) κ light chain locus (**Chapters 7 and 8**). A contracted 3D organization of the *Ig* loci is an important requirement for the generation of a broad antibody repertoire to fight invading pathogens. *Ig* locus topology is organized by TFs and powerful regulatory elements. We show that CTCF, a protein involved in 3D genome organization, allows the powerful *Igκ* enhancers to correctly explore the 3 Mb $V_{κ}$ region by restricting enhancer interactions with the most proximal $V_{κ}$ genes and regions outside the locus (**Chapter 7**). This mechanism ensures the random usage of $V_{κ}$ gene segments and proper antibody diversity as a consequence. We also analyzed the effect of pre-B cell receptor (pre-BCR) signals on *Igκ* locus topology using a series of knockout mouse strains mimicking an *in vivo* pre-BCR signaling gradient (**Chapter 8**). Surprisingly, the *Igκ* locus was already fully contracted independent of pre-BCR signaling. Instead, pre-BCR signals were found to act upon a pre-contracted *Igκ* locus to refine the long-range chromatin interactions directed by the κ

enhancers, a process we call 'enhancer focusing'. We postulate that pre-BCR signaling, through the activation of key TFs, increases *Igκ* locus accessibility to the recombination machinery by a functional redistribution of enhancer-mediated chromatin interactions.

In summary, this thesis contains a collection of novel insights into the biology of hematopoietic TF complexes and how they govern complex cellular processes such as the regulation of gene transcription and *Ig* locus recombination. In **Chapter 9** I discuss the implications of the experiments described here for our understanding of (hematopoietic) gene regulatory mechanisms and how these findings could contribute to the development of new therapeutic approaches for human diseases.

Samenvatting

Complexe ontwikkelingsprocessen waarbij cel proliferatie en differentiatie betrokken zijn worden op het niveau van genexpressie aangestuurd. Daarom moet een strakke regulatie van genexpressie toegepast en gehandhaafd worden om zo een adequate embryogenese en homeostase van volwassen weefsels te kunnen garanderen. Twee belangrijke aspecten van genregulatie zijn intensief bestudeerd in dit proefschrift: 1) de handelingen van gespecialiseerde genregulator eiwitten die transcriptiefactoren (TFen) worden genoemd, en 2) hoe deze TFen cis regulatoire elementen, in het bijzonder enhancers, aansturen om chromatine looping tot stand te laten komen en genexpressie te moduleren.

In **Hoofdstuk 2** beschrijven we de identificatie en karakterisatie van TFen die vroegtijdige gen activatie in rode voorloperbloedcellen voorkomen. In deze voorlopercellen zijn de late erythroïd-specifieke genen al gebonden door het LDB1 TF complex dat verantwoordelijk is voor hun activatie tijdens differentiatie. Echter, hoe gen activatie door het LDB1-complex wordt onderdrukt in rode voorloperbloedcellen is niet geheel duidelijk. Wij tonen aan dat de ETO2 en IRF2BP2 corepressor eiwitten de LDB1-complex doelwitgenen binden en samenwerken om ze in een toestand te brengen die we 'poised' noemen: de genen staan klaar om direct geactiveerd te kunnen worden indien nodig. Daarbij konden we laten zien dat IRF2BP2 belangrijk is voor rode bloedcel vorming *in vivo*. Deze experimenten hebben nieuwe regulatoire mechanismen en eiwitten blootgelegd die de erythroïde ontwikkeling in goede banen leiden.

Regulatoire elementen kunnen zich op grote afstand van hun doelwitgenen bevinden, hetgeen een functionele karakterisatie bemoeilijkt. In **Hoofdstuk 3 en 4** beschrijven we gemultiplexte 3C-Seq technologie en het r3C-Seq data analyse software pakket, die samen een relatief eenvoudige methode bieden voor de snelle karakterisatie van connecties tussen genen en regulatoire elementen op lange afstand.

Vervolgens hebben we 3C-Seq gebruikt om de potentiële regulatoire functie van het intergene gebied tussen de *Hbs1l* en *Myb* genen te onderzoeken. *Myb* codeert voor de c-Myb TF, die cruciaal is voor de hematopoïetische ontwikkeling en het ontstaan van leukemie. Verder is *Myb* een interessant potentieel therapeutisch doelwit, omdat een reductie van *Myb* expressie resulteert in de aanmaak van foetaal hemoglobin (HbF) in humane rode bloedcellen, hetgeen een zeer gunstig effect heeft op het ziektebeeld van de β-hemoglobinopathieën. Onze studies beschreven in **Hoofdstuk 5 en 6** laten zien dat in rode bloedcellen de *Myb-Hbs1l* intergene regio enhancer elementen bevat die *Myb* over lange afstand activeren. Deze enhancers worden aangestuurd door belangrijke erythroïde TFen, en van lokale (onder mensen veelvoorkomende) genetische variatie kon worden aangetoond dat het een negatief effect heeft op de enhancer functie en *MYB* genexpressie. Als gevolg hebben mensen die drager zijn van deze specifieke varianten verlaagde *MYB* expressie en hogere HbF niveaus, suggererende dat modulatie van intergene *MYB* enhancer activiteit een nieuwe therapeutische strategie kan zijn om β-hemoglobinopathieën te behandelen.

We hebben 3C-Seq ook toegepast op zich ontwikkelende B cel populaties en het gebruikt om de driedimensionale (3D) vouwing van het immunoglobuline (*Ig*) κ lichte keten gebied in kaart te brengen (**Hoofdstuk 7 en 8**). Een samengetrokken 3D organisatie van de *Ig* gebieden is een belangrijke vereiste voor de ontwikkeling van een gevarieerd antilichaam repertoire om effectief infecties te kunnen bestrijden. De 3D vouwing van *Ig* gebieden wordt bewerkstelligd door TFen en krachtige regulatoire elementen. We laten zien dat CTCF, een eiwit betrokken bij het 3D organiseren van het genoom, de krachtige *Igκ* enhancers toestaat het 3 Mb grote V_{κ} gebied op de juiste manier ruimtelijk te verkennen door interacties met de meest proximale V_{κ} genen en regio's buiten het *Ig* gebied te beperken (**Hoofdstuk 7**). Dit mechanisme verzekert een nagenoeg willekeurig gebruik van V_{κ} gen segmenten met voldoende antilichaam

diversiteit als gevolg. We hebben ook het effect van pre-B cel receptor (pre-BCR) signalering op de 3D vouwing van het *Igκ* locus geanalyseerd met behulp van een serie knockout muis lijnen die een *in vivo* pre-BCR signalering gradiënt nabootsen (**Hoofdstuk 8**). Verrassend genoeg bleek het *Igκ* gebied al volledig samengetrokken voor de activatie van pre-BCR signalering. De pre-BCR signalen zelf zorgden voor een optimalisatie van het al samengetrokken *Igκ* gebied door de lokale chromatine interacties over lange afstand, die gedirigeerd worden door de κ enhancers, te verfijnen, een proces dat we 'enhancer focusing' hebben genoemd. Wij postuleren dat pre-BCR signalering, via de activatie van belangrijke TFen, de toegankelijkheid van het *Igκ* gebied voor het recombinatie mechanisme vergroot door een functionele re-distributie van enhancer-gestuurde chromatine interacties.

Samengevat bevat dit proefschrift een verzameling van nieuwe inzichten in de biologie van hematopoïetische TF complexen en hoe zij complexe cellulaire processen beheersen, zoals de regulatie van gen transcriptie en de recombinatie van een *Ig* gebied. In **Hoofdstuk 9** bediscussieer ik de implicaties van de beschreven experimenten voor onze kennis van (hematopoïetische) gen regulatoire mechanismen en hoe deze bevindingen bij zouden kunnen dragen aan de ontwikkeling van nieuwe therapeutische invalshoeken voor menselijke ziekten.

Curriculum Vitae

Name:	Ralph Stadhouders
Date of birth:	23 September 1984
Place of birth:	Bergen op Zoom, the Netherlands
Nationality:	Dutch

EDUCATION

2010-2014 **PhD Student**
Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands

2008-2010 **Master of Science** in Molecular Medicine (*Cum Laude*)
Erasmus University, Rotterdam, the Netherlands

2005-2008 **Bachelor of Science** in Biomedical Laboratory Research (*Cum Laude*)
Avans University, Breda, the Netherlands

TRAINING

2010-2014 **PhD research**
Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands
Promoter: Prof. dr. Frank Grosveld & co-promoter: dr. Eric Soler
Subject: Gene regulatory mechanisms controlling blood cell development

2008-2010 **MSc research**
Department of Cell Biology, Erasmus MC, Rotterdam, the Netherlands
Supervisor: Dr. Eric Soler
Subject: Gene regulatory mechanisms controlling blood cell development

2007-2008 **BSc research / technician**
Department of Virology, Erasmus MC, Rotterdam, the Netherlands
Supervisor: Dr. Martin Schutten
Subject: The effect of primer-template mismatches on quantitative PCR

2006-2007 **BSc research**
Department of Cardiology, Erasmus MC, Rotterdam, the Netherlands
Supervisor: Dr. Eric Duckers
Subject: The role of HO-1 in vascular smooth muscle cell biology

AWARDS & HONORS

2010 Royal Netherlands Academy of Arts and Sciences (KNAW) Travel Grant
2009 Royal Netherlands Academy of Arts and Sciences KNAW 1-year 'Academy-Assistant Fellowship' for excellent MSc students
2008 Annual 'Silver Flame' award for best BSc thesis (Dutch Association of Biomedical Laboratory workers)

PUBLICATIONS

1. **Stadhouders R**, Thongjuea S, Kolovos P, Baymaz I, Arcangeli ML, Yu X, Demmers J, Bezstarosti K, Maas A, Kockx C, Van IJcken W, Andrieu-Soler C, Lenhard B, Grosveld F, Soler E. (2014) Control of developmentally poised erythroid genes by combinatorial corepressor actions. *Submitted*

2. **Stadhouders R***, Aktuna S*, Thongjuea S, Aghajanirefah A, Pourfarzad F, Van Ijcken W, Lenhard B, Rooks H, Best S, Menzel S, Grosveld FG, Thein SL, Soler E. (2014) HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *Journal of Clinical Investigation* 124: 1699-1710

***shared first authorship**

3. **Stadhouders R**, de Bruijn MJW, Rother MB, Yuvaraj S, Ribeiro de Almeida C, Kolovos P, Van Zelm MC, Van Ijcken W, Grosveld FG, Soler E, Hendriks RW. (2014) Pre-B Cell Receptor Signaling Induces Immunoglobulin κ Locus Accessibility by Functional Redistribution of Enhancer-Mediated Chromatin Interactions. *PLOS Biology* 12:e1001791

4. Pourfarzad F, Aghajanirefah A, de Boer E, Ten Have S, Bryn van Dijk T, Kheradmandkia S, **Stadhouders R**, Thongjuea S, Soler E, Gillemans N, von Lindern M, Demmers J, Philipsen S, Grosveld F. (2013) Locus-Specific Proteomics by TChP: Targeted Chromatin Purification. *Cell Reports* 15;4(3):589-600

5. Thongjuea S*, **Stadhouders R***, Grosveld FG, Soler E, Lenhard B. (2013) r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Research* 41(13):e132

***shared first authorship**

6. Van der Vries E, Anber J, van der Linden A, Wu Y, Maaskant J, **Stadhouders R**, van Beek R, Rimmelzwaan G, Osterhaus A, Boucher C, Schutten M. (2013) Molecular assays for quantitative and qualitative detection of influenza virus and oseltamivir resistance mutations. *Journal of Molecular Diagnostics* 15(3):347-54

7. **Stadhouders R***, Kolovos P*, Brouwer R*, Zuin J, van den Heuvel A, Kockx C, Palstra RJ, Wendt KS, Grosveld F, van Ijcken W, Soler E. (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature Protocols* 8(3):509-24

***shared first authorship**

8. **Stadhouders R**, van den Heuvel A, Kolovos P, Jorna R, Leslie K, Grosveld F, Soler E. (2012) Transcription regulation by distal enhancers: who's in the loop? *Transcription* 3(4):181-6 *[Review]*

9. Ribeiro de Almeida C*, **Stadhouders R***, Thongjuea S, Soler E, Hendriks RW. (2012) DNA-binding factor CTCF and long-range gene interactions in V(D)J recombination and oncogene activation. *Blood* 119(26):6209-18 *[Review]*

***shared first authorship**

10. **Stadhouders R***, Thongjuea S*, Andrieu-Soler C, Palstra RJ, Bryne JC, van den Heuvel A, Stevens M, de Boer E, Kockx C, van der Sloot A, van den Hout M, van Ijcken W, Eick D, Lenhard B, Grosveld F, Soler E. (2012) Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO Journal* 31(4):986-99

***shared first authorship**

11. Ribeiro de Almeida C, **Stadhouders R**, de Bruijn MJ, Bergen IM, Thongjuea S, Lenhard B, van Ijcken W, Grosveld F, Galjart N, Soler E, Hendriks RW. (2011) The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. *Immunity* 35(4):501-13

12. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, Köchert K, Bouhlel MA, Richter J, Soler E, **Stadhouders R**, Jöhrens K, Wurster KD, Callen DF, Harte MF, Giefing M, Barlow R, Stein H, Anagnostopoulos I, Janz M, Cockerill PN, Siebert R, Dörken B, Bonifer C, Mathas S. (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Medicine* 16(5):571-9

13. Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, **Stadhouders R**, Palstra RJ, Stevens M, Kockx C, van Ijcken W, Hou J, Steinhoff C, Rijkers E, Lenhard B, Grosveld F. (2010) The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes & Development* 24(3):277-89

14. **Stadhouders R**, Pas SD, Anber J, Voermans J, Mes TH, Schutten M. (2010) The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *Journal of Molecular Diagnostics* 12(1):109-17

PhD Portfolio

Name PhD student:	Ralph Stadhouders
Department:	Cell Biology
Research school:	Graduate School MGC
Period:	Sept 2010 - Sept 2014
Promoter:	Prof. dr. Frank Grosveld
Co-promoter:	Dr. Eric Soler

PhD TRAINING

Courses

2013	FEBS 17th International Summer School on Immunology
2013	Molecular Immunology
2011	Safe laboratory techniques
2011	BIOBASE course: Functional Annotation of Experimental Data using TRANSFAC® Professional, HGMD® Professional, and Genome TraxTM
2010	Epigenetic regulation in health and disease
2010	Laboratory animal handling, legislation and management ('Artikel 9')
2008-2009	Molecular and Cell Biology (shared MSc/PhD teaching programme)

Inter(national) conferences & workshops

2013-2014	ESP57 'Genomics in Medicine' course NIHES, Rotterdam, the Netherlands (*2x oral presentation & ENCODE database practical course*)
2014	MolMed course 'A broad spectrum of NGS Applications in Molecular Medicine', Rotterdam, the Netherlands (*oral presentation*)
2014	SyBoSS FP7 Consortium Internal meeting, Hohenkammer, Germany (*oral presentation*)
2012-2014	MGC PhD course 'Technology Facilities', Rotterdam/Leiden, the Netherlands (*3x oral presentation*)
2013	IUAP DevRepair 3rd meeting, Liège, Belgium (*oral presentation*)
2013	FEBS 17th International Summer School on Immunology, Rabac, Croatia (*poster presentation*)
2013	IUAP DevRepair 2nd meeting, Ghent, Belgium (*oral presentation*)
2013	11th B Cell Forum of the German Society of Immunology, Schluchsee, Germany (*poster presentation*)
2010 & 2012	CBC/CGC meeting 'Molecular Mechanisms in Cancer', Amsterdam, the Netherlands (*poster presentation*)
2012	The 18th Conference on Hemoglobin Switching, Monterey (CA), USA
2012	19th MGC PhD Workshop, Düsseldorf, Germany (*poster presentation*)
2011	MolMed Monthly Bridge meeting, Rotterdam, the Netherlands (*oral presentation*)
2011	21st Annual MGC Symposium, Leiden, the Netherlands (*oral presentation*)
2011	18th MGC PhD Workshop, Maastricht, the Netherlands
2011	EMBO Workshop on 'the Operon Model and its impact on modern molecular biology', Paris, France
2010	EuTRACC FP6 Consortium Young Scientist meeting, Dubrovnik, Croatia (*oral presentiation*)

Dankwoord *(Acknowledgements)*

En dan is het nu tijd voor de spreekwoordelijke kers op de taart en veruit meest populaire onderdeel van een proefschrift: het dankwoord. Zeker voor de niet-wetenschappers die dit lezen, moet ik allereerst benadrukken dat een proefschrift, ook al staat er maar 1 auteur op de kaft, een werk van velen is. Ik denk dat mijn proefschrift daarin nog wat extremer is dan gemiddeld: elk experimenteel hoofdstuk had niet in de huidige vorm tot stand kunnen komen zonder het harde werk van de (buitenlandse) onderzoeksgroepen waarmee we intensief samen gewerkt hebben. Daarom moeten er een hoop mensen bedankt worden: aan de slag dus! Allereerst natuurlijk de twee belangrijkste personen, zonder wie dit alles niet mogelijk was geweest: mijn promoter Frank Grosveld en co-promoter Eric Soler.

Beste Frank, bedankt dat je me de kans hebt gegeven om zowel mijn MSc stages als mijn PhD onderzoek bij jou in het laboratorium te kunnen uitvoeren. Je brede en originele kijk op de dingen bood eigenlijk altijd wel weer een oplossing of volgende stap; een gave waar ik erg jaloers op ben. Ik wil je ook bedanken voor alle moeite die je zelf in dit proefschrift en mijn carrière hebt gestoken: papers/thesis hoofdstukken/cover letters/beurs aanvragen corrigeren, aanbevelingsbrieven schrijven en de uitstekende carrière tips (zoals Barcelona). Ik stel voor dat we regelmatig contact houden! Beste Eric a.k.a. 'Monsieur', thanks for everything man. We have now worked together intensively for almost 6 years, and looking back I have to conclude that you have been the kindest, most patient and generous mentor that a MSc/PhD student can ever hope for. We hit it off immediately, and I think it's fair to say we quickly extended our status of 'colleagues' to 'friends'. I always had the feeling we were a team, in which you valued my opinion as if I were your equal. Thanks for teaching me how to pipet, behave and write like an actual scientist! And of course, thanks for all those good times at the bench and outside the lab: we had great fun ('Pastis+Nintendo Wii night', 'Belgian' dance music, all the 'Bert' jokes, speaking 'Alabama', etc. etc.)! I wish you, Charlotte and the ever-growing bunch of complementary kids all the best in la douc(h)e France! Frank, Eric, ik sta bij jullie in het krijt/I owe you big time!

Beste Rudi. Je bent niet alleen lid van mijn kleine commissie en de bewaker van de immunologische correctheid van mijn proefschrift (waarvoor veel dank!), maar toch ook een soort tweede co-promoter. Ik ben erg blij dat jij en Claudia een aantal jaar terug bij Eric en mij op bezoek kwamen met een samenwerkingsvoorstel. En wat een samenwerking is het geworden! Ik verbaas me nog steeds over je literatuur kennis en je immer optimistische kijk op de dingen, die van een dataset die op het eerste oog een tegenvaller lijkt (ik had zelf de hoop al een klein beetje opgegeven) een prachtig artikel kan maken. Dankzij jou kan ik nu een (klein) beetje meepraten met de immunologen (een prachtig vakgebied) en heb ik geweldige congressen in Duitsland en Kroatië mogen meemaken. Dank ook voor de carrière tips en de vele goede en gezellige gesprekken! Ik hoop echt dat we de samenwerking in de toekomst weer voort kunnen zetten.

Mijn dank gaat ook uit naar de overige leden van mijn kleine commissie: beste Sjaak en Danny, dank voor de snelle en uitstekende correcties op het proefschrift. Sjaak, ik wil je ook bedanken voor je vele adviezen, toffe gesprekken en borrel-biertjes door de jaren heen. Danny, ook jij bedankt voor de goede carrière tips, de uitnodigingen voor de IUAP meetings en de gastvrijheid bij jou thuis (Belgisch bier, voetbal kijken en van Vera's uitstekende kookkunsten genieten: men kan het slechter treffen toch?). Succes met het leiden van de afdeling Celbiologie 2.0!

Now a word of thanks to our foreign collaborators, who have been absolutely instrumental for the completion of this thesis. Dear Boris and Supat: many thanks for all the bioinformatical support these years! It's difficult to overestimate your contribution to all the

work described in this thesis. Dear Swee Lay, Suleyman and Stephan: working with you was a great experience, and the final result was fantastic! Swee Lay, I am honoured to have you on my defense committee - especially since I know your schedule tends to be crazy busy. Thanks a lot!

I am also very grateful to the final 2 members of my defense committee. Beste Niels, bedankt dat je mijn boekje wilde lezen en er een (hopelijk niet te) lastige vraag over wilt stellen op 12 september. Dear Thomas, inviting you over for my defense was a great idea (thanks Frank & Danny), and I am honoured to have you on my committee. Thanks a lot, and I look very much forward to our future work together in Barcelona!

Dan zijn daar de paranimfen. Robert-Jan a.k.a. 'Roberto', bedankt dat je mijn paranimf wilt zijn (ondanks dat je nu een verdiepinkje lager zit!). Je bent één van de beste en meest bescheiden wetenschappers waar ik mee heb gewerkt, en je bent altijd m'n sparring-partner geweest als er problemen opgelost of ideeën getest moesten worden. Het allerbeste bij de biochemie, laat dat HIV veld maar eens zien wat een echte transcriptie-held allemaal kan! Joey, broeder van me. Bedankt dat je naast me wilt staan die dag (in jouw Feyenoord stad): zulke morele support is hard nodig! Terwijl ik dit schrijf weet ik dat je rond die tijd ook vader wordt (en ik oom). Ik ben trots op je, weet zeker dat je een goede vader wordt en ik hoop dat ik de kleine Moos (haha) nog even kan zien voordat ik naar Spanje ga!

Tot zo ver de echte kopstukken. Maar, dan zijn er nog een hoop mensen die ik moet bedanken, en een aantal daarvan zijn hartstikke belangrijk (zo niet essentieel) geweest. Bij voorbaat mijn oprechte excuses als ik iemand vergeet…

Ik ben iedereen van de afdeling Celbiologie en het ondersteunend personeel veel dank verschuldigd, maar ik wil er toch een aantal nog even apart noemen. We beginnen bij alle collega's die door de jaren heen in het Grosveld lab hebben gewerkt: bedankt! Bijzonder in deze lange lijst zijn: Anita (m'n lab-maatje, succes met afronden!), Petros (informatics-support), Guillaume (a.k.a. the Octopus), Ruud (recombineering-goeroe), Andrea C. (top chef/host), Andrea M. (prototype cool Italian), Xiao (FACS-support), Charlotte (a.k.a. Chélot), Maureen (Harvard-trained super-mom), Mary (super-tech), Irem (MSc-'colleague'), Farzin and Ali (HEP cell hero's) and Ernie (can simply make anything work). Uit het lab van Rudi wil ik bedanken: Claudia (a true super-talent! thanks so much for coming all the way from Oxford to see me struggle), Marjolein & Ingrid (Rudi's super-techs en FACS experts) en Saravanan (microarray analysis champ).

Daarnaast wil ik bedanken: Raymond (cover letter en allerlei andere adviezen) en 'Poot lab'leden (met name Debby en Erik, bedankt dat ik mijn Celbiologie carrière heb mogen beginnen in het lab van de 'S.O.S. Publishing Group'!), de rode bloedcel mensen in het Philipsen lab (in het bijzonder Nienke, Thamar, Iléana en Sylvia), Harmen (über-whizzkid), Kerstin and Jessica (3C/4C experts, and thanks for all the thesis advice and great chats J!), Jean-Charles & Catherine (FACS phenotyping très bien), de Biomics en Proteomics crews (essentieel!), de mannen en vrouw van lab 7.10 (inclusief Alex de 'mouse-man'), Dorota en Martine (blastocyst harvest/culture expertise), Magda en Menno van de Immunologie (dank voor het delen van de prachtige FISH data), Marie from Paris (FACS data analysis expert), Ruud Delwel voor de goede gesprekken, alle mensen in het EDC die al die jaren trouw voor mijn muizen hebben gezorgd en de vele ondersteunde mensen die het leven van een PhD student zo veel makkelijker maken (Marike, Bep & Jasperina; Leo, Melle, Annet & Koos; de ICT boys). Mijn dank gaat ook uit naar mijn treinmakkers Rini en Michael voor de gezelligheid!

En dan uiteraard is er nog een laatste groep mensen aan wie ik alle dank verschuldigd ben: mijn familie en vrienden. Pa, ma, broers en zusje: ook al zien we elkaar allemaal wat minder de laatste jaren, samenkomen voelt altijd weer net als vroeger! Een veel warmer nest kan ik me niet voorstellen. Door de jaren heen hebben we ook altijd een beroep kunnen doen op de vele handige figuren in mijn familie (vooral mijn vader en opa's): hele verbouwingen konden we met een gerust hart uitbesteden. Oma's, ik ben blij dat jullie allebei nog zo gezond zijn en bij mijn

promotie aanwezig kunnen zijn. Jammer genoeg is dat de opa's niet gelukt. Ik heb veel respect voor hoe jullie met dat verlies omgaan, en ik weet zeker dat die twee oudjes ergens op één of andere manier tevreden zullen meekijken. Ome Rob, Kristel, super bedankt voor het tekenen van de kaft van dit boekje. Ik heb me nooit volledig gerealiseerd wat voor geweldige kunstenaars jullie zijn! Geniet van jullie kersverse gezinsuitbreiding! Ook belangrijk geweest zijn mijn 'schoonouders' Ad en Defi, voor wie het nooit enige moeite kostte om weer eens bij te springen of ons mee te laten eten als we onverwachts langs kwamen. Jullie zijn een grote steun voor mij! Ook niet te vergeten is mijn geweldige vriendengroep. Bij jullie kan ik de wetenschappelijke sores altijd van me af zetten: even de sportschool induiken (laatste tijd een stuk minder, sorry mannen!), ergens een biertje/wijntje drinken of gewoon een paar weken op vakantie gaan. Ik zal jullie een stuk minder gaan zien de komende tijd, maar daar krijgen jullie dan wel een fraai vakantie-adresje voor terug!

Tenslotte is daar Daniëlle. Lieve Daan, ik realiseer me heel goed dat samenleven met een wetenschapper niet altijd even makkelijk is. Zeker het laatste jaar heb je het toch wel regelmatig af moeten leggen tegen mijn werk. Wat ik zo in je bewonder is dat je me juist stimuleert om het beste eruit te halen, terwijl je daarbij vaak jezelf weg moet cijferen. Het feit dat je me alleen een paar jaar naar Barcelona laat gaan om daar verder aan m'n carrière te timmeren (en zelf niet mee kunt omdat jouw carrièrekansen hier in Nederland liggen) illustreert perfect wat voor mazzelaar ik ben. Het gaat ons lukken Daan!